

Машинное обучение

Теоретическое домашнее задание №3

Задача 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:

$$\begin{aligned}b(x_1) &= 0.1, & y_1 &= +1, \\b(x_2) &= 0.8, & y_2 &= +1, \\b(x_3) &= 0.2, & y_3 &= -1, \\b(x_4) &= 0.25, & y_4 &= -1, \\b(x_5) &= 0.9, & y_5 &= +1, \\b(x_6) &= 0.3, & y_6 &= +1, \\b(x_7) &= 0.6, & y_7 &= -1, \\b(x_8) &= 0.95, & y_8 &= +1.\end{aligned}$$

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

Задача 2. Пусть дан классификатор $b(x)$, который возвращает оценку принадлежности объекта x положительному классу. Отсортируем все объекты по неубыванию ответа классификатора: $b(x_{(1)}) \leq \dots \leq b(x_{(\ell)})$. Обозначим истинные ответы на этих объектах через $y_{(1)}, \dots, y_{(\ell)}$.

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.

Задача 3. Пусть дана некоторая выборка X и классификатор $b(x)$, возвращающий в качестве оценки принадлежности объекта x положительному классу 0 или 1 (а не некоторое вещественное число, как предполагалось на семинарах).

1. Постройте ROC-кривую для классификатора $b(x)$ на выборке X .
2. Покажите, что AUC-ROC классификатора $b(x)$ на выборке X может быть выражен через долю правильных ответов и полноту классификатора $a(x; t)$, получающегося при выборе некоторого порога $t \in (0; 1)$. Помимо указанных величин в формулу могут входить только величины ℓ_- , ℓ_+ , ℓ (количество отрицательных, положительных и общее количество объектов в выборке X соответственно).

3. Покажите, что в случае сбалансированной выборки ($\ell_- = \ell_+$) AUC-ROC классификатора $b(x)$ на выборке X совпадает с долей правильных ответов классификатора при выборе некоторого порога $t \in (0; 1)$.

Задача 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 значениями $b(x)$ 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:

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

¹https://en.wikipedia.org/wiki/Mann-Whitney_U_test