# Машинное обучение
# ФКН ВШЭ
# Теоретическое домашнее задание №2

**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:

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

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 $\ell_-$, $\ell_+$, $\ell$ (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–Уилкоксона[1], 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:

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

**Задача 5.** Позволяет ли предсказывать корректные вероятности экспоненциальная функция потерь $L(y, z) = \exp(-yz)$?

**Задача 6.** Рассмотрим постановку оптимизационной задачи метода опорных векторов для линейно разделимой выборки:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min_{w,b}, \\ y_i(\langle w, x \rangle + b) \geq 1, \quad i = \overline{1, \ell}, \end{cases}$$

а также её видоизменёный вариант для некоторого значения $t > 0$:

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min_{w,b}, \\ y_i(\langle w, x \rangle + b) \geq t, \quad i = \overline{1, \ell}. \end{cases}$$

Покажите, что разделяющие гиперплоскости, получающиеся в результате решения каждой из этих задач, совпадают.

**Задача 7.** Вычислите градиент $\frac{\partial}{\partial w} L(x, y; w)$ логистической функции потерь для случая линейного классификатора

$$L(x, y; w) = \log(1 + \exp(-y \langle w, x \rangle))$$

[1] https://en.wikipedia.org/wiki/Mann–Whitney_U_test

и упростите итоговое выражение таким образом, чтобы в нём участвовала сигмоидная функция

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

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

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

**Задача 8.** Ответьте на следующие вопросы:

1. Почему в общем случае распределение $p(y|x)$ для некоторого объекта $x \in \mathbb{X}$ отличается от вырожденного ($p(y|x) \in \{0, 1\}$)?

2. Почему логистическая регрессия позволяет предсказывать корректные вероятности принадлежности объекта классам?

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 $\xi_i$, $i = \overline{1, \ell}$?