**МИНОБРНАУКИ РОССИИ**

**САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ**

**ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ**

**«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)**

**КАФЕДРА САПР**

**ОТЧЕТ**

**по Лабораторной работе № 2**

**«КЛАССИФИКАТОР НА ОСНОВЕ**

**ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ С ГРАДИЕНТНЫМ СПУСКОМ»**

**по дисциплине «Автоматизация схематического проектирования»**

|  |  |  |
| --- | --- | --- |
| Студенты гр. 0302 |  | Сухарев Л. А.. |
| Преподаватель |  | Боброва Ю. О. |

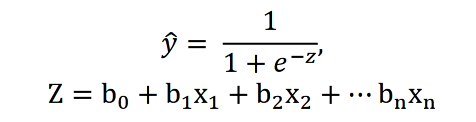
Санкт-Петербург

2024

**Цель:** разработка модели классификатора на основе логистической регрессии, изучение его свойств и принципов работы, получение навыков программирования на Python и использования модуля scikit-learn.

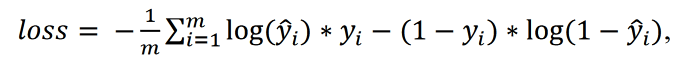
1. **ОСНОВЫНЕ ТЕОРЕТИЧЕСКИЕ ПОЛОЖЕНИЯ**

Результат работы классификатора 𝑦̂ – вероятность наступления события для объекта X, где X – вектор-строка {x1, x2… xn} в n-мерном пространстве, рассчитывается по формуле сигмоиды:



Метод логистической регрессии относится к методам обучения с учителем. Под обучением понимается оптимизационный алгоритм, минимизирующий ошибку предсказания. В результате обучения модели формируется набор коэффициентов, который используется для предсказаний на новых данных.

Ошибка предсказания описывается с помощью функции потерь (loss function). Она может определяться разными способами, наиболее часто используемой является следующая функция:



где m – число объектов в выборке, 𝑦𝑖 – действительное значение класса i-го объекта (0 или 1), 𝑦̂𝑖 – выход сигмоидальной функции (результат предсказания, от 0 до 1).

Метод градиентного спуска – один из методов обучения модели. Более подробно о нем можно узнать в материалах курса.

Чувствительность (истинно положительная пропорция) отражает долю положительных результатов, которые правильно идентифицированы как таковые. Иными словами, чувствительность диагностического теста показывает вероятность того, что больной субъект будет классифицирован именно как больной. Специфичность (истинно отрицательная пропорция) отражает долю отрицательных результатов, которые правильно идентифицированы как таковые, то есть вероятность того, что не больные субъекты будут классифицированы именно как не больные.

Библиотека scikit-learn – одна из самых популярных библиотек для анализа данных. Она включает сотни функций, покрывающие почти все задачи базового анализа. В частности, в библиотеке реализован метод логрегрессии с возможностью настройки гиперпараметров модели, а также выбора оптимизационного алгоритма.

Данная библиотека ориентирована на моделирование данных. Она не ориентирована на их загрузку, манипулирование и суммирование. Для решения этих задач, как правило, применяются библиотеки NumPy и Pandas.

Используемые модули: numpy, scikit-learn, matplotlib

1. **КОД ПРОГРАММЫ. ПОЯСНЕНИЕ ЭТАПОВ**

Для начала импортируем необходимые библиотеки и функции генерации данных, разработанные в предыдущих лабораторных работах. (Рис.1) Также был создан файл estimation, который включает в себя расчёт параметров, нужных для расчета точности, чувствительности и специфичности (Рис.2)

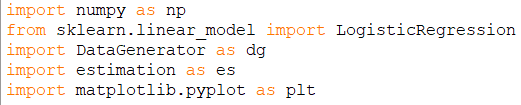


Рис. 1 — Начало работы

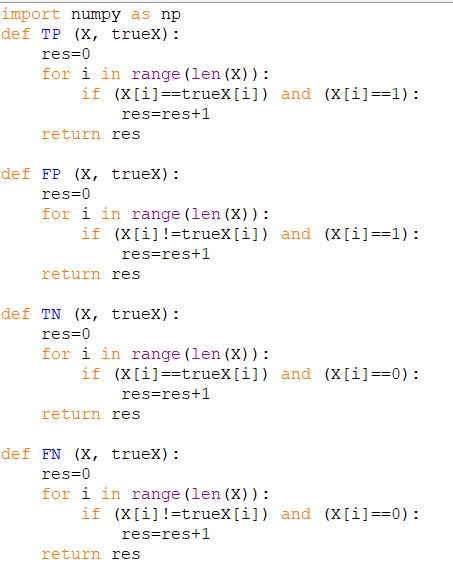


Рис. 2 — Содержание файла estimation

Используя функцию и скрипты, разработанные в лабораторной работе №1, создадим четыре переменных: Xtrain, Ytrain; Xtest, Ytest. Здесь X –объекты двух классов, параметры которых распределенны нормально, а объекты классов хорошо линейно разделимы. Y – метки класса. Для обучения и тестирования модели будут использованы разные выборки.

После ознакомления с документацией по LogisticRegression для модуля scikit-learn, обучим модель на обучающей выборке, используя метод fit(). (Рис.3)

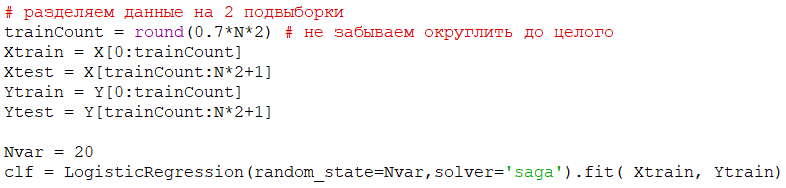


Рис. 3 — Создание переменных и обучение модели

Получите предсказание для новых данных, используя методы predict() и predict\_proba()



Рис. 4 — Предсказание для новых данных

Оценим точность предсказаний, используя метод score() и вручную. (Рис.5)

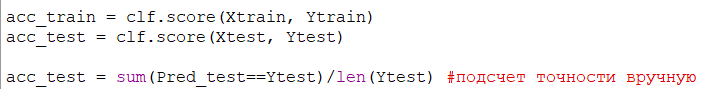


Рис. 5 — Подсчет точности

Визуализируйте результаты классификации для обучающей и тестовой выборок, отобразив на гистограмме распределения вероятности, полученной на выходе из классификатора (Рис. 6)

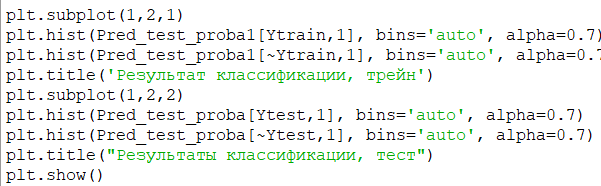


Рис. 6 — Визуализация результатов.

По полученным гистограммам вероятности принадлежности объектов 1 классу для обучающей и тестовой выборок (Рис.7) можно заметить, что для данной выборки результаты классификации не будут идеальны, так как графики перекрываются в середине. Это связанно с тем, что в данной выборке классы частично пересекаются. Точность, чувствительность и специфичность, полученные без использования методов библиотеки, можно увидеть на рисунке 8, так же, как и код для их подсчета.

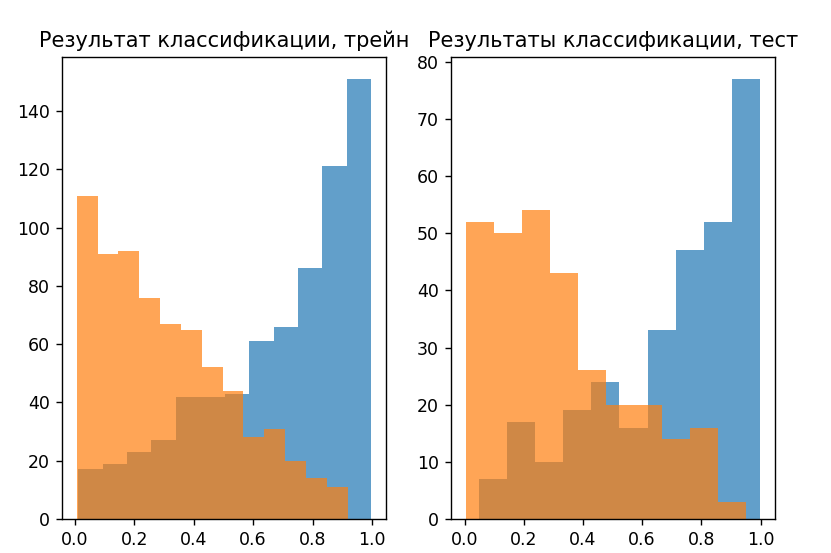


Рис. 7 — Вероятности принадлежности объектов 1 классу для обучающей и тестовой выборок

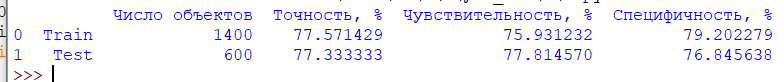
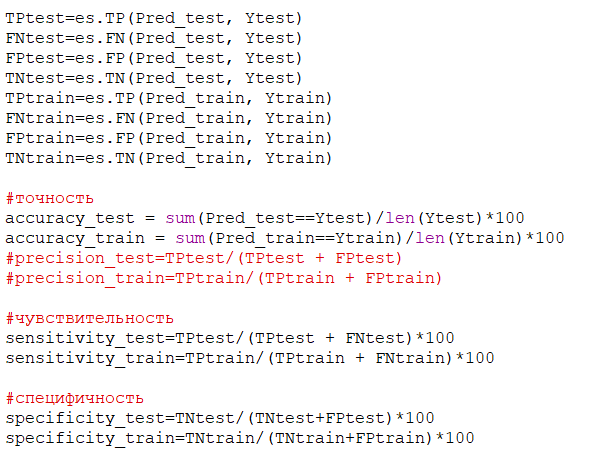
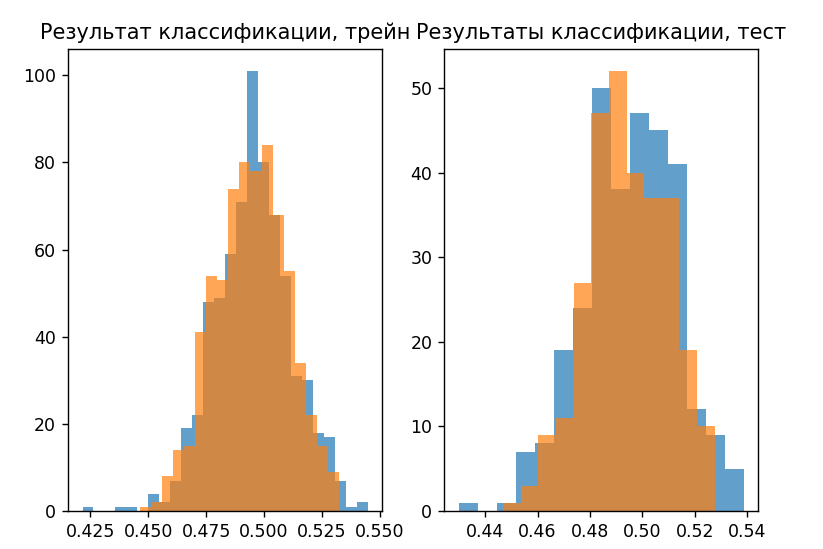


Рис. 8 — Расчет точности, чувствительности и специфичности для обучающей и тестовых выборок

Добьемся более плотного пересечения классов, меняя средние значения mu. (Рис.9)



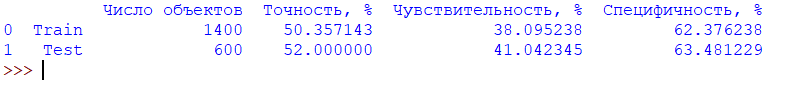


Рис. 9 — Результат классификации плотно пересекаемых классов

Также рассмотрим, как классификатор работает на нелинейно пересекаемых классах (Рис.10). Можно заметить, что на хорошо разделенных классах классификатор работает лучше всего.

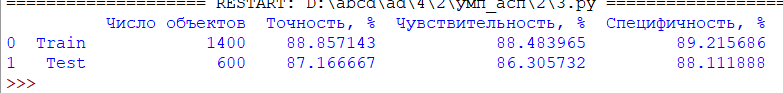
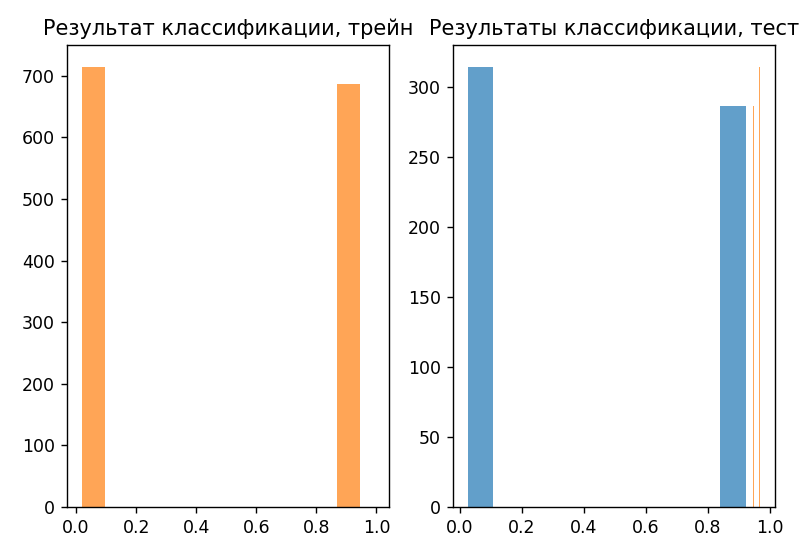


Рис. 10 — Результат работы классификатора

**Выводы**

В ходе данной работе были получены навыки разработки модели классификатора на основе логистической регрессии, изучены его свойств и принципы работы, получены навыки программирования на Python и использования модуля scikit-learn. Было выяснено, что чем лучше разделены классы, тем лучше с их классификацией справляется классификатор. Насколько хорошо разделены классы можно судить по гистограмме результатов классификации, которая строится по вероятностям определения в один или другой класс. Также были изучены и успешно воссозданы формулы для расчета точности, чувствительности и специфичности классификатора, результаты подсчета по которым были сведены в таблицу, используя знания из прошлой лабораторной работы. По полученным таблицам легко можно оценить результаты работы классификатора. После изучения теория, что чем лучше разделены данные, тем точнее работает классификатор, подтвердилась.