

# Машинное обучение, ФКН ВШЭ

## Семинар №5

### 1 AUC-ROC

На предыдущем занятии мы познакомились с такой важной метрикой качества бинарной классификации, как площадь под ROC-кривой (AUC-ROC). Напомним её определение. Рассмотрим задачу бинарной классификации с метками классов  $\mathbb{Y} = \{-1, +1\}$ , и пусть задан некоторый алгоритм  $b(x)$ , позволяющий вычислять оценку принадлежности объекта  $x$  положительному классу. AUC-ROC позволяет оценивать качество классификации для множества алгоритмов следующего вида:

$$a(x; t) = \begin{cases} -1, & b(x) \leq t, \\ +1, & b(x) > t, \end{cases}$$

т.е. алгоритмов, присваивающих метки объектам в соответствии с оценками  $b(x)$ , отсекая их по некоторому порогу  $t$ . Каждый алгоритм (получающийся при фиксации значения порога  $t$ ) представляется точкой на плоскости (FPR, TPR), где

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{\ell_-},$$
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{\text{TP}}{\ell_+},$$

$\ell_-, \ell_+$  — количество объектов отрицательного и положительного классов соответственно. AUC-ROC, в свою очередь, является площадью под получившейся кривой.

Изучим подробнее некоторые важные свойства данной метрики.

Критерий AUC-ROC имеет большое число интерпретаций — например, он равен вероятности того, что случайно выбранный положительный объект окажется позже случайно выбранного отрицательного объекта в ранжированном списке, порожденном  $b(x)$ . Разберем подробнее немного другую формулировку.

**Задача 1.1.** В ранжировании часто используется функционал «доля дефектных пар». Его можно определить и для задачи бинарной классификации.

Пусть дан классификатор  $b(x)$ , который возвращает оценки принадлежности объектов классу  $+1$ . Отсортируем все объекты по неубыванию ответа классификатора  $b$ :  $x_{(1)}, \dots, x_{(\ell)}$ . Обозначим истинные ответы на этих объектах через  $y_{(1)}, \dots, y_{(\ell)}$ . Тогда доля дефектных пар записывается как

$$DP(b, X) = \frac{2}{\ell(\ell-1)} \sum_{i < j}^{\ell} [y_{(i)} > y_{(j)}].$$

Как данный функционал связан с AUC-ROC?

**Решение.** Для начала разберем процедуру построения ROC-кривой. Сперва все объекты сортируются по неубыванию оценки  $b(x)$ , тем самым формируя список  $x_{(1)}, \dots, x_{(\ell)}$ . После этого фиксируется значение порога  $t = b(x_{(\ell)}) + 1$ , в этом случае имеем

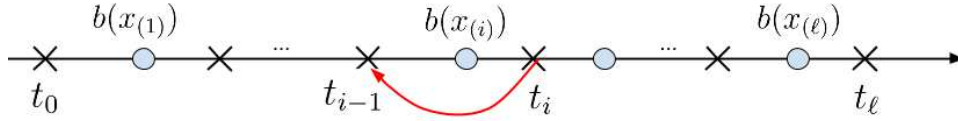
$$\text{FPR} = \frac{\text{FP}}{\ell_-} = \frac{0}{\ell_-} = 0,$$

$$\text{TPR} = \frac{\text{TP}}{\ell_+} = \frac{0}{\ell_+} = 0.$$

Таким образом, алгоритму  $a(x; b(x_{(\ell)}) + 1)$  соответствует точка  $(0; 0)$  на плоскости, откуда начинается построение ROC-кривой.

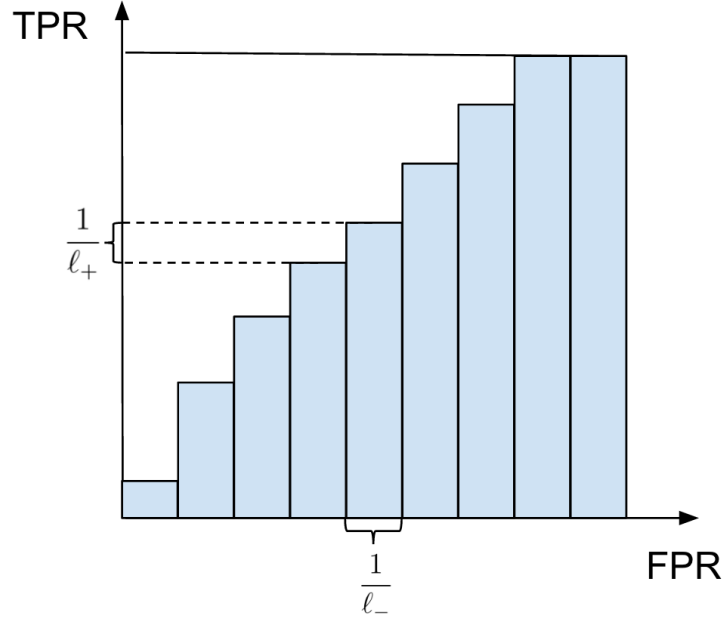
Заметим, что для построения ROC-кривой достаточно рассмотреть  $(\ell + 1)$  различных значений порога  $t$ , соответствующих всем различным способам классификации выборки, порожденным алгоритмом  $b(x)$ , — например, в качестве таких порогов можно рассмотреть следующий набор:

$$\begin{aligned} t_\ell &= b(x_{(\ell)}) + 1, \\ t_i &= \frac{b(x_{(i)}) + b(x_{(i+1)})}{2}, \quad i = \overline{1, \ell - 1}, \\ t_0 &= b(x_{(1)}) - 1. \end{aligned}$$



Будем перебирать пороги в порядке невозрастания их значения, начиная с  $t_\ell$ . Пусть мы хотим уменьшить значение порога с  $t_i$  до  $t_{i-1}$ . При этом классификация объекта  $x_{(i)}$  (и только его) изменится с отрицательной на положительную. Рассмотрим 2 случая.

1.  $y_{(i)} = +1$ . В этом случае классификатор начнет верно классифицировать объект, на котором ранее допускал ошибку, при этом FPR не изменится, а TPR повысится на  $\frac{1}{\ell_+}$ .
2.  $y_{(i)} = -1$ . В этом случае классификатор начнет ошибаться на объекте, который ранее классифицировал верно, при этом TPR не изменится, а FPR повысится на  $\frac{1}{\ell_-}$ .



Теперь рассмотрим, как при этом изменяется AUC-ROC. Заметим, что область под ROC-кривой состоит из непересекающихся прямоугольников, каждый из которых снизу ограничен осью FPR, а сверху — одним из горизонтальных отрезков, соответствующих второму из рассмотренных случаев. Поэтому каждый раз, когда имеет место второй случай, к текущей накопленной площади под кривой (которая изначально в точке  $(0; 0)$  равна 0) добавляется площадь прямоугольника, горизонтальные стороны которого равны  $\frac{1}{\ell_-}$ , а вертикальные равны  $\frac{1}{\ell_+} \sum_{j=i+1}^{\ell} [y_{(j)} = +1]$  (доля уже рассмотренных положительных объектов среди всех положительных), поэтому в этом случае текущее значение AUC-ROC увеличивается на  $\frac{1}{\ell_+ \ell_-} \sum_{j=i+1}^{\ell} [y_{(j)} = +1]$ . Итого, финальное значение AUC-ROC можно посчитать следующим образом:

$$\begin{aligned}
 \text{AUC} &= \frac{1}{\ell_+ \ell_-} \sum_{i=1}^{\ell} [y_{(i)} = -1] \sum_{j=i+1}^{\ell} [y_{(j)} = +1] = \\
 &= \frac{1}{\ell_+ \ell_-} \sum_{i=1}^{\ell} \sum_{j=i+1}^{\ell} [y_{(i)} < y_{(j)}] = \\
 &= \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} (1 - [y_{(i)} > y_{(j)}] - [y_{(j)} = y_{(i)}]) = \\
 &= \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} (1 - [y_{(j)} = y_{(i)}]) - \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} [y_{(i)} > y_{(j)}] = \\
 &= \frac{1}{\ell_+ \ell_-} \frac{\ell(\ell-1)}{2} - \frac{\ell_+(\ell_+-1)}{2\ell_+ \ell_-} - \frac{\ell_-(\ell_--1)}{2\ell_+ \ell_-} - \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} [y_{(i)} > y_{(j)}] = \\
 &= 1 - \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} [y_{(i)} > y_{(j)}].
 \end{aligned}$$

Отсюда получаем, что AUC-ROC и доля дефектных пар связаны следующим соотношением:

$$\text{DP}(b, X) = \frac{2\ell - \ell_+}{\ell(\ell - 1)}(1 - \text{AUC}(b, X)).$$

■

**Задача 1.2.** Пусть даны выборка  $X$ , состоящая из 5 объектов, и классификатор  $b(x)$ , предсказывающий оценку принадлежности объекта положительному классу. Предсказания  $b(x)$  и реальные метки объектов приведены ниже:

$$\begin{aligned} b(x_1) &= 0.2, & y_1 &= -1, \\ b(x_2) &= 0.4, & y_2 &= +1, \\ b(x_3) &= 0.1, & y_3 &= -1, \\ b(x_4) &= 0.7, & y_4 &= +1, \\ b(x_5) &= 0.05, & y_5 &= +1. \end{aligned}$$

Вычислите AUC-ROC для множества классификаторов  $a(x; t)$ , порожденного  $b(x)$ , на выборке  $X$ .

**Решение.** В соответствии с процессом построения ROC-кривой, описанным в предыдущей задаче, отсортируем оценки  $b(x_i)$  в порядке их неубывания:  $(b(x_{(i)}))_{i=1}^{\ell} = (0.05, 0.1, 0.2, 0.4, 0.7)$ . Также составим последовательность реальных меток объектов из этого упорядоченного списка:  $(y_{(i)})_{i=1}^{\ell} = (+1, -1, -1, +1, +1)$ .

Построим ROC-кривую (см. рис. 1), откуда  $\text{AUC-ROC} = \frac{2}{3}$ .

■

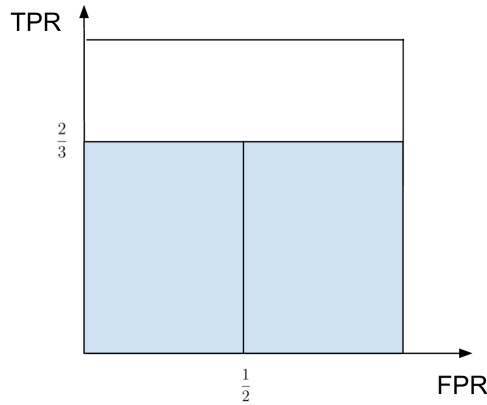


Рис. 1. Иллюстрация к задаче 1.2.

Заметим, что при вычислении AUC-ROC на некоторой выборке  $X$  для итогового классификатора  $a(x; t)$  важны не конкретные значения  $b(x_i)$ ,  $i = \overline{1, \ell}$ , а порядок расположения объектов в отсортированном по неубыванию списке  $b(x_{(1)}), \dots, b(x_{(\ell)})$ ,

порожденным алгоритмом  $b(x)$ . Таким образом, для фиксированной выборки  $X$  алгоритм  $b(x)$  задаёт перестановку на её объектах, которая в дальнейшем используется при расчёте AUC-ROC.

**Задача 1.3.** Пусть  $b(x)$  — классификатор, предсказывающий оценку принадлежности объекта  $x$  классу  $+1$  таким образом, что для некоторой выборки  $X$  он равновероятно выдаёт на её объектах одну из всех возможных перестановок. Чему равно матожидание AUC-ROC этого классификатора?

**Решение.** Как было показано в задаче 1.1, для AUC-ROC верно

$$\text{AUC} = \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} [y_{(i)} = -1][y_{(j)} = +1],$$

поэтому

$$\mathbb{E}\text{AUC} = \frac{1}{\ell_+ \ell_-} \sum_{i < j}^{\ell} \mathbb{E}([y_{(i)} = -1][y_{(j)} = +1]).$$

Заметим, что величина  $[y_{(i)} = -1][y_{(j)} = +1]$  принимает значения 0 и 1, поэтому

$$\mathbb{E}([y_{(i)} = -1][y_{(j)} = +1]) = \mathbb{P}(y_{(i)} = -1, y_{(j)} = +1) = \frac{\ell_- \ell_+ (\ell - 2)!}{\ell!} = \frac{\ell_- \ell_+}{\ell(\ell - 1)}.$$

Отсюда имеем

$$\mathbb{E}\text{AUC} = \frac{1}{\ell_- \ell_+} \sum_{i < j}^{\ell} \frac{\ell_- \ell_+}{\ell(\ell - 1)} = \frac{\ell(\ell - 1)}{2} \frac{1}{\ell(\ell - 1)} = \frac{1}{2}.$$

■

Итого, можем заметить, что значение AUC-ROC, близкое к  $\frac{1}{2}$ , означает, что классификатор близок к случайному, тогда как значение, равное 1, означает, что классификатор безошибочно классифицирует объекты при некотором значении порога.

**Задача 1.4.** Пусть  $b(x)$  — некоторый классификатор, предсказывающий оценку принадлежности объекта  $x$  положительному классу, и при этом AUC-ROC множества классификаторов  $a(x; t)$ , порожденных  $b(x)$ , на некоторой выборке  $X$  принимает значение, меньшее 0.5. Как можно скорректировать прогнозы классификаторов  $a(x; t)$ , чтобы они были более осмысленными по сравнению с прогнозами классификатора, выдающего случайные ответы?

**Решение.** Заметим, что, поскольку  $\text{AUC-ROC} < 0.5$ , то ROC-кривая находится под диагональю единичного квадрата.

Для некоторого классификатора  $a(x; t)$  рассмотрим классификатор  $a^*(x; t)$ , выдающий противоположные метки по сравнению с  $a(x; t)$ , т.е.:

$$a^*(x; t) = -a(x; t).$$

При этом будут выполняться соотношения

$$\begin{aligned}\text{TPR}(a^*, X) &= \text{FPR}(a, X), \\ \text{FPR}(a^*, X) &= \text{TPR}(a, X),\end{aligned}$$

поэтому классификатор  $a^*(x; t)$  будет представлен на плоскости точкой, симметричной точке, отвечающей классификатору  $a(x; t)$ , относительно диагонали единичного квадрата. Заметим, что это верно для всех значений  $t$ , поэтому каждой точке исходной ROC-кривой, расположенной под диагональю единичного квадрата, будет соответствовать точка, расположенная над ней. В связи с этим AUC-ROC множества классификаторов  $a^*(x; t)$  принимает значение, большее 0.5, а прогнозы классификаторов  $a^*(x; t)$  более осмысленны по сравнению со случайным классификатором. ■

## 2 Предсказание вероятностей

Разберемся, каким требованиям должен удовлетворять классификатор, чтобы его выход можно было расценивать как оценку вероятности класса.

Пусть в каждой точке  $x \in \mathbb{X}$  пространства объектов задана вероятность  $p(y = +1 | x)$  того, что данный объект относится к классу  $+1$ , и пусть алгоритм  $b(x)$  возвращает числа из отрезка  $[0, 1]$ . Потребуем, чтобы эти предсказания пытались в каждой точке  $x$  приблизить вероятность положительного класса  $p(y = +1 | x)$ .

Разумеется, выполнение этого требования зависит от функции потерь — минимум ее матожидания в каждой точке  $x$  должен достигаться на данной вероятности:

$$\arg \min_{b \in \mathbb{R}} \mathbb{E} [L(y, b) | x] = p(y = +1 | x).$$

**Задача 2.1.** Покажите, что квадратичная функция потерь  $L(y, z) = ([y = +1] - z)^2$  позволяет предсказывать корректные вероятности.

**Решение.** Заметим, что поскольку алгоритм возвращает числа от 0 до 1, то его ответ должен быть близок к единице, если объект относится к положительному классу, и к нулю — если объект относится к отрицательному классу.

Запишем матожидание функции потерь в точке  $x$ :

$$\mathbb{E} [L(y, b) | x] = p(y = +1 | x)(b - 1)^2 + (1 - p(y = +1 | x))(b - 0)^2.$$

Продифференцируем по  $b$ :

$$\frac{\partial}{\partial b} \mathbb{E} [L(y, b) | x] = 2p(y = +1 | x)(b - 1) + 2(1 - p(y = +1 | x))b = 2b - 2p(y = +1 | x) = 0.$$

Легко видеть, что оптимальный ответ алгоритма действительно равен вероятности:

$$b = p(y = +1 | x).$$

■

**Задача 2.2.** Покажите, что абсолютная функция потерь  $L(y, z) = |[y = +1] - z|$ ,  $z \in [0; 1]$ , не позволяет предсказывать корректные вероятности.

**Решение.** Запишем матожидание функции потерь в точке  $x$ :

$$\begin{aligned}\mathbb{E}[L(y, b)|x] &= p(y = +1|x)|1 - b| + (1 - p(y = +1|x))|b| = \\ &= p(y = +1|x)(1 - b) + (1 - p(y = +1|x))b.\end{aligned}$$

Продифференцируем по  $b$ :

$$\frac{\partial}{\partial b} \mathbb{E}[L(y, b)|x] = 1 - 2p(y = +1|x) = 0.$$

Рассмотрим 2 случая:

1.  $p(y = +1|x) = \frac{1}{2}$ . Тогда  $\mathbb{E}[L(y, b)|x] = \frac{1}{2} \quad \forall b \in [0; 1]$ , а потому классификатор не позволяет предсказывать корректную вероятность в точке  $x$ .
2.  $p(y = +1|x) \neq \frac{1}{2}$ . В этом случае интервал  $(0; 1)$  не содержит критических точек, а потому минимум матожидания достигается на одном из концов отрезка  $[0; 1]$ :

$$\begin{aligned}\min_{b \in [0; 1]} \mathbb{E}[L(y, b)|x] &= \min(\mathbb{E}[L(y, 0)|x], \mathbb{E}[L(y, 1)|x]) = \\ &= \min(p(y = +1|x), 1 - p(y = +1|x)).\end{aligned}$$

Отсюда  $\arg \min_{b \in [0; 1]} \mathbb{E}[L(y, b)|x] \in \{0, 1\}$ , а потому классификатор также не позволяет предсказывать корректную вероятность в точке  $x$ .

■

### 3 SVM

Будем рассматривать задачу бинарной классификации с метками  $\mathbb{Y} = \{-1, +1\}$  и линейные классификаторы вида

$$a(x) = \text{sign}\langle w, x \rangle + b, w \in \mathbb{R}^d, b \in \mathbb{R}.$$

На лекции были приведены оптимизационные задачи метода опорных векторов (SVM) для случаев линейно разделимой и неразделимой выборки. Напомним, как выглядит задача для второго случая:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C \sum_{i=1}^{\ell} \xi_i \rightarrow \min_{w, b, \xi}, \\ y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad i = \overline{1, \ell}, \\ \xi_i \geq 0, \quad i = \overline{1, \ell}. \end{cases}$$

Поскольку выборка не является линейно разделимой, решить исходную оптимизационную задачу не представляется возможным, в связи с чем вводятся штрафы  $\xi_i$ ,  $i = \overline{1, \ell}$ , для объектов за попадание внутрь разделяющей полосы. При этом излишний акцент на максимизации отступа приводит к большой ошибке на обучении, а «подгонка» под обучающую выборку, как правило, приводит к маленькой ширине разделяющей полосы. В связи с этим гиперпараметр  $C$  отвечает за то, какая из указанных целей является более приоритетной, — чем больше значение  $C$ , тем сильнее модель будет настраиваться на обучающую выборку.