МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)

Институт №8 «Информационные технологии и прикладная математика»
Кафедра 810Б «Информационные технологии в моделировании и управлении»

Практическое задание №2
по курсу «Интелектуальный анализ данных»

Линейные алгоритмы классификации.

Выполнил: Щербаков В.С.
Группа: М8О-110М-19
Преподаватель: Абгарян К.К.

Москва, 2020

Содержание

Задание 1.

Пусть даны выборка X, состоящая из 8 объектов, и классификатор b(x), предсказывающий оценку принадлежности объекта положительному классу. Предсказания b(x) и реальные метки объектов приведены ниже:

b(x1) = 0.1, y1 = +1,

b(x2) = 0.8, y2 = +1,

b(x3) = 0.2, y3 = −1,

b(x4) = 0.25, y4 = −1,

b(x5) = 0.9, y5 = +1,

b(x6) = 0.3, y6 = +1,

b(x7) = 0.6, y7 = −1,

b(x8) = 0.95, y8 = +1.

Постройте ROC-кривую и вычислите AUC-ROC для множества классификаторов a(x; t), порожденных b(x), на выборке X.

Решение:

Для построения ROC-кривой выставим все предсказания по мере убывания вероятности и посчитаем, сколько наблюдений положительного класса выше каждого наблюдения отрицательного класса.

b(xn)	yn	Количество положительных над отрицательным классом
0.95	+1	
0.9	+1	
0.8	+1	
0.6	-1	3
0.3	+1	
0.25	-1	4
0.2	-1	4
0.1	+1	

Код отрисовки ROC-кривой на зяыке python 3.7 с использованием библиотек sklearn, matplotlib:

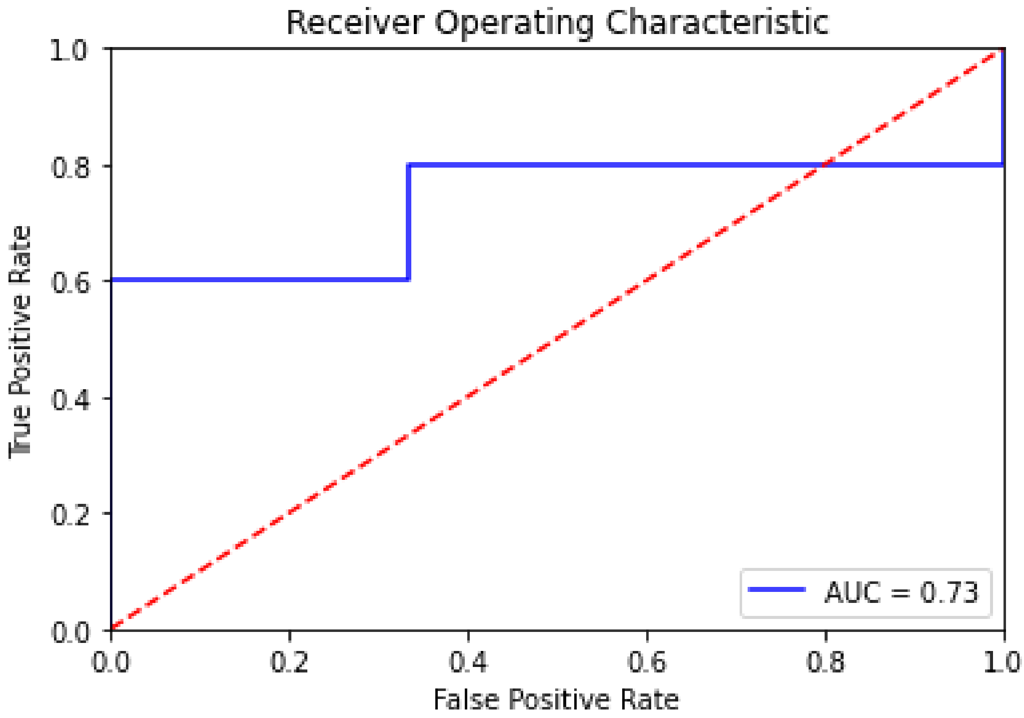
```
import sklearn.metrics as metrics
import matplotlib.pyplot as plt

y_test = [1, 1, 1, -1, 1, -1, -1, 1]
preds = [0.95, 0.9, 0.8, 0.6, 0.3, 0.25, 0.2, 0.1]
fpr, tpr, threshold = metrics.roc_curve(y_test, preds)
```

```
roc_auc = metrics.auc(fpr, tpr)

plt.title('Receiver Operating Characteristic')
plt.plot(fpr, tpr, 'b', label = 'AUC = %0.2f' % roc_auc)
plt.legend(loc = 'lower right')
plt.plot([0, 1], [0, 1],'r--')
plt.xlim([0, 1])
plt.ylim([0, 1])
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

Полученная ROC- кривая:



Для вычисления AUC-ROC для данного классификатора применим данную формулу:

AUC=1PNj=1N(sj-j)=1PNj=1Nt=1sj-j

Получим:

AUC=135(3+4+4)=11150,733

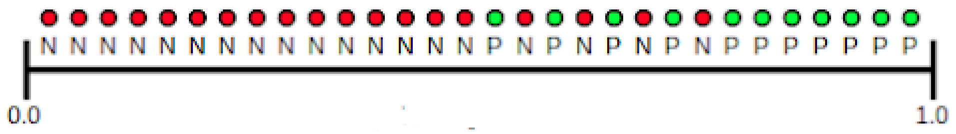
Такой показатель AUC означает, что у множества классификаторов a(x; t), порожденных b(x), хорошая дискриминирующая способность.

Задание 2.

Пусть дан классификатор b(x), который возвращает оценку принадлежности объекта x положительному классу. Отсортируем все объекты по неубыванию ответа классификатора: $b(x(1)) \leq \cdot\cdot\cdot \leq b(x(\ell))$. Обозначим истинные ответы на этих объектах через $y(1), \ldots, y(\ell)$. Покажите, что AUC-ROC для данной выборки будет равен вероятности того, что случайно выбранный положительный объект окажется в отсортированном списке не раньше случайно выбранного отрицательного объекта.

Решение:

AUC предоставляет совокупное измерение производительности при всех возможных значениях классификационного порога. AUC выдает результаты в периоде от 0 до 1. Модель, чьи прогнозы 100% ошибочны, имеет AUC равную 0.0, а модель со 100% верными прогнозами имеет AUC равную 1.0. Предположим, что прогнозы нашего классификатора будут распределены следующим образом:



Где N - отрицательный прогноз, P - положительный прогноз.
AUC предоставляет вероятность того, что случайный позитивный (зеленый) пример будет расположен правее случайного отрицательного (красного) примера.

Из расположений предсказаний на данной прямой можно сделать вывод, что что AUC-ROC для данной выборки будет равен вероятности того, что случайно выбранный положительный объект окажется в отсортированном списке не раньше случайно выбранного отрицательного объекта.

Задание 4.

В анализе данных для сравнения среднего значения некоторой величины у объектов двух выборок часто используется критерий Манна–Уитни–Уилкоксона1, основанный на вычислении U-статистики.

Пусть у нас имеется выборка X и классификатор $b(x)$, возвращающий оценку принадлежности объекта x положительному классу. Тогда вычисление U-статистики для подвыборки X, состоящей из объектов положительного класса, производится следующим образом: объекты обеих выборок сортируются по неубыванию значения $b(x)$, после чего каждому объекту в полученном упорядоченном ряду $x_{(1)}, \dots, x_{(\ell)}$ присваивается ранг — номер позиции $r_{(i)}$ в ряду (начиная с 1, при этом для объектов с одинаковыми значением $b(x)$ в качестве ранга присваивается среднее значение ранга для таких объектов). Тогда U-статистика для объектов положительного класса равна:

$$U_+ = \sum_{\substack{i=1\\ y_{(i)}=+1}}^{\ell} r_{(i)} - \frac{\ell_+(\ell_+ + 1)}{2}.$$

Покажите, что для значения AUC-ROC классификатора $b(x)$ на выборке X и U-статистики верно следующее соотношение:

$$\mathrm{AUC} = \frac{U_+}{\ell_-\ell_+}.$$

Для начала представим ранги элементов в аналитическом виде. Ранг элемента - это число элементов, не больших данного (включая его самого). Для удобства обозначений будем считать, что индексация ведется по возрастанию ответов классификатора. Тогда можно записать следующее:

$$r_{(i)} = \sum_{j \le i} 1 = \sum_{j \le i} [y_{(j)} = -1] + \sum_{j \le i} [y_{(j)} = +1]$$

Теперь посчитаем U_+ статистику:

$$\begin{aligned}
U_+ = \sum_{i:y_{(i)}=+1} r_{(i)} - \frac{l_+(l_+ + 1)}{2} = \sum_{i=1}^{l} [y_{(i)} = +1]\, r_{(i)} - \frac{l_+(l_+ + 1)}{2} = \\
= \sum_{i=1}^{l} \left([y_{(i)} = +1] \sum_{j<i} [y_{(j)} = -1] \right) + \\
+ \sum_{i=1}^{l} \left([y_{(i)} = +1] \sum_{j \le i} [y_{(j)} = +1] \right) - \frac{l_+(l_+ + 1)}{2} = \\
= \sum_{j<i} [y_{(j)} < y_{(i)}] + \sum_{i=1}^{l_+} i - \frac{l_+(l_+ + 1)}{2} = \\
= \sum_{j<i} [y_{(j)} < y_{(i)}]
\end{aligned}$$

Таким образом, $\frac{U_+}{l_+ l_-} = \frac{\sum_{j<i} [y_{(j)} < y_{(i)}]}{l_+ l_-}$, а это и есть не что иное, как AUC-ROC, то есть число пар объектов из разных классов, верно разделенных классификатором, к общему числу пар.

Задание 7.

Вычислите градиент $L(x, y; w)$ логистической функции потерь для случая линейного классификатора $L(x, y; w) = \log(1 + \exp(-y\,(w, x)))$ и упростите итоговое выражение таким образом, чтобы в нём участвовала сигмоидальная функция

$$\sigma(z) = 1 /(1 + \exp(-z))$$

При решении данной задачи вам может понадобиться следующий факт (убедитесь, что он действительно выполняется):

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)).$$

Посчитаем производную сигмоиды:

$$\sigma'(z) = -\frac{1}{\left(1+e^{-z}\right)^2}\left(-e^{-z}\right) = \frac{e^{-z}}{\left(1+e^{-z}\right)^2} =$$
$$= \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) = \sigma(z)(1-\sigma(z))$$

Немного преобразуем функцию потерь и посчитаем градиент:

$$L(x, y, w) = \log\left(1 + \exp\left(-y < w, x >\right)\right) = -\log\left(\sigma(-y < w, x >)\right)$$

$$\frac{\partial L}{\partial w} = -\frac{1}{\sigma(-y < w, x >)}\sigma(-y < w, x >))\left(1 - \sigma(-y < w, x >)\right)(-yx) =$$
$$= yx\left(1 - \sigma(-y < w, x >)\right)$$

Задание 8.

Ответьте на следующие вопросы:

1. Почему в общем случае распределение p(y|x) для некоторого объекта $x \in X$ отличается от вырожденного ($p(y|x) \in \{0, 1\}$)?
2. Почему логистическая регрессия позволяет предсказывать корректные вероятности принадлежности объекта классам?

   Логистическая регрессия позволяет предсказывать корректные вероятности за счет того, что внутри нее используется сигмоидальная функция. Это позволяет избавится от ограничений, характерных, к примеру, для линейной регрессии.

3. Рассмотрим оптимизационную задачу hard-margin SVM. Всегда ли в обучающей выборке существует объект x_i, для которого выполнено $y_i(\langle w, x_i\rangle + b) = 1$? Почему?
4. С какой целью в постановке оптимизационной задачи soft-margin SVM вводятся переменные ξ_i, $i = \overline{1, \ell}$?
   Метод опорных векторов с жестким зазором справляется с своей задачей до тех пор, пока у классы линейно разделимы. Чтобы алгоритм смог работать и с линейно неразделимых данными, необходимо немного преобразовать систему постановке задачи. Необходимо позволить алгоритму допускать ошибки на обучающих объектах, но при этом постараться, чтобы ошибок