# Школа анализа данных
# Машинное обучение, часть 1
# Домашнее задание №3

Решите предложенные задачи. Решения необходимо оформить в виде PDF документа. Каждая задача должна быть подробно обоснована, задачи без обоснования не засчитываются. Решения пишутся в свободной форме, однако так, чтобы проверяющие смогли разобраться в нем. Если проверяющие не смогут разобраться в решении какой-нибудь задачи, то она автоматически не засчитывается.

**Задача 1 (1 балл)**. Определим понятие доли дефектных пар ответов классификатора. Пусть дан классификатор $a(x)$, который возвращает оценки принадлежности объектов классам: чем больше ответ классификатора, тем более он уверен в том, что данный объект относится к классу «+1». Отсортируем все объекты по неубыванию ответа классификатора $a$: $x_{(1)}, \dots, x_{(\ell)}$. Обозначим истинные ответы на этих объектах через $y_{(1)}, \dots, y_{(\ell)}$. Тогда доля дефектных пар записывается как

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

Как данный функционал связан с AUC-ROC (площадью под ROC-кривой)? Ожидается, что ответ будет дан в виде формулы, связывающей DP и AUC.

**Задача 2 (2 балла)**. Метод главных компонент (PCA) ранга $m$ находит малоранговое приближение матрицы объекты-признаки $F \approx GU^T$, где $F \in \mathbb{R}^{l\times n}$, $G \in \mathbb{R}^{l\times m}$, $U \in \mathbb{R}^{n\times m}$, которое является решением задачи:

$$\|F - GU^T\| \to \min_{G,U}.$$

Положим, что у нас есть алгоритм, реализующий PCA ранга $m = 1$. Результатом его работы PCA являются векторы $g_1 \in \mathbb{R}^l$ и $u_1 \in \mathbb{R}^n$, являющиеся решением задачи

$$\|F - g_1 u_1^T\| \to \min_{g_1,u_1}.$$

Давайте попробуем построить PCA ранга 2 используя лишь одноранговый алгоритм выше. Для этого рассмотрим одноранговый PCA разности $F - g_1 u_1^T$ и обозначим результат его работы за $g_2$ и $u_2$. Докажите, что матрицы $[g_1, g_2] \in \mathbb{R}^{l\times 2}$ и $[u_1, u_2] \in \mathbb{R}^{n\times 2}$ являются решением задачи PCA ранга 2. Здесь $[a, b]$ обозначает горизонтальную конкатенацию векторов $a$ и $b$. Также предложите способ построения PCA ранга $m$ с использованием только PCA ранга 1.

**Задача 3 (1 балл)** Верно ли, что после перехода в спрямляющее пространство гауссовского ядра $K(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|^2}{2}\right)$ качество классификации методом 1-NN может повыситься? Ответ необходимо пояснить. Считается, что 1-NN использует метрику, порожденную нормой в соответствующем пространстве.

**Задача 4 (2 балла)** . Костя участвует в конкурсе по анализу данных, в котором нужно решить задачу бинарной классификации функционалом ошибки Mean Absolute Error (MAE):

$$Q(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = \frac{1}{l}\sum_{i=1}^{l} |y_i - \tilde{y}_i|, \quad \boldsymbol{y} \in \{0,1\}^l, \quad \tilde{\boldsymbol{y}} \in [0,1]^l,$$

где $l$ — количество обучающих объектов, $\boldsymbol{y}$ — вектор истинных классов объектов, $\tilde{\boldsymbol{y}}$ — вектор предсказанных «степеней принадлежности» классу 1, качество которого и оценивается. Костя заметил, что качество предсказания на скрытой выборке, которая доступна только организаторам конкурса, всегда улучшается, если особым образом сдвинуть каждый прогноз $\tilde{y}_i$ в один из концов отрезка $[0,1]$ для каждого объекта $i$. Объясните, почему так происходит.

Этот факт показывает, что MAE является неудачным функционалом для оценки качества вектора промежуточных степеней принадлежности из $[0,1]$ в задачах бинарной классификации.

**Подсказка**: Сделайте предположение, что объект с номером $i$ принадлежит классу 1 с истинной вероятностью $p_i$.

**Задача 5 (1 балл)**. Рассмотрим модель временного ряда $\tilde{y}_t = c + \varepsilon_t$, где $\varepsilon_t$ – шум с центром в нуле и дисперсией $\sigma^2$. Покажите, как зависит дисперсия прогноза экспоненциальным сглаживанием от параметра сглаживания $\alpha$.

**Задача 6 (1 балла)** Для двух выборок ниже запишите уравнения разделяющих поверхностей, которые будут построены в результате обучения наивного байесовского классификатора с предположением, что значения каждого признака имеют нормальное распределение.

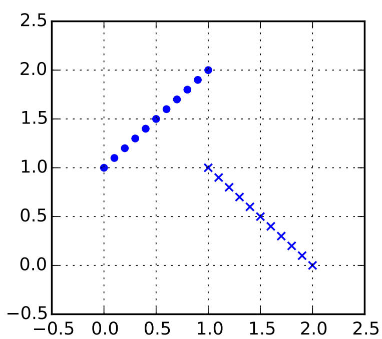


Рис. 1: Первая выборка в задаче 8

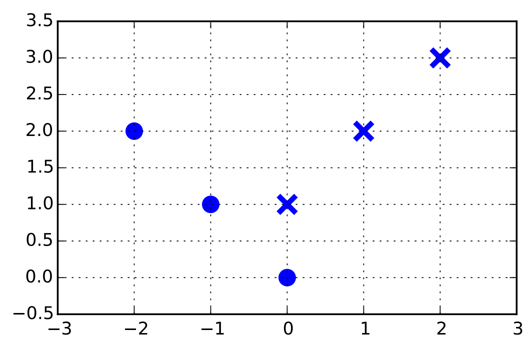


Рис. 2: Вторая выборка в задаче 8

**Задача 7 (1 балл)**. Доказать, что наивный байесовский классификатор в случае бинарных признаков $\xi_j \in \{0,1\}$ является линейным разделителем: $a(\xi_1, \ldots, \xi_n) = [\alpha_0 + \alpha_1\xi_1 + \ldots + \alpha_n\xi_n > 0]$. Выпишите формулы для вычисления коэффициентов $\alpha_j,\ j = 0, \ldots n$ по обучающей выборке.

**Задача 8 (1 балл)**. Дана выборка

$$X = \{\boldsymbol{x}_i\}_{i=1}^{l},\ \boldsymbol{x}_i \in \mathbb{R}^2,\ y_i \in Y = \{0,1\}.$$

Обозначим за $\lambda_y$ штраф за ошибку в классе $y$. Известно, что

$$\ln \lambda_y P(y) = C_y, \quad y \in Y.$$

Положим, что функции правдоподобия классов имеют гауссовские распределение со средними

$$\boldsymbol{\mu}_0 = \begin{pmatrix} a \\ b \end{pmatrix}, \ \boldsymbol{\mu}_1 = \begin{pmatrix} -a \\ -b \end{pmatrix}$$

соответственно и одинаковыми ковариационными матрицами

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & S \end{pmatrix}.$$

Выпишите байесовский алгоритм классификации и уравнение разделяющей поверхности.