МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт No8 «Информационные технологии и прикладная математика»

Кафедра 810 «Информационные технологии в моделировании и управлении»

**Теоретическое задание No2**

**по курсу «Интеллектуальный анализ данных»**

**Логистическая регрессия. Метод опорных векторов. Калибровочные кривые.**

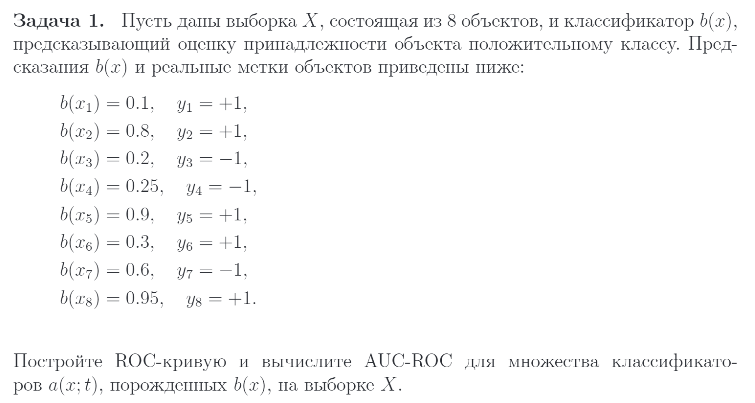
Выполнил: Д.А. Аникин

Группа: М8O-103М-19

Преподаватель: К.К. Абгарян

Москва, 2020

**Теоритическая часть**

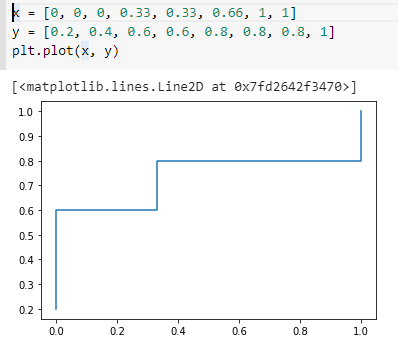


**Решение:**

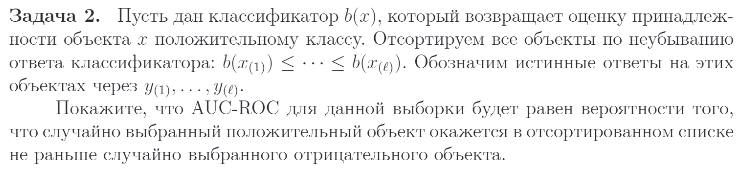
Упорядочим объекты:

|  |  |  |
| --- | --- | --- |
| 1 | 0.95 | 1 |
| 2 | 0.9 | 1 |
| 3 | 0.8 | 1 |
| 4 | 0.6 | -1 |
| 5 | 0.3 | 1 |
| 6 | 0.25 | -1 |
| 7 | 0.2 | -1 |
| 8 | 0.1 | 1 |

ROC – кривая:



AUC - ROC = 0.732



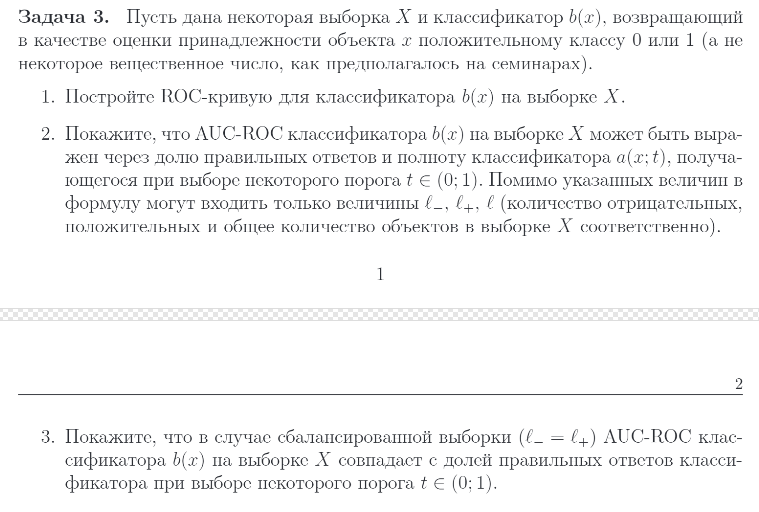
**Решение:**

ROC – это график TPR от FPR.

Запишем ниже площадь под графиком, как раз наш ROC-AUC в виде интеграла от этой функции:

где функция плотности вероятности:

и производная функции распределения:

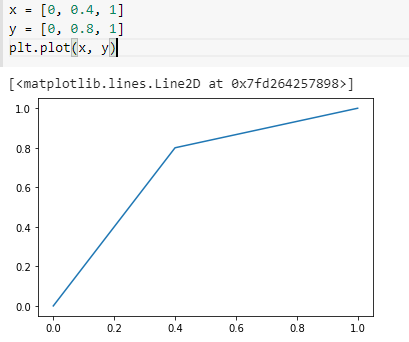


**Решение:**

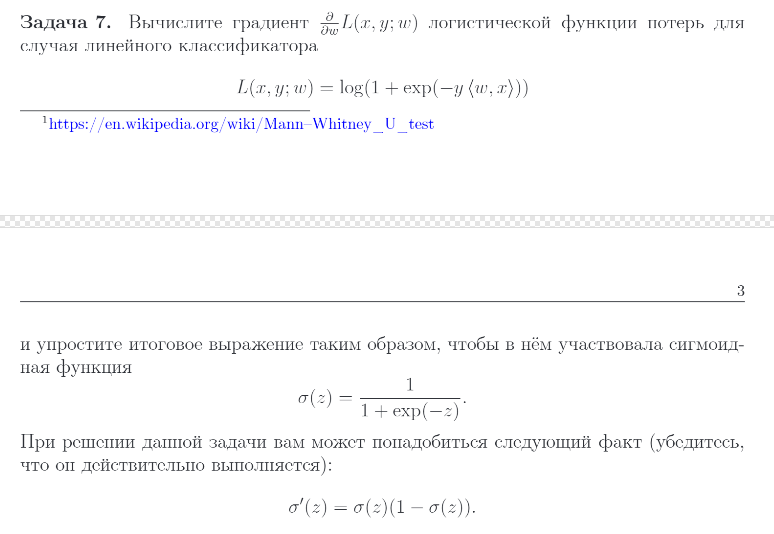
Возьмем выборку из задания 1 с порогом 0,8:

|  |  |  |  |
| --- | --- | --- | --- |
| № | Вероятность | Бинарное значение с порогом 0,4 | Верный ответ |
| 1 | 0.95 | 1 | 1 |
| 2 | 0.9 | 1 | 1 |
| 3 | 0.8 | 1 | 1 |
| 4 | 0.6 | 1 | -1 |
| 5 | 0.3 | -1 | 1 |
| 6 | 0.25 | -1 | -1 |
| 7 | 0.2 | -1 | -1 |
| 8 | 0.1 | -1 | 1 |

1. ROC – кривая:



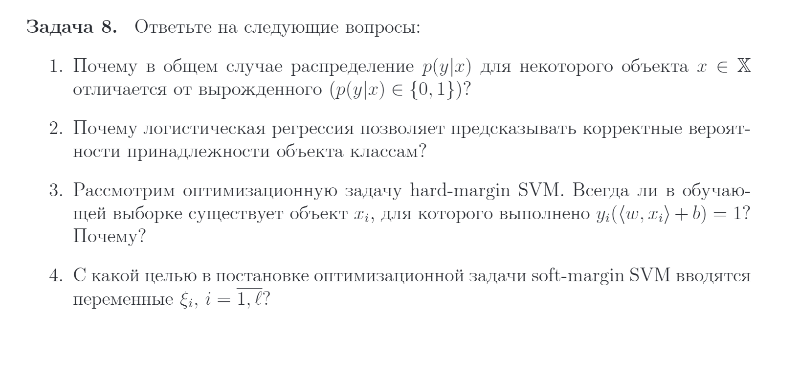
2 и 3) Из графика ROC – кривой зависимости TPR от FPR:



**Решение:**

Посчитаем производную сигмоиды:

Немного преобразуем функцию потерь и посчитаем градиент:



**Решение:**

1. Так как нескольким объектам с одинаковым признаковым описанием могут соответствовать разные классы.
2. Выведем оптимальный ответ алгоритма с log-loss  
   Данная функция потерь называется логарифмической (log-loss). Покажем, что она также позволяет корректно предсказывать вероятности. Запишем матожидание функции потерь в точке x:  
   Продифференцируем по b:

Легко видеть, что оптимальный ответ алгоритма равен вероятности положительного класса:

1. Да, потому что это условие соответствует принадлежности точки к полосе разделения. А сам алгоритм SVM максимизирует зазор между гиперплоскостью и объектами классов, которые расположены ближе всего к ней. При этом в каждом из классов найдётся хотя бы один "граничный" объект обучающей выборки, отступ которого равен этому минимуму: иначе можно было бы сместить гиперплоскость в сторону класса с большим отступом, тем самым увеличив минимальное расстояние от гиперплоскости до объектов обучающей выборки.
2. Эти переменные являются штрафом для объектов x, что позволяет перейти к задачам линейно неразделимыми множествами.

Выводы

По итоге выполнения лабораторной работы удалось изучить метод опорных векторов и логиситческую регрессию. Были пристально рассмотрены различные метрики, в особенности ROC-кривая, и ее интегральный показатель – ROC-AUC (площадь под этой кривой).