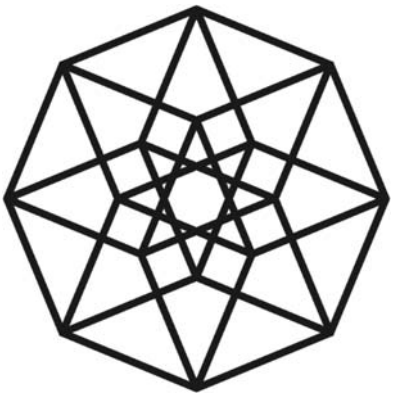
МЕЖГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«БЕЛОРУССКО-РОССИЙСКИЙ УНИВЕРСИТЕТ»

Кафедра «Высшая математика»

МЕТОДЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Методические рекомендации к лабораторным работам для студентов направления подготовки 01.03.04 «Прикладная математика» очной формы обучения



Могилев 2022

УДК 004.65
ББК 32.973.26-018.2
М54

Рекомендовано к изданию
учебно-методическим отделом
Белорусско-Российского университета

Одобрено кафедрой «Высшая математика» «29» сентября 2022 г., протокол № 1

Составители: доц. Д. В. Роголев;
доц. А. А. Романенко;
ст. преподаватель Т. Ю. Орлова

Рецензент канд. физ.-мат. наук, доц. И. И. Маковецкий

Методические рекомендации разработаны на основе рабочей программы по дисциплине «Методы анализа больших данных» для студентов направления подготовки 01.03.04 «Прикладная математика» очной формы обучения и предназначены для использования при выполнении лабораторных работ по дисциплине в седьмом семестре.

Учебно-методическое издание

МЕТОДЫ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Ответственный за выпуск	В. Г. Замураев
Корректор	И. В. Голубцова
Компьютерная верстка	Н. П. Полевничая

Подписано в печать . Формат 60×84/16. Бумага офсетная. Гарнитура Таймс.
Печать трафаретная. Усл. печ. л. . Уч.-изд. л. . Тираж 56 экз. Заказ №

Издатель и полиграфическое исполнение:
Межгосударственное образовательное учреждение высшего образования
«Белорусско-Российский университет».
Свидетельство о государственной регистрации издателя,
изготовителя, распространителя печатных изданий
№ 1/156 от 07.03.2019.
Пр-т Мира, 43, 212022, г. Могилев.

Содержание

1 Лабораторная работа № 1. Предварительная обработка значений временных рядов

Цель работы: изучение и применение для решения практических задач методов выявления аномальных значений, сглаживания временных рядов, определения наличия тренда временного ряда. Приобретение навыков статистической обработки временных рядов.

1.1 Введение

Динамические процессы, происходящие в системах, обычно представляются в виде ряда значений некоторого показателя, последовательно расположенных в хронологическом порядке. Изменение этого показателя отражает ход развития изучаемого процесса. Последовательность наблюдений одного показателя (признака), упорядоченная в зависимости от последовательно возрастающих или убывающих значений другого показателя, называется динамическим рядом, или рядом динамики. Если в качестве признака, в зависимости от которого происходит упорядочивание, берётся время, то такой динамический ряд называется временным рядом.

Элементами рядов динамики являются значения наблюдаемого показателя, называемые уровнями ряда, и моменты или интервалы времени, к которым относятся уровни. Временные ряды, в которых заданы значения показателя, относящиеся к определённым моментам времени, называются моментными. Если уровни временного ряда образуются суммированием, усреднением или каким-либо другим методом агрегирования за некоторый промежуток времени, то такие ряды называют интервальными временными рядами. Примерами могут служить ряды объёмов произведённой продукции по месяцам и ряды средних заработных плат работников по месяцам.

Длина временного ряда – время, прошедшее от начального момента наблюдений до конечного, или число уровней ряда.

1.2 Указания к выполнению

Предварительная обработка временных рядов состоит в выявлении аномальных значений ряда и сглаживании ряда.

Аномальные значения временного ряда не отвечают потенциалу исследуемой системы, и их использование для построения трендовой модели может сильно исказить получаемые результаты. Причинами появления аномальных уровней могут быть технические ошибки при сборе, обработке и передаче информации. Такие ошибки называются ошибками первого рода, их можно выявить и устранить или принять меры к их недопущению. Кроме того, аномальные уровни могут возникать из-за воздействия факторов, имеющих объективный характер, но действующих эпизодически. Такие ошибки называются ошибками второго рода, их невозможно устранить, но можно исключить из рассмотрения, заменив аномальное значение на среднеарифметическое двух соседних уровней.

Для выявления аномальных значений ряда используется критерий Ирвина,

согласно которому аномальной считается точка Y_t, отстоящая от предыдущей точки Y_{t-1} на величину, большую среднеквадратичного отклонения:

$$\lambda_i = \frac{|Y_t - Y_{t-1}|}{\sigma},$$

где λ_i – критерий Ирвина;

σ – среднеквадратичное отклонение, вычисляемое по формуле

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(Y_t - Y)^2}{n-1}}.$$

Точка считается аномальной, если $\lambda_i > \lambda_{\text{таб}}$. Табличные значения $\lambda_{\text{таб}}$ уменьшаются с ростом длины ряда, их значения приведены в таблице 1.1.

Таблица 1.1 – Табличные значения критерия Ирвина

n	10	20	30	50	100
$\lambda_{\text{таб}}$	1,5	1,3	1,2	1,1	1,0

Очень часто уровни ряда динамики колеблются, так что тенденция развития процесса скрыта случайными отклонениями. Сглаживание временного ряда позволяет отфильтровать мелкие случайные колебания и выявить основную тенденцию изменения исследуемой величины. При механическом сглаживании выравнивание отдельных уровней производится с использованием значений соседних уровней. Для сглаживания используются следующие методы.

1 Простая (среднеарифметическая) скользящая средняя

$$\widetilde{Y}_t = \frac{\sum_{i=t-p}^{t+p} Y_i}{2p+1}, \quad p < t < n-p.$$

Сглаженное значение $\widetilde{Y}_t$ является среднеарифметическим из $2p+1$ соседних точек. Наиболее часто используется сглаживание по 5 точкам.

2 Взвешенная (средневзвешенная) скользящая средняя

$$\widetilde{Y}_t = \frac{\sum_{i=t-p}^{t+p} \rho_i Y_i}{\sum_{i=t-p}^{t+p} \rho_i}, \quad p < t < n-p.$$

В этом методе каждая из точек входит в общую сумму с весовым коэффициентом ρ_i. Для сглаживания по пяти точкам используют весовые коэффициенты (–3, 12, 17, 12, –3). Для сглаживания по 7 точкам используются коэффици-

енты (–2, 3, 6, 7, 6, 3, –2) или (5, –30, 75, 131, 75, –30, 5).

3 Среднехронологическая скользящая средняя

$$\widetilde{Y}_t = \frac{\frac{Y_{t-T/2}}{2} + \sum\limits_{i=t-T/2+1}^{t+T/2-1} Y_i + \frac{Y_{t+T/2}}{2}}{T-1}, \quad \frac{T}{2} < t < n - \frac{T}{2}.$$

Эта формула используется для моментных временных рядов.

4 Экспоненциальное сглаживание.

В этом методе для сглаживания текущей точки используются все предшествующие точки, причём значения весовых коэффициентов экспоненциально убывают по мере удаления от текущей точки. Выражение для текущей сглаженной точки представляет собой функцию от текущей несглаженной точки и предыдущей сглаженной

$$\widetilde{Y}_t = \alpha Y_t + (1-\alpha)\widetilde{Y}_{t-1}, \quad 0 < \alpha < 1,$$

где $\alpha -$ параметр сглаживания.

Фиктивное начальное значение сглаженного ряда принимают равным первой точке $\widetilde{Y}_0 = \widetilde{Y}_1$ или среднеарифметическому первых трех точек $\widetilde{Y}_0 = \frac{Y_1 + Y_2 + Y_3}{3}$.

При сглаживании временного ряда по $2p+1$ соседним точкам p точек в начале и в конце ряда остаются несглаженными. Эти точки следует либо исключить из рассмотрения, либо использовать для них специальные формулы сглаживания для крайних точек. В частности, для сглаживания по трём точкам можно использовать формулы

$$\widetilde{Y}_1 = \frac{5Y_1 + 2Y_2 - Y_3}{6}; \quad \widetilde{Y}_n = \frac{5Y_n + 2Y_{n-1} - Y_{n-2}}{6}.$$

5 Метод проверки разностей средних уровней.

Исходный ряд из n точек делится на два ряда с примерно одинаковым числом точек n_1 и n_2 $(n = n_1 + n_2)$, и для каждой из частей вычисляются средние значения и дисперсия. Затем проверяется гипотеза об однородности дисперсий частей ряда с помощью критерия Фишера:

$$F = \begin{cases} \sigma_1^2 / \sigma_2^2, \text{если} & \sigma_1^2 > \sigma_2^2; \\ \sigma_2^2 / \sigma_1^2, \text{если} & \sigma_2^2 > \sigma_1^2. \end{cases}$$

Если полученное значение $F < F_{\text{табл}}$, то гипотеза об однородности дисперсий принимается. Если $F \geq F_{\text{табл}}$, то гипотеза об однородности дисперсий отклоняется, а метод не даёт ответа на вопрос о наличии или отсутствии тренда.

Табличное значение $F_{\text{табл}}$ зависит от уровня значимости и длины сравниваемых рядов. Значения критерия Фишера для 5-процентного уровня ошибки приведены в таблице 1.2.

Таблица 1.2 – Значения критерия Фишера

n	20	25	30
$F_{\text{табл}}$	2,12	1,96	1,84

Окончательная проверка гипотезы об отсутствии тренда производится с использованием t-критерия Стьюдента, вычисляемого по формуле

$$t = \frac{\left|\overline{Y_1} - \overline{Y_2}\right|}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

где σ – среднеквадратическое отклонение разности средних уровней, $\sigma = \sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 2}}$.

Если расчётное значение $t < t_{\text{табл}}$, то гипотеза принимается, т. е. тренда нет; в противном случае тренд есть. Для определения табличного значения число степеней свободы принимается равным $n_1 + n_2 - 2$. Значения статистики Стьюдента приведены в таблице 1.3.

Таблица 1.3 – Значения критерия Стьюдента

n	28	30	50	58	60
$t_{\text{табл}}(\alpha = 0,05)$	2,048	2,042	2,009	2,002	2,000
$t_{\text{табл}}(\alpha = 0,01)$	2,763	2,750	2,678	2,663	2,666
$t_{\text{табл}}(\alpha = 0,2)$	1,303	1,310	1,290	1,295	1,296

6 Метод Фостера – Стюарта.

Этот метод даёт более надёжные результаты по сравнению с методом разностей средних уровней. Кроме самого тренда, он позволяет установить наличие тренда дисперсии. При отсутствии тренда дисперсии разброс уровней ряда постоянен, при наличии тренда дисперсии дисперсия увеличивается или уменьшается. Метод предполагает построение двух последовательностей:

$$k_t = \begin{cases} 1, \text{если } Y_t \text{ больше всех предыдущих } Y_i; \\ 0, \text{в противном случае;} \end{cases}$$

$$l_t = \begin{cases} 1, \text{если } Y_t \text{ меньше всех предыдущих } Y_i; \\ 0, \text{в противном случае,} \end{cases}$$

где $t = \overline{2,n}$.

Для выявления изменчивости временного ряда и дисперсии вычисляются величины s и d по формулам

$$s = \sum_{t=2}^{n} (k_t + l_t);\ d = \sum_{t=2}^{n} (k_t - l_t).$$

Величина s характеризует изменение временного ряда, она может принимать значение от 0 (когда все уровни ряда равны) до $n-1$ (ряд монотонный). Величина d характеризует изменение дисперсии временного ряда и изменяется от $-(n-1)$ (когда ряд монотонно убывает) до $n-1$ (когда ряд монотонно возрастает). Эти величины являются случайными с математическим ожиданием μ для значения s и 0 для значения d.

Для проверки гипотез о случайности отклонения величины s от её математического ожидания μ, а также о случайности отклонения величины d от нуля с помощью критерия Стьюдента для средней и для дисперсии вычисляются статистики t_s, t_d по формулам

$$t_s = \frac{|s-\mu|}{\sigma_1};\quad \sigma_1 = \sqrt{2\ln n - 3{,}4253};\quad t_d = \frac{|d-0|}{\sigma_2};\quad \sigma_2 = \sqrt{2\ln n - 0{,}8456},$$

где μ – математическое ожидание величины s для случайного временного ряда;

σ_1 – среднеквадратичное отклонение s;

σ_2 – среднеквадратичное отклонение d.

Полученные значения t_s, t_d необходимо сравнить с табличными значениями критерия Стьюдента $t_{\text{табл}}$. Если $t < t_{\text{табл}}$, то соответствующий тренд отсутствует: т. е. если $t_s > t_{\text{табл}}$, а $t_d < t_{\text{табл}}$, то тренд ряда есть, а тренда дисперсии нет.

1.3 Задание

1 Выявить в заданном временном ряду аномальные значения по критерию Ирвина. Обнаруженные аномальные значения заменить путём интерполирования по соседним точкам.

2 Выполнить сглаживание заданного ряда следующими методами:

1) простая (среднеарифметическая) скользящая средняя по 5 точкам;
2) взвешенная (средневзвешенная) скользящая средняя по 5 точкам;
3) взвешенная (средневзвешенная) скользящая средняя по 7 точкам;
4) среднехронологическая по 12 точкам;
5) экспоненциальное сглаживание.

3 На одной диаграмме построить графики исходного ряда и все сглаженные ряды.

4 По заданным значениям временного ряда определить наличие тренда методами:

1) проверки разностей средних уровней;
2) Фостера – Стюарта.

5 Сделать окончательный вывод о наличии или отсутствии тренда временного ряда.

1.4 Содержание отчёта

Отчёт должен содержать:

1) исходный временной ряд, значения оценок математического ожидания и среднеквадратического отклонения;

2) значения критерия Ирвина для каждой пары точек, выявленные аномальные значения по критерию Ирвина, полученный после замены аномальных значений временной ряд;

3) сглаженные временные ряды;

4) графики исходного ряда и всех сглаженных рядов;

5) фактические и критические значения статистик критериев проверки гипотез о наличии тренда;

6) вывод о наличии или отсутствии тренда.

2 Лабораторная работа № 2. Корреляционный и регрессионный анализ

Цель работы: закрепление теоретических знаний и приобретение практических навыков в построении регрессионных моделей объекта по экспериментальным данным (ЭД).

2.1 Введение

Одной из типовых задач обработки многомерных результатов экспериментов является определение зависимости между различными показателями.

Решение общей задачи построения регрессионной модели разбивается на несколько этапов:

1) вывод соотношений для оценки параметров заданных регрессионных моделей;

2) оценка параметров регрессионных моделей;

3) проверка адекватности регрессионной модели;

4) оценка точности регрессионных моделей;

5) формирование выводов о возможности применения разработанных регрессионных моделей.

2.2 Указания к выполнению

Каждый студент обрабатывает свой вариант экспериментальных данных (таблицы 2.1–2.3). Для выполнения вычислений можно разработать соответствующие процедуры с использованием любого универсального языка программирования, можно воспользоваться возможностями табличного процессора, например Excel. Обработка данных ведётся применительно к трём видам уравнений регрессии, исходные данные приведены в таблицах: в таблице 2.1 – линейная регрессия $y = a_1 x + a_0$ и параболическая регрессия второго порядка $y = a_2 x^2 + a_1 x + a_0$; в таблице 2.2 – множественная регрессия $y = a_2 x_2 + a_1 x_1 + a_0$; в таблице 2.3 приведены варианты индивидуального задания.

Парные коэффициенты корреляции между признаками вычисляются \по формуле

$$\rho_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \mu_x\right)\left(y_i - \mu_y\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \mu_x\right)^2 \sum_{i=1}^{n}\left(y_i - \mu_y\right)^2}}.$$

Значение коэффициента корреляции лежит в пределах от – 1 до + 1. Если случайные величины независимы, то коэффициент корреляции обязательно равен нулю, обратное же утверждение неверно. Коэффициент корреляции характеризует значимость линейной связи между случайными величинами (параметрами):

$\rho_{xy} = 1$ – значения x_i и y_i полностью совпадают. Иначе говоря, имеет место функциональная зависимость: зная значение одного параметра, можно однозначно указать значение другого параметра;

$\rho_{xy} = -1$ – величины x_i и y_i принимают противоположные значения. В этом случае имеет место функциональная зависимость;

$\rho_{xy} = 0$ – величины x_i и y_i практически не связаны друг с другом линейным соотношением. Это не означает отсутствия каких-то других (например, нелинейных) связей между параметрами;

$0 < \left|\rho_{xy}\right| < 1$ – однозначной линейной связи величин x_i и y_i нет. И чем меньше абсолютная величина коэффициента корреляции, тем в меньшей степени по значениям одного параметра можно предсказать значение другого.

Вывод соотношений для расчётов оценок параметров регрессионных моделей производится с использованием метода наименьших квадратов.

Метод наименьших квадратов (МНК) как вычислительная процедура был описан Лагранжем в 1806 г. Им также было предложено название этого метода.

В основе МНК лежат следующие положения:

1) значения величин ошибок и факторов независимы, а значит, и некоррелированы, т. е. предполагается, что механизмы порождения помехи не связаны с механизмом формирования значений факторов;

2) математическое ожидание ошибки a_0 должно быть равно нулю (постоянная составляющая входит в коэффициент ε), иначе говоря, ошибка является центрированной величиной;

3) выборочная оценка дисперсии ошибки должна быть минимальна.

Пусть уравнение линейной регрессии имеет вид

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{ij} + \varepsilon_i, \quad i = \overline{1,n}.$$

Обозначим

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix};\ X=\begin{pmatrix} 1 & x_{11} & \dots & x_{1m} \\ 1 & x_{21} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nm} \end{pmatrix};\ \alpha=\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \dots \\ \alpha_m \end{pmatrix};\ \varepsilon=\begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \dots \\ \varepsilon_m \end{pmatrix}.$$

Тогда в матричной форме модель примет вид $Y=X\alpha+\varepsilon$. Оценкой является уравнение $Y=Xa+e$. Для оценки вектора неизвестных параметров α применяется МНК. Условие минимизации остаточной суммы может быть записано в виде

$$S=\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2=\sum_{i=1}^{n}e_i^2=e^Te=\left(Y-Xa\right)^T\left(Y-Xa\right)\to\min.$$

На основании необходимого условия экстремума функции нескольких переменных необходимо приравнять частные производные по этим переменным к нулю или в матричном виде

$$\frac{\partial S}{\partial a}=\begin{pmatrix} \frac{\partial S}{\partial a_0} & \frac{\partial S}{\partial a_1} & \dots & \frac{\partial S}{\partial a_m} \end{pmatrix}=O.$$

Оценка параметров производится путём решения систем уравнений

$$\frac{\partial S}{\partial a_0}=0;\ \frac{\partial S}{\partial a_1}=0;\ \dots;\ \frac{\partial S}{\partial a_m}=0.$$

Средняя относительная ошибка аппроксимации δ вычисляется по формуле

$$\delta=\frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i-\hat{y}_i\right|}{y_i},$$

где y_i- экспериментальные значения результативного признака;

$\hat{y}_i-$ рассчитанные по построенному уравнению регрессии значения результативного признака;

$n-$ число элементов выборки.

Для выполнения проверки на 5-процентном уровне значимости гипотезы H_0 о равенстве нулю коэффициентов уравнения регрессии необходимо рассчитать показатели по формулам

$$s_{\text{регр}}=\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2;\ s_{\text{ост}}=\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^2;\ s=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2,$$

где y_i- экспериментальные значения результативного признака;

$\hat{y}_i-$ рассчитанные по построенному уравнению регрессии значения результативного признака;

$\overline{y}-$ среднее значение результативного признака, рассчитанное как среднее

арифметическое по экспериментальным данным;

n – число элементов выборки.

Оценка коэффициента линейной детерминации, модуль оценки множественного коэффициента корреляции и оценка нормированного коэффициента линейной детерминации вычисляются соответственно по формулам

$$\widehat{R}^2 = 1 - \frac{s_{\text{ост}}}{s};\ \ \widehat{R} = \sqrt{\widehat{R}^2};\ \ \widetilde{R}^2 = 1 - \frac{s_{\text{ост}}}{n-m-1} \cdot \frac{n-1}{s}.$$

Нормированный коэффициент детерминации $\widetilde{R}^2$, в отличие от коэффициента линейной детерминации $\widehat{R}^2$, увеличивающегося при увеличении числа m регрессоров, может и уменьшаться. Чем больше его значение, тем качественнее уравнение регрессии.

Проверка гипотезы $H_0 : a_1 = a_2 = ... = a_m = 0$ производится на основе анализа статистики, вычисляемой по формуле

$$F_{m;n-m-1} = \frac{s_{\text{регр}}}{m} \cdot \frac{n-m-1}{s_{\text{ост}}} = \frac{\widehat{R}^2 \left(n-m-1\right)}{m\left(1-\widehat{R}^2\right)}$$

и имеющей (в предположении справедливости H_0) распределение Фишера – Снедекора с m и $(n-m-1)$ степенями свободы. Вычисленное значение сравнивается с критическим значением $f_{0,05;m;n-m-1}$, найденным по статистическим таблицам распределения Фишера – Снедекора. Если наблюдаемое значение статистики $F_{m;n-m-1} > f_{0,05;m;n-m-1}$, то гипотеза H_0 отвергается на 5-процентном уровне значимости, в противном случае гипотеза принимается.

При проверке гипотез о равенстве нулю полученных коэффициентов регрессии в качестве критерия используется t-критерий Стьюдента, определяемый по формуле

$$t_{a_i} = \frac{a_i}{\sqrt{\dfrac{\sigma^2_{\text{ост}} \sum\limits_{j=1}^{n} x_{ij}}{n \sum\limits_{j=1}^{n} \left(x_{ij} - \overline{x}_i\right)^2}}},$$

где t_{a_i} – значения критерия Стьюдента для коэффициентов a_i;

$\sigma^2_{\text{ост}}$ – остаточная дисперсия уравнения регрессии, $\sigma^2_{\text{ост}} = \sqrt{\dfrac{\sum\limits_{i=1}^{n} \left(y_i - \overline{y}\right)^2}{n-m-1}} -$;

n – число элементов в выборке;

m – число переменных в уравнении регрессии, для парной линейной регрессии $m = 1$.

Полученные фактические значения критерия Стьюдента сравниваются с табличными значениями $t_{m;\frac{\alpha}{2};n-m-1}$. Если оказывается, что рассчитанное значение больше критического, то соответствующий коэффициент статистически значим; в противном случае коэффициент незначим.

Оценка точности производится путём расчёта суммы квадратов отклонений фактических значений от расчётных значений функций, полученных по уравнениям регрессии. Если эти отклонения много меньше (более чем на порядок) значений функции во всех точках, то уравнение регрессии хорошо описывает ЭД. При отклонениях, хотя бы в одной точке сравнимых со значениями функции, уравнения регрессии непригодны для описания всей совокупности ЭД.

По результатам выполнения двух предыдущих пунктов задания формируются выводы о возможности аппроксимации данных указанными уравнениями регрессии.

Наряду с точечными оценками a_i коэффициентов регрессии α_i регрессионный анализ позволяет получать и интервальные оценки коэффициентов с доверительной вероятностью γ.

Интервальная оценка с доверительной вероятностью γ для параметра α_i имеет вид

$$a_i - t_{1-\gamma} \cdot \hat{s}_{a_i} \le \alpha_i \le a_i + t_{1-\gamma} \cdot \hat{s}_{a_i},$$

где $t_{1-\gamma}$ находят по таблице *t*-распределения при вероятности $1 - \gamma$ и числе степеней свободы $v = n - m - 1$, а $\hat{s}_{a_i}$ вычисляется по формуле

$$\hat{s}_{a_{i-1}} = \frac{1}{n-m-1}\left(Y - \hat{Y}\right)^T \left(Y - \hat{Y}\right) x_{ii}^*, \quad i = \overline{1, m+1},$$

где x_{ii}^* – *i*-й диагональный элемент матрицы $\left(X^T X\right)^{-1}$.

2.3 Задание

Построить регрессионные модели объектов по заданным ЭД (см. таблицы 2.1–2.3).

Задание предусматривает построение трёх регрессионных моделей:

1) построение уравнения линейной регрессии $y = a_1 x + a_0$;

2) построение уравнения параболической регрессии второго порядка $y = a_2 x^2 + a_1 x + a_0$;

3) построение уравнения множественной регрессии $y = a_2 x_2 + a_1 x_1 + a_0$.

Значения результатов наблюдений величин *x*, *y* и *z* задаются в таблицах.

Требуется:

1) рассчитать парные коэффициенты корреляции между признаками и сделать вывод о силе линейной связи результативного признака с регрессорами и вывод о силе линейной связи между результативным признаком и регрессорами;

2) вычислить оценки параметров модели линейной регрессии, среднюю относительную ошибку аппроксимации δ.

Предположив выполнение условий линейного регрессионного анализа:

1) оценить статистическую значимость уравнения регрессии (проверить на 5-процентном уровне значимости гипотезу H_0 о равенстве нулю коэффициентов уравнения регрессии);

2) проверить на 5-процентном уровне значимости гипотезы $H_0^{(j)} : a_j \neq 0$.

Систематизировать результаты, рассчитав:

1) коэффициент линейной детерминации R^2 (*R*-квадрат), нормированный $\hat{R}^2$, ошибку аппроксимации δ, *F*-статистику и критическую точку $f_{0,05;k,n-k-1}$;

2) 95-процентные доверительные интервалы для коэффициентов уравнения регрессии;

3) числовые значения *t*-статистик и критическую точку $t_{0,05;n-m-1}$ для коэффициентов уравнения регрессии.

2.4 Содержание отчёта

Отчёт должен содержать:

1) заданные выборки в исходном виде, значения оценок математических ожиданий и среднеквадратических отклонений по каждому показателю;

2) системы уравнений для оценок коэффициентов уравнений регрессии;

3) значения коэффициентов уравнений регрессии;

4) фактические и критические значения статистик критерия проверки гипотез для линейной регрессии;

5) заключение о возможности применения полученной модели;

6) отклонения расчётных значений от фактических значений функций;

7) выводы по результатам обработки экспериментальных данных.

Таблица 2.1 – Линейная регрессия

x	*y*										
	1	2	3	4	5	6	7	8	9	10	11
0,32	2,62	2,54	2,47	3,35	3,18	3,11	3,12	13,24	8,57	7,13	3,19
0,08	1,05	0,92	0,82	1,53	1,44	1,48	1,26	12,38	7,49	6,49	1,20
1,02	7,21	7,26	7,24	8,30	8,31	8,42	8,58	15,65	12,13	8,88	8,69
0,19	1,90	1,72	1,43	2,34	2,29	2,23	2,21	12,78	8,10	6,77	2,14
0,00	0,70	0,38	0,25	0,94	0,97	0,91	0,78	12,04	7,07	6,02	0,57
1,06	7,55	7,47	7,50	8,61	8,54	8,72	8,93	15,66	12,18	9,06	8,96
0,26	2,34	2,16	2,11	2,94	2,81	2,86	2,67	12,96	8,30	6,86	2,63
0,16	1,58	1,41	1,39	2,11	2,10	2,09	1,88	12,59	7,86	6,68	1,84
0,45	3,50	3,52	3,31	4,30	4,30	4,11	4,22	13,58	9,25	7,43	4,20
1,58	10,74	11,00	11,07	12,25	12,35	12,53	12,80	17,56	14,63	10,52	13,02
1,32	9,24	9,31	9,27	10,31	10,50	10,67	10,92	16,63	13,46	9,78	11,04
1,47	10,15	10,32	10,36	11,35	11,59	11,81	12,08	17,04	14,26	10,26	12,11
0,55	4,08	4,09	3,94	4,87	4,94	4,85	4,98	13,99	9,80	7,53	4,96
0,63	4,77	4,70	4,62	5,41	5,42	5,61	5,49	14,24	10,21	7,95	5,51
1,89	12,88	13,03	13,26	14,50	14,73	15,03	15,30	18,48	16,17	11,32	15,58

Таблица 2.2 – Множественная регрессия

Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
Австралия	80	17800	15	8	16848	85
Австрия	79	8000	12	11	18396	58
Аргентина	75	33900	20	9	3408	86
Бангладеш	53	125000	35	11	202	16
Беларусь	76	10300	13	11	6500	65
Бельгия	79	10100	12	11	17912	96
Бразилия	67	156600	21	9	2354	75
Буркина-Фасо	50	10000	47	18	357	15
Великобритания	80	58400	13	11	15974	89
Вьетнам	68	73100	27	8	230	20
Гаити	47	6500	40	19	383	29
Германия	79	81200	11	11	17539	85
Гондурас	70	5600	35	6	1030	44
Гонконг	80	5800	13	6	14641	94
Египет	63	60000	29	9	748	44
Замбия	45	9100	46	18	573	42
Индия	59	911600	29	10	275	26
Ирландия	78	3600	14	9	12170	57
Испания	81	39200	11	9	13047	78
Италия	81	58100	11	10	17500	69
Канада	81	29100	14	8	19904	77
Китай	69	1205200	21	7	377	26
Колумбия	75	35600	24	6	1538	70
Коста-Рика	79	3300	26	4	2031	47
Куба	78	11100	17	7	1382	74
Малайзия	72	19500	29	5	2995	43
Марокко	70	28600	29	6	1062	46
Мексика	77	91800	28	5	3604	73
Нидерланды	81	15400	13	9	17245	89
Новая Зеландия	80	3524	16	8	14381	84
Норвегия	81	4300	13	10	17755	75
ОАЭ	74	2800	28	3	14193	81
Польша	77	38600	14	10	4429	62
Португалия	78	10500	12	10	9000	34
Россия	74	149200	13	11	6680	74
Саудовская Аравия	70	18000	38	6	6651	77
Северная Корея	73	23100	24	6	1000	60
Сингапур	79	2900	16	6	14990	100
США	79	260800	15	9	23474	75
Таиланд	72	59400	19	6	1800	22
Турция	73	62200	26	6	3721	61
Украина	75	51800	12	13	2340	67
Филиппины	68	69800	27	7	867	43

Окончание таблицы 2.2

Страна	Y	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$	$x^{(5)}$
Финляндия	80	5100	13	10	15877	60
Франция	82	58000	13	9	18944	73
Чили	78	14000	23	6	2591	85
Швейцария	82	7000	12	9	22384	62
Швеция	81	8800	14	11	16900	84
Эфиопия	54	55200	45	14	122	12
ЮАР	68	43900	34	8	3128	49
Южная Корея	74	45000	16	6	6627	72
Япония	82	125500	11	7	19860	77

Здесь Y – ожидаемая продолжительность жизни женщины (в годах);

$x^{(1)}$ – численность населения (в тыс. чел.);

$x^{(2)}$ – рождаемость (на 1000 чел.);

$x^{(3)}$ – смертность (на 1000 чел.);

$x^{(4)}$ – ВВП на душу населения (в долл. США по покупательной способности валют);

$x^{(5)}$ – процент городского населения.

Таблица 2.3 – Номера признаков для регрессионного анализа

Вариант	Номер факторного признака	
1	1	2
2	1	3
3	1	4
4	1	5
5	2	3
6	2	4
7	2	5
8	3	4
9	3	5
10	4	5

3 Лабораторная работа № 3. Дискриминантный анализ

Цель работы: закрепление теоретических знаний и приобретение практических навыков в обработке данных с помощью дискриминантного анализа.

3.1 Введение

Дискриминантный анализ является разделом многомерного статистического анализа, который включает в себя методы классификации многомерных наблюдений по принципу максимального сходства при наличии обучающих признаков.

Напомним, что в кластерном анализе рассматриваются методы многомерной классификации без обучения. В дискриминантном анализе новые кластеры не образуются, а формулируется правило, по которому объекты подмножества,

подлежащего классификации, относятся к одному из уже существующих (обучающих) подмножеств (классов) на основе сравнения величины дискриминантной функции классифицируемого объекта, рассчитанной по дискриминантным переменным, с некоторой константой дискриминации.

Предположим, что существуют две или более совокупности (группы) и что мы располагаем множеством выборочных наблюдений над ними. Основная задача дискриминантного анализа состоит в построении с помощью этих выборочных наблюдений правила, позволяющего отнести новое наблюдение к одной из совокупностей.

3.2 Задание и указания к выполнению

1 В файле firm.sta имеются данные по 12 предприятиям, которые характеризуются тремя экономическими показателями: labor – производительность труда, defect – удельный вес потерь от брака (%) и fund – фондоотдача активной части основных производственных фондов. Из этих предприятий выделены две обучающие выборки (переменная firm), первая из которых включает 4 предприятия группы A, а вторая – 5 предприятий группы B. Требуется классифицировать в одну из групп A или B оставшиеся три предприятия.

2 Перед выполнением дискриминантного анализа необходимо убедиться в том, что переменные, характеризующие предприятия, являются нормально распределёнными и дисперсии, и ковариации этих переменных внутри групп однородны. Для этого используется дисперсионный анализ. Необходимые опции реализованы в модуле ANOVA (рисунок 3.1).

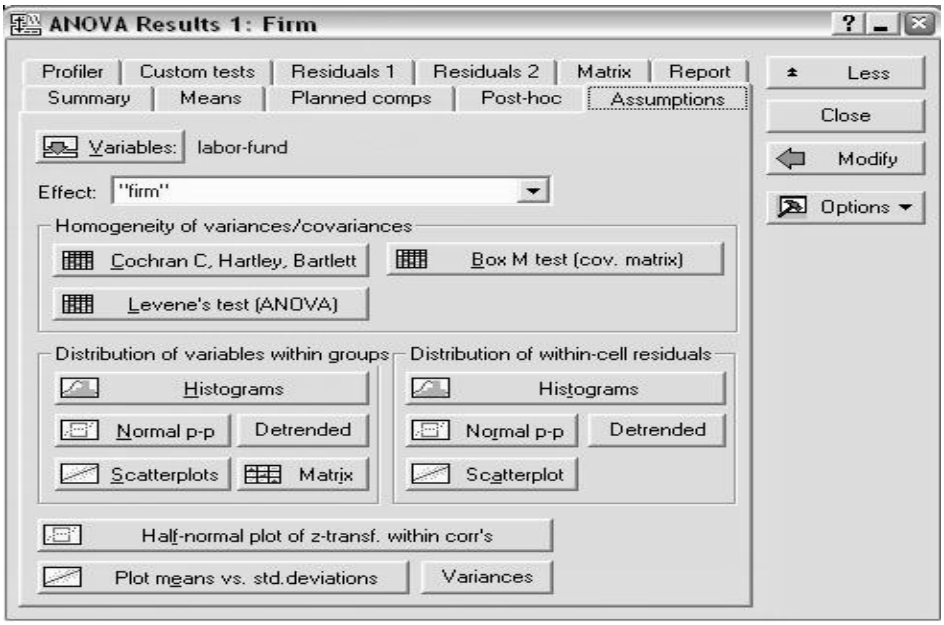


Рисунок 3.1

Выполнив опцию Statistics \ ANOVA и выбрав One-way ANOVA, задайте лист зависимых (dependent) переменных labor, defect, fund и независимую переменную (factor) firm. Нажав OK, выберите внизу появившегося окна опцию More results и затем вкладку Assumptions.

Затем, выбрав один из тестов в группе Homogeneity of variances / covariances (например, М-тест Бокса) путём нажатия соответствующей кнопки, получим результаты, которые убеждают нас в однородности дисперсий и ковариаций внутри двух групп.

Для проверки на нормальность распределения воспользуйтесь группой кнопок Distribution of variables within groups, например, графиками поля рассеяния: Scatterplots.

3 Выполните дискриминантный анализ имеющихся 9 предприятий, воспользовавшись меню Statistics\ Multivariate Exploratory Techniques\ Discriminant Analysis и указав в появившемся окне в качестве группировочной переменной (Grouping variable) firm, а в качестве независимых (Independent variable list) остальные labor, defect и fund. В появившемся окне нажмите кнопку Summary. Получим результаты дискриминантного анализа по каждой переменной, в частности, лямбды Уилкса как для всей дискриминации, так и отдельно для каждой переменной и значимость переменных для классификации.

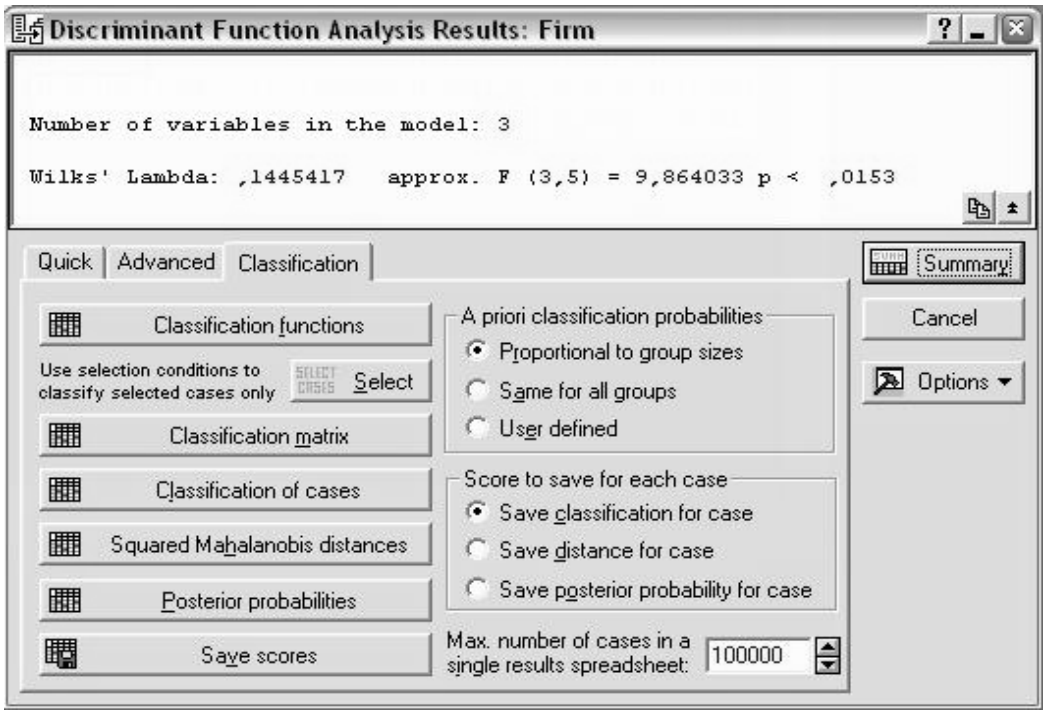


Рисунок 3.2

Вернувшись в окно Discriminant Function Analysis Results (рисунок 3.2), выберите вкладку Classification и затем Classification functions. Получим значения коэффициентов дискриминантных функций с послеопытными вероятностями попадания предприятия в одну из групп. Должны получиться следующие дискриминантные функции:

$$f_A = -54{,}87 + 9{,}13 labor + 6{,}39 defect + 10{,}56 fund;$$

$$f_B = -25{,}18 + 6{,}42 labor + 11{,}07 defect + 3{,}35 fund.$$

Если в этом же окне выбрать опцию Classification matrix, получим матрицу,

по строкам которой фактическая классификация, а по столбцам – полученная по модели. В идеальном случае они должны совпадать, и матрица должна иметь диагональный вид при проценте корректных наблюдений 100.

Далее, использовав опции Classification of cases или Posterior probabilities, получим соответственно классификацию по наблюдениям и вероятности отнесения каждого наблюдения к каждой из двух групп (A или B) (рисунок 3.3). Причём классифицированы будут и последние 3 наблюдения, для которых мы не имели первоначально информации о том, к какой из групп они относятся (наблюдения 10 и 11 – к группе B, а 12 – к группе A). Также можно получить квадрат расстояния Махаланобиса от центра каждой из групп с помощью опции Squared Mahalanobis distances.

	Posterior Probabilities (Firm) Incorrect classifications are marked with *		
Case	**Observed Classif.**	A p=,44444	B p=,55556
1	A	0,999868	0,000132
2	A	0,999568	0,000432
3	A	0,999740	0,000260
4	A	0,999989	0,000011
5	B	0,000478	0,999522
6	B	0,000000	1,000000
7	B	0,009825	0,990175
8	B	0,010251	0,989749
9	B	0,000000	1,000000
10	---	0,000492	0,999508
11	---	0,000935	0,999065
12	---	1,000000	0,000000

Рисунок 3.3

4 Выполните пошаговый дискриминантный анализ. Вернитесь к первоначальному окну дискриминантного анализа (Statistics\ Multivariate Exploratory Techniques\ Discriminant Analysis) и поставьте галочку напротив опции Advanced options. Нажмите OK. Выберите вкладку Advanced (обратите внимание на возможности изменения значений F критерия для включения/исключения переменной и вида отображения – конечного результата или результатов по шагам) и метод (Method) пошагового анализа: Forward (включения) или Backward (исключения). При этом опция Standard относится к стандартному алгоритму анализа, выполненному нами в предыдущем пункте. Нажмите OK и получите окно результатов, имеющее такой же, как и в п. 3, вид.

Выполните пошаговый анализ методом последовательного включения переменных и методом исключения.

5 Получите результаты в п. 2–4.

3.3 Содержание отчёта

Отчёт должен представлять собой содержательные выводы по результатам всех выполненных расчётов.

4 Лабораторная работа № 4. Метод главных компонент

Цель работы: лабораторная работа по методу главных компонент выполняется на демонстрационной версии статистического пакета Statgraphics. Продемонстрировать реализацию метода.

4.1 Введение

В статистической работе часто встречаются ситуации, когда общее число признаков, регистрируемых на каждом из множества исследуемых объектов, довольно велико. Представление каждого многомерного наблюдения в виде вектора некоторых вспомогательных показателей с существенно меньшим числом компонент может быть обусловлено следующими причинами:

1) необходимость наглядного представления (визуализации) исходных данных, что достигается их проецированием на специально подобранное трёхмерное пространство плоскость или числовую прямую;

2) стремление к лаконизму исследуемых моделей, обусловленному упрощением интерпретации полученных статистических выводов;

3) ограниченные возможности человека в одновременном охвате большого числа частных критериев;

4) необходимость существенного сжатия объёмов хранимой статистической информации (без видимых потерь в её информативности), если речь идёт о записи и хранении данных в специальной базе данных.

Имеется, по крайней мере, три основных типа принципиальных предпосылок, обуславливающих переход от большого числа исходных показателей состояния анализируемой системы к существенно меньшему числу наиболее информативных переменных. Это, во-первых, дублирование информации, доставляемой сильно взаимосвязанными признаками; во-вторых, неинформативность признаков, мало меняющихся при переходе от одного объекта к другому; в-третьих, возможность агрегирования, т. е. простого или «взвешенного» суммирования, по некоторым признакам.

В задачах классификации исследователя интересуют, в первую очередь, лишь те признаки, которые обнаруживают наибольшую изменчивость при переходе от одного объекта к другому. С другой стороны, не обязательно для описания состояния объекта использовать какие-то из исходных, непосредственно замеренных на объекте, признаков. Именно эти принципиальные установки заложены в сущность того линейного преобразования исходной системы признаков, которое приводит к главным компонентам.

4.2 Указания к выполнению

В качестве примера поставим задачу: оценить уровень развития здравоохранения в субъектах Центрального федерального округа России (ЦФО) в 2000 г. по следующим показателям: обеспеченность населения врачами, человек на 10000 населения (X1); обеспеченность средним медицинским персоналом, человек на 10000 населения (X2); обеспеченность больничными койками, число на 10000

населения (X3); обеспеченность врачебными амбулаторно-поликлиническими учреждениями, число посещений в смену на 10000 населения (X4). Далее приводится таблица 4.1 с исходными статистическими данными.

Исследование начинается с ввода исходных статистических данных в новую электронную таблицу. Таблица организована таким образом, что её строкам должны соответствовать объекты (наблюдения), а столбцам – признаки. Для именования переменных (признаков) и задания их типа нужно маркировать требуемую колонку и, щёлкнув правой кнопкой мыши, выбрать команду контекстного меню Modify Column. В появившемся окне диалога нужно указать требуемые параметры.

Таблица 4.1 – Исходные статистические данные

Субъект ЦФО	X1	X2	X3	X4
Белгородская область	38,2	117,4	125	199,8
Брянская область	36,6	109,6	123,1	202,9
Владимирская область	36,2	104,3	121	293,9
Воронежская область	49,3	113,3	118,9	192
Ивановская область	51,3	138,8	142	188,7
Калужская область	39,5	104,1	114,4	242,7
Костромская область	35,9	120,1	145,3	187
Курская область	45,5	107	115,3	206
Липецкая область	39,2	123,3	130,7	291,4
Московская область	33,9	87,9	106,7	246,7
Орловская область	36,7	112,2	118,9	180
Рязанская область	51,7	120,3	131,7	215,3
Смоленская область	57,6	109	130,6	230,9
Тамбовская область	33	104,1	126,5	224,5
Тверская область	48,9	101,7	121,7	183,7
Тульская область	32,8	114,9	142,4	219,7
Ярославская область	54,5	106,7	125,9	250,3
г. Москва	86	121,6	122	416,9

После ввода данных можно приступить непосредственно к анализу, выбрав следующий пункт меню: Special/Multivariate Methods/Principal components … . Система отобразит окно диалога для задания переменных. Анализируемые переменные необходимо переместить в поле Data, а наименования объектов – в поле Point Labels. После нажатия кнопки OK система выдаст окно с исходной сводкой результатов анализа по методу главных компонент.

Из полученной сводки можно получить справочную информацию об анализируемых переменных и количестве наблюдений (объектов), подвергающихся анализу. Информация самого метода главных компонент заключена в таблице. В ней представлены характеристики новых компонент: собственные значения главных компонент, упорядоченные по величине (Eigenvalue); процент дисперсии, приходящийся на каждую выделенную главную компоненту (Percent of Variance); накопленный процент дисперсии (Cumulative Percentage).

Полученные для нашего примера цифры говорят о том, что две первые главные компоненты описывают почти 85 % дисперсии исходных данных. Третья

главная компонента добавляет ещё 10 % дисперсии, но в нашем случае предпочтительно ими пожертвовать в целях более лаконичного представления данных. Таким образом, для дальнейшего анализа оставим в рассмотрении две первые главные компоненты.

Для более детального анализа нажмём кнопку табличных опций (вторая слева в верхнем ряду) и в соответствующем окне диалога установим флажок компонентных весов (Components Weights).

Как следует из полученных цифр, в первой главной компоненте наибольшие по величине положительные коэффициенты имеют обеспеченность средним медперсоналом и больничными койками, во второй же компоненте превалирует обеспеченность врачами и учреждениями. В соответствии с этим можно предложить содержательную интерпретацию новым переменным.

На основании сделанных выводов перейдём к рассмотрению диаграммы рассеяния совокупности объектов в двухмерном пространстве (в соответствии с числом главных компонент, оставленных в рассмотрении). Для этого нажмём на кнопку графических отображений и инициализируем соответствующее двухмерное отображение. Данная визуализация позволяет выделить из всей совокупности субъектов ЦФО:

– Москву, характеризующуюся очень большой (по сравнению с другими субъектами) обеспеченностью врачами и учреждениями;

– Ивановскую область, выделяющуюся по самому высокому уровню обеспеченности средним медперсоналом и больничными койками при довольно низкой обеспеченности врачами и учреждениями;

– Московскую область с самой низкой обеспеченностью средним медперсоналом и больничными койками;

– оставшиеся объекты, занимающее среднее положение в общей совокупности.

Известно, что уникальные свойства первой главной компоненты позволяют использовать её в качестве интегрального показателя, с максимальной информативностью характеризующего исследуемую совокупность объектов и используемого в качестве основы для проведения ранжировок. Информацию о значениях главных компонент по каждому из исследуемых объектов можно получить во вкладке Data Table табличных отображений. Однако в нашем случае использование первой главной компоненты в качестве интегрального показателя, заключающего в себе информацию обо всех исходных переменных, вряд ли оправдано, т. к. в ней заключено менее 50 % дисперсии исходных данных.

4.3 Задания

Имеются данные, представляющие собой статистическое обследование множества объектов по ряду характеризующих их признаков. Требуется:

1) на основе критерия информативности метода главных компонент вынести решение о числе главных компонент, оставляемых в рассмотрении для дальнейшего анализа;

2) на основании уравнений для главных компонент дать компонентам со-

держательную интерпретацию;

3) графически представить объекты в пространстве соответствующей мерности;

4) предложить и обосновать метод вычисления интегрального показателя, основанного на знании значений исходных показателей. Произвести ранжировку объектов по данному интегральному показателю;

5) предложить интерпретацию полученных результатов.

4.4 Варианты заданий

Вариант 1

Имеются исходные статистические данные (таблица 4.2) по двадцати сельскохозяйственным районам: число колёсных тракторов на 100 га (X1); число зерноуборочных комбайнов на 100 га (X2); число орудий поверхностной обработки почвы на 100 га (X3); количество удобрений, расходуемых на гектар (X4); количество средств защиты растений, расходуемых на гектар (X5).

Таблица 4.2

Номер с/х района	X1	X2	X3	X4	X5
1	1,59	0,21	2,05	0,32	0,09
2	0,34	0,23	2,78	0,87	0,11
3	2,53	0,3	2,96	0,45	0,21
4	1,32	0,25	3,1	0,34	0,24
5	2,45	0,17	2,75	0,19	0,32
6	3,7	0,33	3,02	0,37	0,27
7	2,21	0,89	2,48	0,76	0,25
8	3,4	0,34	3,46	0,44	0,22
9	1,45	0,8	3,76	0,56	0,14
10	0,23	0,22	2,98	028	0,08
11	2,8	0,23	1,67	0,53	0,16
12	0,56	0,11	3,45	0,56	0,16
13	4	2,01	6,23	0,98	0,07
14	2,33	0,1	2,89	0,75	0,4
15	0,78	0,44	1,99	0,5	0,27
16	1,09	0,19	2,04	0,47	0,21
17	2,43	0,25	3,9	0,37	0,16
18	1,35	0,28	2,76	0,64	0,11
19	0,89	0,16	1,87	0,44	0,18
20	1,2	0,56	3,64	0,68	0,25

Вариант 2

В таблице 4.3 представлены общие затраты на рубль товарной продукции (X1) и фондоотдача (X2) по 10 предприятиям.

Вариант 3

Условия жизни населения 10 стран характеризуются тремя показателями: X1 – оценка ВВП по паритету покупательной способности на душу населения (в процентах к США); X2 – расходы на здравоохранение (в процентах от ВВП);

X3 – численность врачей на 10000 населения (таблица 4.4).

Таблица 4.3

Номер предприятия	X1	X2
1	0,92	0,51
2	0,93	0,59
3	0,83	0,99
4	0,81	1,03
5	0,95	1,21
6	0,88	0,68
7	0,85	0,77
8	0,77	0,59
9	0,78	0,86
10	0,82	1,34

Таблица 4.4

Страна	X1	X2	X3
Россия	20,4	3,2	44,5
Австралия	71,4	8,5	32,5
Австрия	78,7	9,2	33,9
Азербайджан	12,1	3,3	38,8
Армения	10,9	3,2	34,4
Беларусь	20,4	5,4	43,6
Бельгия	79,8	8,9	41
Болгария	17,3	5,4	36,4
Великобритания	69,7	7,1	17,9
Венгрия	24,5	6	32,1

Вариант 4

Уровень цен в 12 городах сравнивался по следующим видам продовольственных товаров: говядина (X1), растительное масло (X2), сахар-песок (X3) и хлеб белый в/с (X4). Статистические данные по ценам (в рублях) представлены в таблице 4.5.

Таблица 4.5

Город	X1	X2	X3	X4
Брянск	12500	7726	3410	4875
Владимир	13857	7880	3183	7125
Иваново	14150	6128	3209	4998
Калуга	12697	8237	3400	5170
Кострома	13000	8750	3600	5476
Москва	14120	11024	4418	6466
Орел	10678	8456	3634	4200
Рязань	12163	9172	4033	4720
Смоленск	12833	8320	3909	4354
Тверь	14400	7083	3416	5440
Тула	12083	8259	3486	5140
Ярославль	14379	7991	3938	5283

Вариант 5

Собраны статистические данные (таблица 4.6) федеральных округов России в 2000 г. по следующим относительным показателям: X1 – оборот розничной торговли на душу населения; X2 – объем платных услуг на душу населения; X3 – объём бытовых услуг на душу населения; X4 – объём инвестиций в основной капитал на душу населения; X5 – денежные доходы на душу населения.

Таблица 4.2

Федеральный округ	X1	X2	X3	X4	X5
Центральный округ	26663	6868	971	8097	3279
Северо-западный округ	14747	4044	461	8019	2170
Южный округ	10130	3244	429	5960	1329
Приволжский округ	11891	1902	554	5985	1624
Уральский округ	13234	2631	426	20848	2558
Сибирский округ	12192	2328	484	4504	1766
Дальневосточный округ	13086	3673	458	7082	2228

4.5 Содержание отчёта

Отчёт должен содержать:

1) данные, на основании которых сделаны выводы о размерности нового признакового пространства и дана содержательная интерпретация главных компонент, оставленных в рассмотрении;

2) графическое представление объектов в пространстве выбранной мерности;

3) таблицу проведённой ранжировки;

4) характеристику предложенной для анализа совокупности объектов на основании полученных результатов.

Вопросы для самоконтроля

1 На основании каких данных сделан вывод о размерности нового признакового пространства? Что обозначают эти данные?

2 Почему было оставлено в рассмотрении n главных компонент, а не $n+1$, ведь большее признаковое пространство даёт большее значение критерия информативности?

3 На основании каких факторов главной компоненте придаётся определённый содержательный смысл? Всегда ли это возможно?

4 О чем свидетельствует отрицательный знак одного из коэффициентов в уравнении для главной компоненты?

5 Чем, по сравнению с другими объектами, характеризуется та или иная группа объектов на новом признаковом пространстве?

6 Чему соответствует первенство объекта в списке ранжировки?

5 Лабораторная работа № 5. Кластерный анализ. Иерархические методы

Цель работы: получение навыков практического применения иерархических и итеративных методов кластерного анализа.

5.1 Введение

Термин «кластерный анализ», впервые введённый Трионом (Tryon) в 1939 г., включает в себя более 100 различных алгоритмов.

В отличие от задач классификации, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Кластерный анализ позволяет сокращать размерность данных, делать её наглядной.

Кластерный анализ может применяться к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

Кластерный анализ параллельно развивался в нескольких направлениях, таких как биология, психология и других, поэтому у большинства методов существует по два названия и более.

Задачи кластерного анализа можно объединить в следующие группы:

1) разработка типологии или классификации;

2) исследование полезных концептуальных схем группирования объектов;

3) представление гипотез на основе исследования данных;

4) проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Как правило, при практическом использовании кластерного анализа одновременно решается несколько из указанных задач.

5.2 Указания к выполнению

Название «кластерный анализ» происходит от английского слова cluster – гроздь, скопление. Кластерный анализ – широкий класс процедур многомерного статистического анализа, позволяющих произвести автоматизированную группировку наблюдений в однородные классы – кластеры.

Кластер имеет следующие математические характеристики:

- центр;
- радиус;
- дисперсия кластера;
- среднеквадратическое отклонение.

Центр кластера – это среднее геометрическое место точек в пространстве переменных.

Радиус кластера – максимальное расстояние точек от центра кластера.

Дисперсия кластера – это мера рассеяния точек в пространстве относительно центра кластера.

Среднеквадратичное отклонение (СКО) объектов относительно центра кластера – квадратный корень из дисперсии кластера.

Методы кластерного анализа.

Методы кластерного анализа можно разделить на две группы:

1) иерархические;

2) неиерархические.

Каждая группа включает множество подходов и алгоритмов.

Используя различные методы кластерного анализа, аналитик может получить различные решения на одних и тех же данных. Это считается нормальным явлением.

Иерархические методы кластерного анализа.

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES).

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (*делимые*) *методы* (DIvisive ANAlysis, DIANA).

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о «схожести» объектов при их объединении в группу.

Меры сходства.

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний.

Евклидово расстояние является геометрическим расстоянием в многомерном пространстве и вычисляется по формуле

$$R(x,y)=\sqrt{\sum_i (x_i-y_i)^2}\ .$$

Евклидово расстояние (и его квадрат) вычисляется по исходным, а не по стандартизованным данным.

Манхэттенское расстояние (расстояние городских кварталов), также называемое «хэмминговым» или «сити-блок» расстоянием, рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным расчётам евклидова расстояния. Однако для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат. Манхэттенское расстояние вычисляется по формуле

$$R(x,y)=\sum_i |x_i - y_i|.$$

Расстояние Чебышёва стоит использовать, когда необходимо определить два объекта как «различные», если они отличаются по какому-то одному измерению. Расстояние Чебышёва вычисляется по формуле

$$R(x,y)=\max|x_i - y_i|.$$

Степенное расстояние используется в тех случаях, когда желают прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по формуле

$$R(x,y)=\left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{r}},$$

где r и p – параметры, определяемые пользователем.

Параметр p отвечает за постепенное взвешивание разностей по отдельным координатам, параметр r – за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра r и p равны двум, то это расстояние совпадает с расстоянием Евклида.

Процент несогласия используется в тех случаях, когда данные являются категориальными. Это расстояние вычисляется по формуле

$$R(x,y)=\frac{\text{Количество } x_i \neq y_i}{n}.$$

Методы объединения или связи.

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Однако, когда связываются вместе несколько объектов, необходимо использовать другие методы определения расстояния между кластерами. Существует множество методов объединения кластеров:

1) одиночная связь (метод ближайшего соседа) – расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах;

2) полная связь (метод наиболее удалённых соседей) – расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. «наиболее удалёнными соседями»);

3) невзвешенное попарное среднее – расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них;

4) взвешенное попарное среднее – метод идентичен методу *невзвешенного попарного среднего*, за исключением того, что при вычислениях размер соответствующих кластеров (т. е. число объектов, содержащихся в них) используется в качестве весового коэффициента;

5) невзвешенный центроидный метод – расстояние между двумя кластерами определяется как расстояние между их центрами тяжести;

6) взвешенный центроидный метод (медиана) – метод идентичен невзвешенному центроидному методу, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т. е. числами объектов в них);

7) метод Варда – расстояние между кластерами определяется как прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. Метод отличается от всех других методов, поскольку он использует методы дисперсионного анализа для оценки расстояний между кластерами. Метод минимизирует сумму квадратов для любых двух (гипотетических) кластеров, которые могут быть сформированы на каждом шаге.

Метод ближайшего соседа.

Расстояние между двумя классами определяется как расстояние между ближайшими их представителями.

Перед началом работы алгоритма рассчитывается *матрица расстояний* между объектами. Согласно критерию классификации объединение происходит между кластерами, расстояние между ближайших представителей которых наименьшее: выбираются два объекта с наименьшим расстоянием в один кластер. После этого необходимо произвести перерасчёт матрицы расстояний с учётом нового кластера. На каждом шаге в матрице расстояний ищется минимальное значение, соответствующее расстоянию между двумя наиболее близкими кластерами. Найденные кластеры объединяются, образуя новый кластер. Эта процедура повторяется до тех пор, пока не будут объединены все кластеры.

При использовании метода ближайшего соседа особое внимание следует уделять выбору меры расстояния между объектами. На основе неё формируется начальная матрица расстояний, которая и определяет весь дальнейший процесс классификации.

5.3 Задания

1 Разработать алгоритмы методов ближнего соседа и k-средних и реализовать их в виде компьютерных программ. С помощью генератора случайных чисел (ГСЧ) сгенерировать 50 реализаций $x = (x_1, x_2)$ – случайной 2-мерной величины, координаты которой распределены равномерно в интервале (3;8). Распределить их с помощью разработанных программ на минимальное число кластеров, каждый из которых помещается в сфере радиуса 0,15.

2 Имеются данные, представляющие собой статистическое обследование множества объектов по ряду характеризующих их признаков.

Требуется:

1) с помощью иерархического агломеративного алгоритма провести классификацию объектов при использовании обычной евклидовой метрики – методом:

а) «ближайшего соседа»;

б) «дальнего соседа»;

в) «центра тяжести»;

г) «средней связи»;

2) на основании анализа полученных дендрограмм и диаграмм рассеяния для каждого алгоритма выбрать предпочтительное разбиение объектов на кластеры;

3) используя наиболее устойчивое разбиение из всех четырёх вариантов, а также априорные представления об исследуемой совокупности, вынести окончательное решение о разбиении объектов на классы;

4) по результатам кластер-анализа дать характеристику каждому сформированному классу.

5.4 Варианты заданий

Вариант 1

Уровень жизни населения двадцати стран за 1994 г. характеризуется следующими показателями: X1 – потребление мяса и мясопродуктов на душу населения, кг; X2 – смертность населения по причине болезни органов кровообращения на 100000 населения; X3 – оценка ВВП по паритету покупательной способности на душу населения (в процентах к США); X4 – расходы на здравоохранение (в процентах от ВВП); X5 – потребление фруктов и ягод на душу населения, кг; X6 – потребление хлебопродуктов на душу населения, кг (таблица 5.1).

Таблица 5.1

Страна	X1	X2	X3	X4	X5	X6
Россия	55	84,98	20,4	3,2	28	124
Австралия	100	30,58	71,4	8,5	121	87
Австрия	93	38,42	78,7	9,2	146	74
Азербайджан	20	60,34	12,1	3,3	52	141
Армения	20	60,22	10,9	3,2	72	134
Беларусь	72	60,7	20,4	5,4	38	120
Бельгия	85	29,82	79,7	8,3	83	72
Болгария	85	70,57	17,3	5,4	92	156
Великобритания	67	34,51	69,7	7,1	91	91
Венгрия	73	64,73	24,5	6	73	106
Германия	88	36,63	76,2	8,6	138	73
Греция	83	32,84	44,44	5,7	99	108
Грузия	21	62,64	11,3	3,5	55	140
Дания	98	34,07	79,2	6,7	89	77
Ирландия	99	39,27	57	6,7	87	102
Испания	89	28,46	54,8	7,3	103	72
Италия	84	30,27	72,1	8,5	169	118
Казахстан	61	69,04	13,4	3,3	10	191
Канада	98	25,42	79,9	10,2	123	77
Киргизия	46	53,13	11,2	3,4	20	134

Вариант 2

Рост промышленного производства в Ивановской области в апреле 2002 г. по основным отраслям промышленности характеризуется следующими показателями: X1 – индекс физического объёма выпуска продукции к предыдущему месяцу, %; X2 – индекс физического объёма выпуска продукции к соответствующему месяцу прошлого года, %; X3 – индекс физического объёма выпуска про-

дукции к соответствующему периоду прошлого года, % (таблица 5.2).

Таблица 5.2

Отрасль промышленности	X1	X2	X3
Электроэнергетика	58,9	111,9	99,2
Топливная	105,1	99	86
Черная металлургия	85,6	45	89,7
Химическая и нефтехимическая (без химико-фармацевтической)	86,9	79,1	78,5
Отрасль промышленности	X1	X2	X3
Машиностроение и металлообработка	57	118	92
Лесная, деревообрабатывающая и целлюлозно-бумажная	80,2	69,7	76,9
Стройматериалов	106,5	100,6	97
Лёгкая	103,4	130,7	119,1
Пищевая	105,7	125,8	114,3
Мукомольно-крупяная и комбикормовая	100	78	68,8
Медицинская	101,1	93,7	94,1
Полиграфическая	152,6	179	103,8

Вариант 3

Уровень медицинского обслуживания субъектов РФ центрального федерального округа характеризуется показателями: X1 – обеспеченность населения врачами, человек на 10000 населения; X2 – обеспеченность средним медицинским персоналом, человек на 10000 населения; X3 – обеспеченность больничными койками, число на 10000 населения; X4 – обеспеченность врачебными амбулаторно-поликлиническими учреждениями, число посещений в смену на 10000 населения (таблица 5.3).

Таблица 5.3

Субъект ЦФО	X1	X2	X3	X4
Белгородская область	38,2	117,4	125	199,8
Брянская область	36,6	109,6	123,1	202,9
Владимирская область	36,2	104,3	121	293,9
Воронежская область	49,3	113,3	118,9	192
Ивановская область	51,3	138,8	142	188,7
Калужская область	39,5	104,1	114,4	242,7
Костромская область	35,9	120,1	145,3	187
Курская область	45,5	107	115,3	206
Липецкая область	39,2	123,3	130,7	291,4
Московская область	33,9	87,9	106,7	246,7
Орловская область	36,7	112,2	118,9	180
Рязанская область	51,7	120,3	131,7	215,3
Смоленская область	57,6	109	130,6	230,9
Тамбовская область	33	104,1	126,5	224,5
Тверская область	48,9	101,7	121,7	183,7
Тульская область	32,8	114,9	142,4	219,7
Ярославская область	54,5	106,7	125,9	250,3
г. Москва	86	121,6	122	416,9

Вариант 4

Анализируется движение населения за 2000 г. в субъектах РФ центрального федерального округа по следующим показателям: X1 – естественный прирост населения, на 1000 населения; X2 – миграционный прирост населения, на 1000 населения (таблица 5.4).

Таблица 5.4

Субъект ЦФО	X1	X2
Белгородская область	-7,4	100
Брянская область	-10,2	7
Владимирская область	-11,3	21
Воронежская область	-10,2	29
Ивановская область	-13,2	22
Калужская область	-10,6	16
Костромская область	-10,4	24
Курская область	-10,1	5
Липецкая область	-8,6	47
Московская область	-10	56
Орловская область	-9,7	27
Рязанская область	-12,1	9
Смоленская область	-12,5	-4
Тамбовская область	-10,7	-3
Тверская область	-13,5	11
Тульская область	-14,2	2
Ярославская область	-10,7	25
г. Москва	-6,7	78

5.5 Содержание отчёта

Отчёт должен содержать:

1) варианты разбиения по каждому из предложенных алгоритмов;

2) дендрограмму работы одного из алгоритмов, приводящего к выбранному варианту разбиения, с указанием расстояний, на которых происходит объединение кластеров;

3) принадлежность объектов к кластерам, центроиды.

Вопросы для самоконтроля

1 Каковы основные положения любого критерия качества разбиения?

2 На сколько информативно представление результатов кластерного анализа на диаграмме рассеяния?

3 Чем отличается один метод от другого? Одинаковы ли дендрограммы для разных методов, дающих одинаковое разбиение?

4 Охарактеризуйте тот или иной получившийся кластер.

5 Опишите в самом общем виде работу любого алгоритма кластерного анализа, иерархических процедур.

6 Как можно придать дополнительную классифицирующую силу той или иной исходной переменной?

7 На основании каких параметров сделан вывод об окончательном варианте разбиения?

8 Опишите работу иерархического алгоритма на конкретном примере по шагам.

6 Лабораторная работа № 6. Неиерархические методы кластерного анализа

Цель работы: получение навыков практического применения неиерархических методов кластерного анализа.

6.1 Указания к выполнению

Итеративные методы.

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении исходных кластеров на другие кластеры, и которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определённое количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т. е. определение кластера там, где имеется большое «сгущение точек». Второй подход заключается в минимизации меры различия объектов.

В отличие от иерархических методов классификации итеративные методы могут привести к образованию пересекающихся кластеров, когда один объект может одновременно принадлежать нескольким кластерам.

К итеративным методам относятся, например, метод *k*-средних, метод поиска сгущений и др. Итеративные методы относятся к быстродействующим, что позволяет использовать их для обработки больших массивов исходной информации.

Алгоритм k-средних (*k-means*).

Среди итеративных методов наиболее популярным методом является метод *k*-средних Мак-Кина. В отличие от иерархических методов в большинстве реализаций этого метода сам пользователь должен задать искомое число конечных кластеров, которое обычно обозначается «*k*». Алгоритм *k*-средних строит *k* кластеров, расположенных на возможно больших расстояниях друг от друга. Основное требование к типу задач, которые решает алгоритм *k*-средних, – наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа *k* может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Как и в иерархических методах кластеризации, пользователь при этом может выбрать тот или иной тип меры сходства. Разные алгоритмы метода *k*-средних отличаются и способом выбора начальных центров задаваемых кластеров.

В некоторых вариантах метода сам пользователь может (или должен) задать такие начальные точки, либо выбрав их из реальных наблюдений, либо задав координаты этих точек по каждой из переменных. В других реализациях этого метода выбор заданного числа k начальных точек производится случайным образом, причем эти начальные точки (центры кластеров) могут в последующем уточняться в несколько этапов. Можно выделить четыре основных этапа таких методов:

1) выбираются или назначаются k наблюдений, которые будут первичными центрами кластеров;

2) при необходимости формируются промежуточные кластеры приписыванием каждого наблюдения к ближайшим заданным кластерным центрам;

3) после назначения всех наблюдений отдельным кластерам производится замена первичных кластерных центров на кластерные средние;

4) предыдущая итерация повторяется до тех пор, пока изменения координат кластерных центров не станут минимальными.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Описание алгоритма.

1 Первоначальное распределение объектов по кластерам.

2 Выбирается число k и k точек. На первом шаге эти точки считаются «центрами» кластеров. Каждому кластеру соответствует один центр. Выбор начальных центроидов может осуществляться следующим образом:

- выбор k-наблюдений для максимизации начального расстояния;
- случайный выбор k-наблюдений;
- выбор первых k-наблюдений.

3 Затем каждый объект назначается определённому наиболее близкому кластеру.

Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

1) кластерные центры стабилизировались, т. е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации. В некоторых вариантах этого метода пользователь может задать числовое значение критерия, трактуемого как минимальное расстояние для отбора новых центров кластеров. Наблюдение не будет рассматриваться как претендент на новый центр кластера, если его расстояние до заменяемого центра кластера превышает заданное число. Такой параметр в ряде программ называется «радиусом». Кроме этого параметра, возможно задание обычно достаточно малого числа, с которым сравнивается изменение расстояния для всех кластерных центров. Этот параметр обычно называется «конвергенцией», т. к. отражает сходимость итерационного процесса кластеризации;

2) число итераций равно максимальному числу итераций.

Проверка качества кластеризации.

После получений результатов кластерного анализа методом *k*-средних следует проверить правильность кластеризации (т. е. оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Достоинства алгоритма *k*-средних:

1) простота использования;

2) быстрота использования;

3) понятность и прозрачность алгоритма.

Недостатки алгоритма *k*-средних:

1) алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма – алгоритм *k*-медианы;

2) алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

6.2 Содержание отчёта

Отчёт должен содержать:

1) описание и блок-схемы алгоритмов;

2) исходные тексты программных модулей;

3) результаты работы алгоритмов в виде графиков.

7 Лабораторная работа № 7. Компонентный анализ

Цель работы: получение навыков практического применения методов компонентного анализа.

7.1 Введение

Компонентный анализ является методом определения структурной зависимости между случайными переменными. В результате его использования получается сжатое описание малого объёма, несущее почти всю информацию, содержащуюся в исходных данных. Главные компоненты получаются из исходных переменных путём целенаправленного вращения, т. е. как линейные комбинации исходных переменных. Вращение производится таким образом, чтобы главные компоненты были ортогональны и имели максимальную дисперсию среди возможных линейных комбинаций исходных переменных *X*. При этом переменные не коррелированы между собой и упорядочены по убыванию дисперсии (первая компонента имеет наибольшую дисперсию). Кроме того, общая дисперсия после преобразования остаётся без изменений.

7.2 Задание

1 В файле stocks.sta содержатся еженедельные доходности акций пяти американских компаний Allied Chemical, Du Pont, Union Carbide, Exxon и Texaco на

Нью-Йоркской фондовой бирже (всего 100 наблюдений) с января 1975 г. по декабрь 1976 г. Требуется исследовать возможность объяснения динамики доходности акций на фондовом рынке менее чем пятью признаками (акциями).

2 Выполните расчёт описательных статистик по переменным выборки. Сделайте выводы. Рассчитайте корреляционную матрицу переменных. Почему коэффициенты корреляции для доходностей акций ненулевые?

3 Анализ главных компонент выполняется в программе Statistica с помощью модуля Statistics \Multivariate Exploratory Techniques\Principal Components & Classification Analysis.

В появившемся окне Principal Components and Classification Analysis укажите переменные классификации (укажите все имеющиеся переменные, кроме номера недели – Week) и нажмите OK (рисунок 7.1).

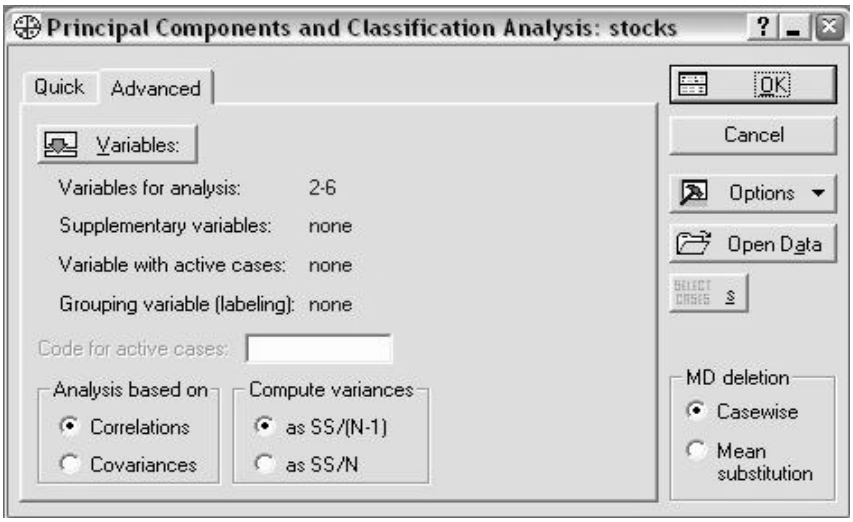


Рисунок 7.1

В верхней части появившегося окна результатов компонентного анализа даны общая информация и рассчитанные собственные числа корреляционной матрицы переменных (рисунок 7.2).

Выберите число факторов (Number of factors) и получите соответствующий процент объяснённой этими факторами вариации (Quality of representation). Просмотрите коэффициенты факторных нагрузок Factor& variable correlations. Далее можно получить график Screeplot собственных чисел и сами собственные числа Eigenvalues, соответствующие им собственные векторы Eigenvectors (рисунок 7.3). Дайте интерпретацию факторам по значениям полученных собственных векторов.

Рекомендуется далее самостоятельно ознакомиться с другими опциями в представлении результатов компонентного анализа.

4 Для имеющихся данных определите и обоснуйте оптимальное число факторов и дайте их интерпретацию. Какой процент вариации они объясняют?

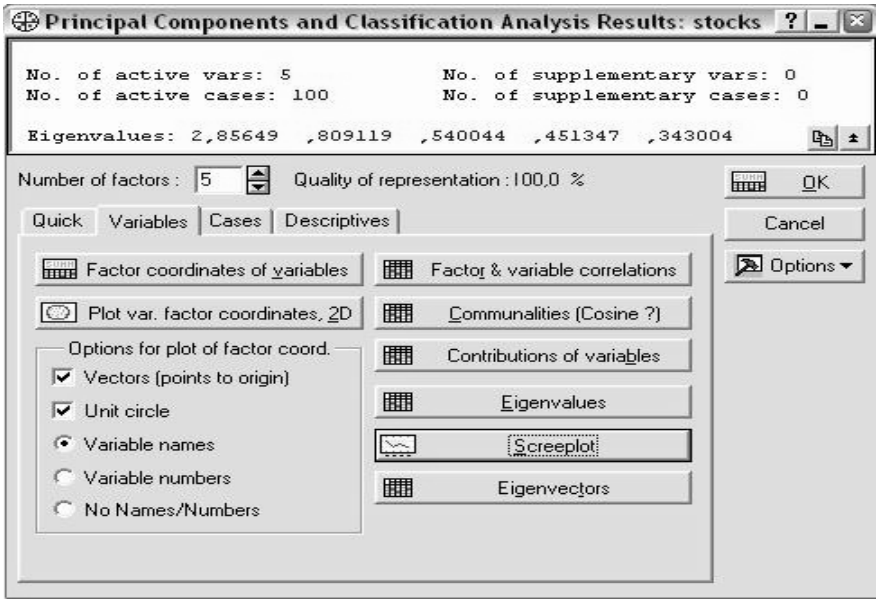


Рисунок 7.2

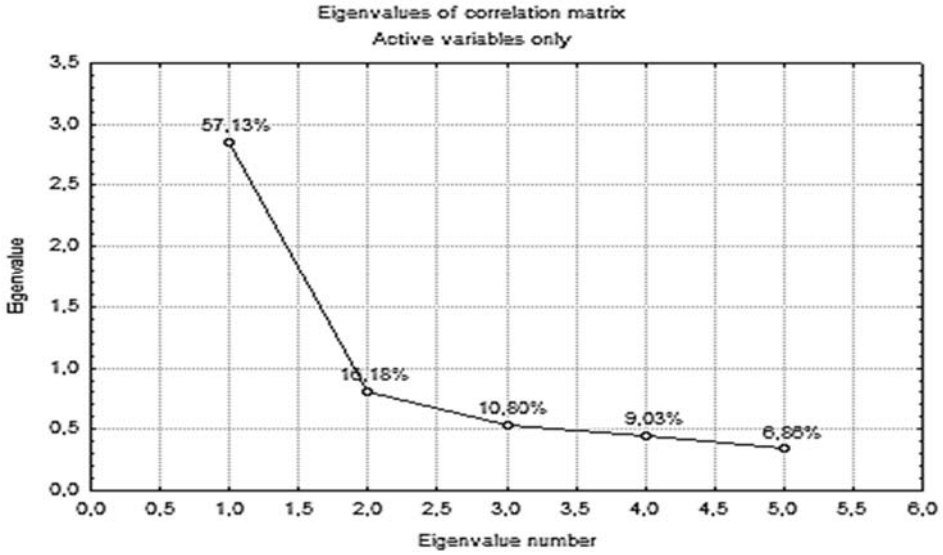


Рисунок 7.3

7.3 Содержание отчёта

Отчёт должен представлять собой содержательные выводы по результатам всей выполненной работы.

8 Лабораторная работа № 8. Факторный анализ

Цель работы: получение навыков практического применения методов факторного анализа.

8.1 Введение

Факторный анализ – это совокупность методов, которые на основе реально существующих связей объектов (признаков) позволяют выявить латентные (неявные) обобщающие характеристики организационной структуры. При этом

предполагается, что наблюдаемые переменные являются линейной комбинацией факторов. Под *фактором* понимается гипотетическая непосредственно не измеряемая, скрытая (латентная) переменная, в той или иной мере связанная с исходными наблюдаемыми переменными. К факторному анализу относятся: метод главных компонент; методы многомерного шкалирования, применяемые для формирования факторного пространства по информации о близости объектов; методы кластерного анализа, применяемые для описания неколичественных факторов.

Основные цели факторного анализа:

1) сокращение числа переменных (редукция данных);

2) определение структуры взаимосвязей между переменными (классификация переменных);

3) косвенные оценки признаков, неподдающихся непосредственному измерению;

4) преобразование исходных переменных к более удобному для интерпретации виду.

Если кратко охарактеризовать факторный анализ, то наиболее важными являются следующие моменты:

1) факторный анализ, в противоположность контролируемому эксперименту, опирается в основном на наблюдения над естественным варьированием переменных;

2) при использовании факторного анализа совокупность переменных, изучаемых с точки зрения связей между ними, не выбирается произвольно: сам метод позволит выявить основные факторы, оказывающие существенное влияние в данной области;

3) факторный анализ не требует предварительных гипотез, наоборот, он сам может служить методом выдвижения гипотез, а также выступать критерием гипотез, опирающихся на данные, полученные другими методами;

4) факторный анализ не требует априорных предположений относительно того, какие переменные независимы, а какие зависимы, метод не преувеличивает причинно-следственные связи и решает вопрос об их мере в процессе дальнейших исследований.

Метод факторного анализа первоначально был разработан в психологии с целью выделения отдельных компонентов человеческого интеллекта из многомерных данных по измерению различных проявлений умственных способностей. Однако очень быстро этот метод завоевал и такие области применения, как социология, экономика, география и многие другие.

Переменные, значения которых можно измерить, имеют для исследуемого объекта нередко достаточно условный характер, лишь опосредованно отражая его внутреннюю структуру, движущие механизмы или факторы. Например, исследователь ставит цель: провести сравнительный анализ темпов экономического роста отдельных регионов (соответствующий пример будет в дальнейшем рассмотрен). Закономерен вопрос: чем измерить экономическое развитие и какие показатели следует включить в исследование?

Когда неизвестный фактор проявляется в изменении нескольких переменных, в процессе анализа можно наблюдать существенную корреляцию между пе-

ременными. Тем самым факторов может быть существенно меньше, чем измеряемых переменных, число которых выбирается исследователем достаточно субъективно.

Степень влияния фактора на некоторый показатель (переменную) статистически характеризуется величиной дисперсии этого показателя при изменении значений фактора. Если расположить оси исходных переменных ортогонально друг к другу, то можно обнаружить, что в этом многомерном пространстве объекты группируются в виде эллипса рассеяния, более вытянутого в одних направлениях и почти плоского в других. Если теперь провести новые оси соответственно осям эллипса рассеяния, то можно говорить о выделении скрытых факторов и оценивать сравнительную значимость этих факторов в терминах дисперсии. При этом оказывается, что толщина такого эллипса по некоторым осям настолько не велика, что можно исключить их из исследования.

8.2 Указания к выполнению

Как правило, применение методов факторного анализа включает три этапа:

1) выделение первоначальных факторов;

2) вращение выделенных факторов с целью облегчения их интерпретации в терминах исходных переменных (в частности, для исключения отрицательных значений);

3) содержательная интерпретация новых факторов в предметных терминах, что является творческой задачей исследователя, выходящей за рамки предлагаемого формального метода.

Наиболее часто факторный анализ используется для выявления в наблюдаемых признаках $x_1,...,x_k$ некоторых латентных (скрытых) переменных f_m, называемых *факторами.* Гипотеза о наличии этих факторов основана на предположении о существовании чего-то общего в наблюдаемых признаках. Выводимые гипотетические факторы обладают следующими свойствами:

1) они образуют линейно независимый набор переменных, т. е. ни один из факторов (компонент) не выводится как линейная комбинация остальных;

2) переменные, являющиеся гипотетическими факторами, можно разделить на два основных вида – общие и характерные факторы. Они отличаются структурой весов в линейном уравнении, которое выводит значение наблюдаемой переменной из гипотетических факторов. Общий фактор имеет несколько переменных с ненулевым весом или факторной нагрузкой, соответствующей этому фактору. При этом фактор называется *общим*, если хотя бы две его нагрузки значительно отличаются от нуля. Характерный фактор имеет только одну переменную с ненулевым весом (т. е. только одна переменная от него зависит);

3) всегда предполагается, что общие факторы не коррелируют с характерным фактором, также характерные факторы не коррелированы между собой;

4) обычно предполагается, что число общих факторов меньше, чем число наблюдаемых переменных, однако число характерных факторов принимают равным числу наблюдаемых переменных.

8.3 Задание

1 Рабочий файл данных тот же, что и в предыдущей лабораторной работе: stocks.sta.

2 Выполним факторный анализ по имеющимся выборочным данным. В программе Statistica запустите модуль Statistics\ Multivariate Exploratory Techniques\ Factor Analysis (рисунок 8.1) и в появившемся окне укажите переменные для анализа (укажите все имеющиеся переменные, кроме номера недели – Week) и нажмите OK. Далее во вкладке Advanced укажите метод поиска латентных факторов: Principal components – метод главных факторов, а также максимальное количество факторов, например 2, и минимальное собственное число фактора для включения его в анализ – 0. Нажмите OK.

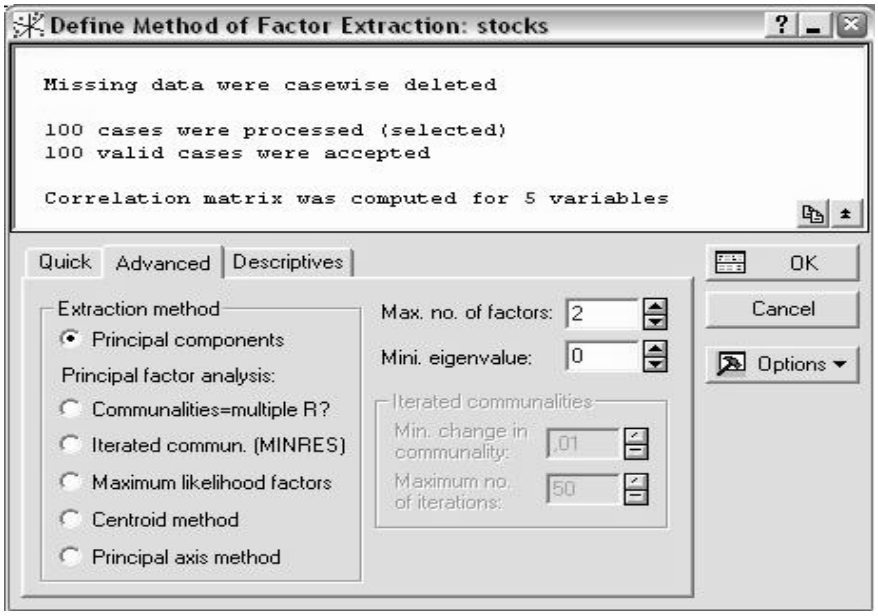


Рисунок 8.1

В появившемся окне результатов (рисунок 8.2) после нажатия кнопки Summary на вкладке Quick получим факторные нагрузки, причём красным цветом выделены коэффициенты, большие по модулю 0,7. Также можно получить (опция Eigenvalues) значения собственных чисел и долю объяснённой факторами вариации. Воспользовавшись вкладкой Explained variance и опцией Communalities, получите значения общностей для каждого из фактора. С помощью вкладки Scores и кнопки Factor scores получим значения факторов для каждого наблюдения (всего 100 наблюдений).

3 Осуществим вращение факторов. Для этого выберите в меню Factor rotation один из методов вращения: варимакс, биквартимакс, квартимакс, эквимакс, например, Varimax normalized. Посмотрите, как изменились факторные нагрузки и другие результаты факторного анализа.

Дайте интерпретацию факторов.

Поэкспериментируйте, выбирая различные значения числа факторов на начальной стадии анализа и различные методы вращения факторов. Выберите оптимальный, с вашей точки зрения, результат.

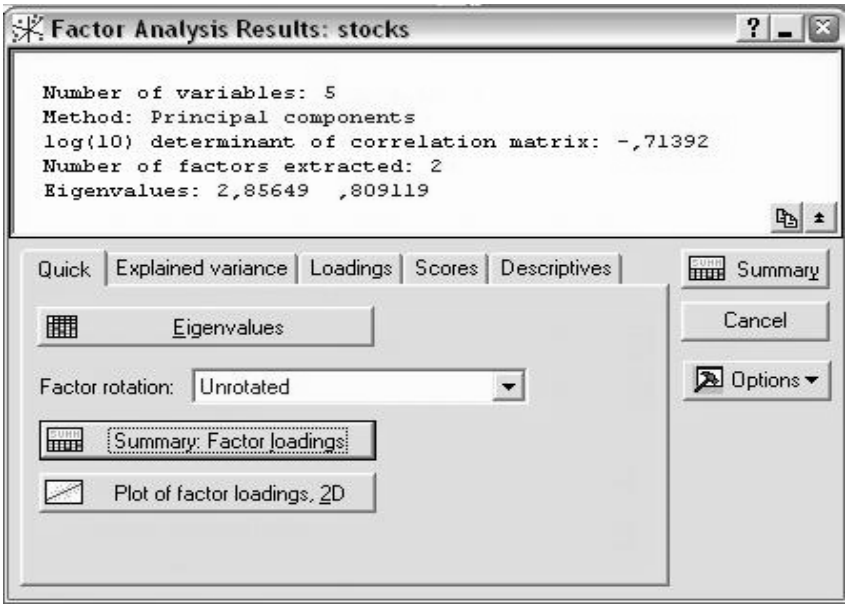


Рисунок 8.2

4 Выполните кластерный анализ наблюдений (по неделям) по двум факторам, полученным после варимакс вращения в предыдущем пункте. Для этого выполните следующие действия.

В окне исходных данных (рисунок 8.3), щёлкнув правой кнопкой мыши по заголовку любой переменной, выберите Add variables и добавьте две новые переменные в конец списка переменных.

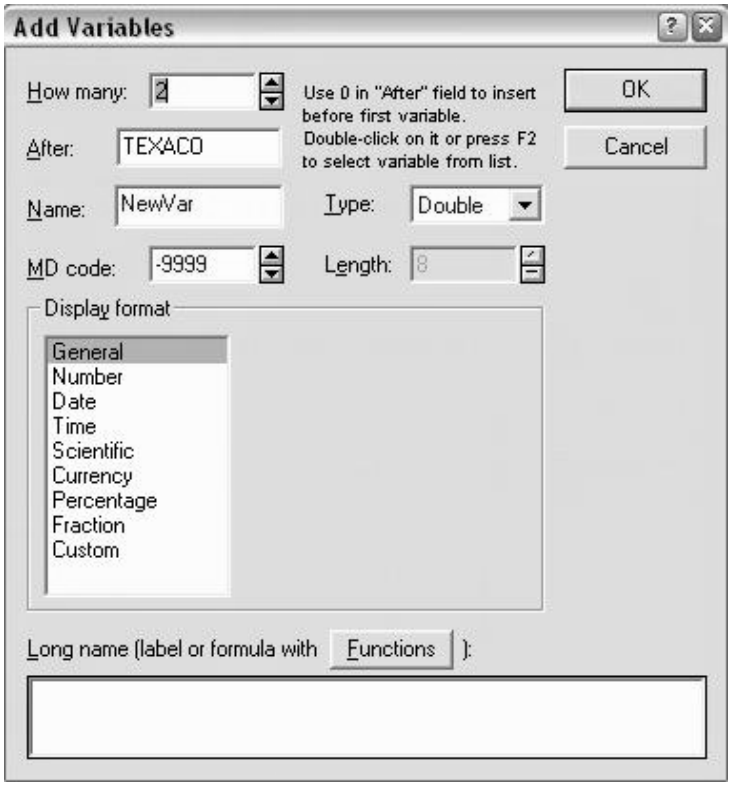


Рисунок 8.3

Вернувшись в окно факторного анализа с полученными, как в п. 4, двумя факторами, с помощью вкладки Scores и кнопки Factor scores получите таблицу

значений факторов для каждого наблюдения (всего 100 наблюдений). Скопируйте два столбца значений таблицы в окно исходных данных на место добавленных двух новых переменных (с именами по умолчанию NewVar1 и NewVar2). Далее воспользуйтесь меню Statistics\ Multivariate Exploratory Techniques\ Cluster Analysis и методом автоматической классификации *k*-средних (см. лабораторную работу № 6) для двух вновь полученных переменных.

5 Сделайте содержательные экономические выводы по результатам факторного и кластерного анализа.

8.4 Содержание отчёта

Отчёт должен содержать экономические выводы по результатам факторного и кластерного анализа.

9 Лабораторная работа № 9. Многомерный статистический анализ

Цель работы: выполнение контрольного задания (обобщение методов многомерного статистического анализа) по указанию преподавателя.

Содержание отчёта.

Отчёт должен содержать:

1) описательные статистики исходных данных и их краткий анализ;

2) расчёт парных и множественных коэффициентов корреляции для переменных выборки;

3) выполнение многомерного статистического анализа согласно заданию;

4) экономическую интерпретацию результатов и выводы.

10 Лабораторная работа № 10. Анализ данных с помощью технологии Data Mining

Цель работы: усвоение основ интеллектуального анализа данных и применение наивного байесовского классификатора для задачи категоризации текстовых документов.

10.1 Введение

Интеллектуальный анализ данных (Data Mining) – это процесс обнаружения в множестве данных ранее неизвестных, полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Английский термин Data Mining не имеет однозначного перевода на русский язык (интеллектуальный анализ данных, добыча данных, вскрытие данных, информационная проходка, извлечение данных/информации) поэтому в большинстве случаев используется в оригинале.

Методы Data Mining разделяются на две группы:

1) статистические (дескриптивный анализ, корреляционный и регрессион-

ный анализ, факторный анализ, дисперсионный анализ, компонентный анализ, дискриминантный анализ, анализ временных рядов);

2) кибернетические (искусственные нейронные сети, байесовские сети, эволюционное программирование, генетические алгоритмы, ассоциативная память, нечёткая логика, деревья решений, системы обработки экспертных знаний).

Визуальные инструменты Data Mining позволяют проводить анализ данных предметными специалистами (аналитиками), не владеющими соответствующими математическими знаниями.

10.2 Указания к выполнению

Основные задачи, решаемые с помощью интеллектуального анализа данных (ИАД):

1) классификация – отнесение входного вектора (объекта, события, наблюдения) к одному из заранее известных классов;

2) кластеризация – разделение множества входных векторов на группы (кластеры) по степени «похожести» друг на друга;

3) сокращение описания – для визуализации данных, лаконизма моделей, упрощения счета и интерпретации, сжатия объёмов собираемой и хранимой информации;

4) ассоциация – поиск повторяющихся образцов. Например, поиск «устойчивых связей в корзине покупателя» (*market basket analysis*) – вместе с пивом часто покупают орешки.

5) прогнозирование – предсказание следующего состояния системы по наблюдаемым;

6) анализ отклонений – выявление нетипичной сетевой активности позволяет обнаружить вредоносные программы.

В литературе можно встретить ещё ряд классов задач. Базовыми задачами являются первые три. Остальные задачи сводятся к ним тем или иным способом.

Алгоритмы решения задач ИАД.

1 Для задач классификации характерно «обучение с учителем», при котором построение (обучение) модели производится по выборке, содержащей входные и выходные векторы.

2 Для задач кластеризации и ассоциации применяется «обучение без учителя», при котором построение модели производится по выборке, в которой нет выходного параметра. Значение выходного параметра («относится к кластеру…», «похож на вектор…») подбирается автоматически в процессе обучения.

3 Для задач сокращения описания характерно *отсутствие разделения на входные и выходные векторы*. Начиная с классических работ К. Пирсона по методу главных компонент, основное внимание здесь уделяется аппроксимации данных.

Этапы обучения.

Можно выделить типичный ряд этапов решения задач методами ИАД:

1) формирование гипотезы;

2) сбор данных;

3) подготовка данных (фильтрация);

4) выбор модели;

5) подбор параметров модели и алгоритма обучения;

6) обучение модели (автоматический поиск остальных параметров модели);

7) анализ качества обучения, если неудовлетворительный переход на п. 5 или п. 4;

8) анализ выявленных закономерностей, если неудовлетворительный переход на п. 1, 4 или 5.

Далее приведено решение задачи классификации документов на основе наивной байесовской модели (naive Bayesian model).

Классификация документов – одна из задач информационного поиска, заключающаяся в отнесении документа к одной из нескольких категорий на основании содержания документа. Классификация может осуществляться полностью вручную, либо автоматически с помощью созданного вручную набора правил, либо автоматически с применением методов машинного обучения. Следует отличать классификацию текстов от кластеризации, в последнем случае тексты также группируются по некоторым критериям, но заранее заданные категории отсутствуют.

Области применения задачи классификации текстов:

- фильтрация спама;
- составление интернет-каталогов;
- подбор контекстной рекламы;
- в системах документооборота;
- автоматическое реферирование (составление аннотаций);
- снятие неоднозначности при автоматическом переводе текстов;
- ограничение области поиска в поисковых системах;
- определение кодировки и языка текста.

Выделяют три подхода к задаче классификации текстов.

Во-первых, классификация не всегда осуществляется с помощью компьютера. Например, в обычной библиотеке тематические рубрики присваиваются книгам вручную библиотекарем. Подобная *ручная классификация* дорога и неприменима в случаях, когда необходимо классифицировать большое количество документов с высокой скоростью.

Другой подход заключается в *написании правил*, по которым можно отнести текст к той или иной категории. Например, одно из таких правил может выглядеть следующим образом: «если текст содержит слова производная и уравнение, то отнести его к категории математика». Специалист, знакомый с предметной областью и обладающий навыком написания регулярных выражений, может составить ряд правил, которые затем автоматически применяются к поступающим документам для их классификации. Этот подход лучше предыдущего, поскольку процесс классификации автоматизируется и, следовательно, количество обрабатываемых документов практически не ограничено. Более того, построение правил вручную может дать лучшую точность классификации, чем при машинном. Однако создание и поддержание правил в актуальном состоянии требует постоянных усилий специалиста.

Наконец, третий подход основывается на *машинном обучении*. В этом подходе набор правил критерий принятия решения текстового классификатора вычисляется автоматически из обучающих данных (другими словами, производится обучение классификатора). Обучающие данные – это некоторое количество хороших образцов документов из каждого класса. В машинном обучении сохраняется необходимость ручной разметки (термин *разметка* означает процесс приписывания класса документу). Но разметка является более простой задачей, чем написание правил. Кроме того, разметка может быть произведена в обычном режиме использования системы. Например, в программе электронной почты может существовать возможность помечать письма как спам, тем самым формируя обучающее множество для классификатора – фильтра нежелательных сообщений. Таким образом, классификация текстов, основанная на машинном обучении, является примером обучения с учителем, где в роли учителя выступает человек, задающий набор классов и размечающий обучающее множество.

Постановка задачи классификации текстов.

1 Имеется множество категорий (классов, меток) $C = \left\{c_1,...,c_{|C|}\right\}$.

2 Имеется множество документов $D = \left\{d_1,...,d_{|D|}\right\}$.

3 Неизвестная целевая функция $F : C \times D \to \left\{0,1\right\}$.

4 Необходимо построить классификатор F^*, максимально близкий к F.

5 Имеется некоторая начальная коллекция размеченных документов $R \subset C \times D$, для которых известны значения F. Обычно её делят на «обучающую» и «проверочную» части. Первая используется для обучения классификатора, вторая – для независимой проверки качества его работы.

6 Классификатор может выдавать точный ответ $F^* : C \times D \to \left\{0,1\right\}$ или степень подобия $F^* : C \times D \to \left[0,1\right]$.

Этапы решения задачи классификации текстов.

Индексация документов. Построение некоторой числовой модели текста, например, в виде многомерного вектора слов и их веса в документе. Уменьшение размерности модели.

Построение и обучение классификатора. Могут использоваться различные методы машинного обучения: решающие деревья, наивный байесовский классификатор, нейронные сети, метод опорных векторов и др.

Оценка качества классификации. Можно оценивать по критериям полноты, точности, сравнивать классификаторы по специальным тестовым наборам.

Применение наивной байесовской модели к задаче классификации текстов.

Наивная байесовская модель является вероятностным методом обучения. Вероятность того, что документ *d* попадёт в класс *c* записывается как $P\left(c\middle|d\right)$. Поскольку цель классификации – найти самый подходящий класс для данного документа, то в наивной байесовской классификации задача состоит в нахождении наиболее вероятного класса $c_m = \arg\max_{c \in C} P\left(c\middle|d\right)$.

Вычислить значение этой вероятности напрямую невозможно, поскольку для этого нужно, чтобы обучающее множество содержало все (или почти все) возможные комбинации классов и документов. Однако, используя формулу Байеса, можно переписать: $c_m = \arg\max_{c \in C} \frac{P(c)P(d|c)}{P(d)} = \arg\max_{c \in C} P(c)P(d|c)$, где знаменатель $P(d)$ опущен, т. к. не зависит от *c* и, следовательно, не влияет на нахождение максимума; $P(c)-$ вероятность того, что встретится класс *c* независимо от рассматриваемого документа; $P(d|c)-$ вероятность встретить документ *d* среди документов класса *c*.

Используя обучающее множество, вероятность $P(c)$ можно оценить как $\hat{P}(c) = \frac{N_c}{N}$, где $N_c -$ количество документов в классе *c*, $N -$ общее количество документов в обучающем множестве. Здесь использован другой знак для вероятности, $\hat{P}$, поскольку с помощью обучающего множества можно лишь оценить вероятность, но не найти её точное значение.

Чтобы оценить вероятность $P(d|c) = P(t_1,...,t_{n_d}|c)$, где $t_k -$ терм из документа *d*, $n_d -$ общее количество значимых термов в документе, необходимо ввести упрощающие предположения о условной независимости термов и о независимости позиций термов. Другими словами, мы пренебрегаем, во-первых, тем фактом, что в тексте на естественном языке появление одного слова часто тесно связано с появлением других слов (например, вероятнее, что слово интеграл встретится в одном тексте со словом уравнение, чем со словом бактерия), и, во-вторых, что вероятность встретить одно и то же слово различна для разных позиций в тексте. Именно из-за этих грубых упрощений рассматриваемая модель естественного языка называется наивной (хотя она является достаточно эффективной в задаче классификации). Итак, в свете сделанных предположений, используя правило умножения вероятностей независимых событий, можно записать $P(d|c) = P(t_1,...,t_{n_d}|c) = P(t_1|c)...P(t_{n_d}|c) = \prod_{k=1}^{n_d} P(t_k|c)$.

Оценка вероятностей $P(t|c)$ с помощью обучающего множества будет $\hat{P}(t|c) = \frac{T_{ct}}{T_c}$, где $T_{ct} -$количество вхождений терма *t* во всех документах класса *c* (и на любых позициях – здесь существенно используется второе упрощающее предположение, иначе пришлось бы вычислить эти вероятности для каждой позиции в документе, что невозможно сделать достаточно точно из-за разреженности обучающих данных – трудно ожидать, чтобы каждый терм встретился в каждой позиции достаточное количество раз); $T_c -$ общее количество термов в документах класса *c*. При подсчёте учитываются все повторные вхождения.

После того, как классификатор «обучен», т. е. найдены величины $\hat{P}(c)$

и $\hat{P}(t|c)$, можно отыскать класс документа

$$c_m = \arg\max_{c \in C} \hat{P}(c)\hat{P}(d|c) = \arg\max_{c \in C} \hat{P}(c)\prod_{k=1}^{n_d} \hat{P}(t_k|c).$$

Чтобы избежать в последней формуле переполнения снизу, когда из-за большого числа сомножителей выражение $\hat{P}(c)\prod_{k=1}^{n_d} \hat{P}(t_k|c) = 0$, на практике вместо произведения обычно используют сумму логарифмов. Логарифмирование не влияет на нахождение максимума, т. к. логарифм является монотонно возрастающей функцией. Поэтому в большинстве реализаций вместо последней формулы используется

$$c_m = \arg\max_{c \in C}\left[\lg\left(\hat{P}(c)\hat{P}(d|c)\right)\right] = \arg\max_{c \in C}\left[\lg\hat{P}(c) + \sum_{k=1}^{n_d}\lg\hat{P}(t_k|c)\right].$$

Эта формула имеет простую интерпретацию. Шансы классифицировать документ часто встречающимся классом выше, и слагаемое $\lg\hat{P}(c)$ вносит в общую сумму соответствующий вклад. Величины же $\lg\hat{P}(t_k|c)$ тем больше, чем важнее терм *t* для идентификации класса *c*, и, соответственно, тем весомее их вклад в общую сумму.

Необходимо также отметить возможность использования различных терминов, как и их количества, для описания различных классов документов. В этом случае некоторые значения вероятностей $\hat{P}(t_k|c)$ будут равны нулю и не будут вовлечены в вычисление c_m в связи с сингулярностью выражения $\lg\hat{P}(t_k|c)$. При возникновении подобной ситуации, вероятности $\hat{P}(t_k|c)$ будут игнорироваться и финальное значение c_m должно быть нормировано на фактическое значение числа слагаемых $\lg\hat{P}(t_k|c)$.

10.3 Задание

Разработать систему классификации текстов на основе наивного байесовского классификатора. Например, создание почтового фильтра, отмечающего, к какой категории отнести входящее сообщение (спам, личное или рабочее письмо). Для определения категоризирующих терминов необходимо произвести частотный анализ текста как всех сообщений из обучающего множества, так и предклассифицированных.

10.4 Содержание отчёта

Отчёт должен представлять собой содержательные выводы по результатам всей выполненной работы.

Список литературы

1 **Бендат, Дж.** Прикладной анализ случайных данных / Дж. Бендат, А. Пирсол. – Москва: Мир, 1989. – 540 с.

2 **Тихонов, А. Н.** Статистическая обработка результатов экспериментов: учебное пособие / А. Н. Тихонов, М. В. Уфимцев. – Москва: МГУ, 1988. – 174 с.

3 **Себер, Дж.** Линейный регрессионный анализ / Дж. Себер. – Москва: Мир, 1980. – 456 с.

4 **Дайнтбегов, Д. М.** Программное обеспечение статистической обработки данных: учебное пособие / Д. М. Дайнтбегов, О. В. Калмыкова, А. И. Черепанов. – Москва: Финансы и статистика, 1984. – 192 с.

5 **Лоусон, Ч.** Численное решение задач методом наименьших квадратов / Ч. Лоусон, Р. Хенсон. – Москва: Наука, 1986. – 232 с.

6 **Отнес, Р.** Прикладной анализ временных рядов / Р. Отнес, Л. Эноксон. – Москва: Мир, 1982. – 432 с.