УДК 00.4

# ИСПОЛЬЗОВАНИЕ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ПРИ АНАЛИЗЕ МЕДИЦИНСКИХ ДАННЫХ

В. В. ПАСЕДЬКО, Н. В. ВЫГОВСКАЯ
Белорусско-Российский университет
Могилев, Беларусь



Логистическая регрессия или логит регрессия (англ. logit model) – это статистическая модель, используемая для предсказания вероятности возникновения некоторого события путём подгонки данных к логистической кривой. Метод, основанный на применении логистической регрессии, является одним из самых используемых при решении проблемы классификации.

Была поставлена задача проанализировать медицинские данные пациенток с опухолями в грудной массе, которые были собраны в штате Висконсин, США. Данные были взяты из Kaggle – системы организации конкурсов по исследованию данных, а также социальной сети специалистов по обработке данных и машинному обучению. Данные были получены из оцифрованного изображения биопсии грудной массы, также они были собраны доктором Уильямом Х. Вольбергом в университете Висконсин, больница Мэдисон, США. Они представляют собой характеристики ядер клеток, присутствующих на вышеуказанном изображении.

Для анализа данных использовался язык программирования Python и его библиотеки для визуализации. В качестве метода для построения предиктивной математической модели был выбран метод логистической регрессии, который также был имплементирован на языке Python. Целью разработки программного обеспечения (ПО) было построение модели, способной прогнозировать вероятность рака груди на новых данных. Для прогнозирования требуется взять биопсию и оцифровать её изображение. По этим новым данным можно будет прогнозировать вероятность рака груди у пациента.

В процессе работы был проведён разведывательный анализ данных, а именно:

– рассчитаны статистические показатели для каждой переменной-предиктора;

– рассчитано распределение целевой переменной;

– построены графики распределений переменных-предикторов;

– рассчитана ядерная оценка плотности для переменных-предикторов;

– построены графики корреляций между переменными-предикторами, а также между предикторами и целевой переменной;

– построен график тепловой карты, который отображает корреляции между переменными.

Найдены позитивно и негативно коррелирующие между собой предикторы.

Далее был выполнен этап подготовки к построению модели, а именно:

– определен метод логистической регрессии как оптимальный для выполнения задачи классификации;

– определен метод оценки результата выполнения модели.

Следующим шагом была подготовка датасета:

– определена матрица переменных-предикторов, а также вектор целевой переменной;

– данные стандартизированы (Feature Scaling);

– датасет разделён на тестовую и обучающую выборки;

– подобраны оптимальные гиперпараметры для построения модели;

– произведено обучение модели на обучающей выборке, в процессе обучения подобраны коэффициенты (веса) для каждой переменной-предиктора.

После применения модели на тестовой выборке были получены следующие результаты:

– точность модели на тестовой выборке – 99,4 %;

– точность рассчитывалась по формуле

$$accuracy = (TP + TN)/(TP + TN + FP + FN),$$

где $TP$ – количество правильно идентифицированных злокачественных опухолей; $TN$ – количество правильно идентифицированных доброкачественных опухолей; $FP$ – количество ложноотрицательных результатов; $FN$ – количество ложноположительных результатов.

Была построена матрица (рис. 1) выполнения модели на тестовой выборке.

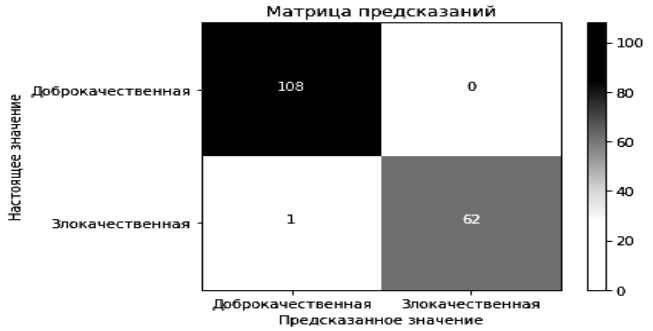


Рис. 1. Матрица модели тестовой выборки

Разработанное ПО может быть использовано при диагностике раковых заболеваний у пациенток с опухолями в грудной массе наряду с другими методами медицинской диагностики.