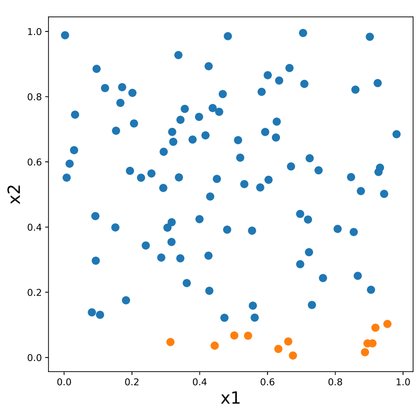
Машинное обучение 1
Контрольная работа
Вариант 0.1

Задача 1 (2.5 балла). Для начала напомним, как выглядят некоторые функции потерь для задачи классификации:

$$L_a(y, z) = \log\left(1 + \exp(-yz)\right); \quad L_b(y, z) = \max(0, 1 - yz); \quad L_c(y, z) = (yz - 1)^2.$$

Ответьте на вопросы о линейных методах (и не только):

1. Кратко опишите идею метода one-vs-all для многоклассовой классификации. Запишите формулу, по которой определяется объекта класс на основе выходов моделей.
2. Рассмотрим функции потерь L_a, L_b и L_c. Для каждой из них найдите минимальное по модулю значение z, при котором функция достигает своего минимума на объекте отрицательного класса ($y = -1$).
3. Допустим, мы обучили линейный классификатор $a(x) = \text{sign}\langle w, x\rangle$. Добавим теперь в эту модель порог: $\tilde{a}(x) = \text{sign}(\langle w, x\rangle - t)$. Положим $t = 10$ и рассмотрим только объекты положительного класса. Может ли быть такое, что из-за добавления порога не изменится значение функции потерь L_a ни на одном положительном объекте? А для L_b? Ответ обоснуйте.
4. Рассмотрим следующую выборку с бинарной целевой переменной:



   При обучении на ней SVM с гиперпараметрами по умолчанию ($C = 1$) получается модель, которая относит все объекты к синему классу. Из-за чего это может происходить? Аргументируйте свою гипотезу.

Задача 2 (2.5 балла). Ответьте на вопросы о решающих деревьях и композициях моделей:

1. Что такое out-of-bag оценка в бэггинге? Опишите её идею и объясните, для чего её можно использовать.
2. Рассмотрим следующий способ обучения базовой модели в градиентном бустинге для функции потерь $L(y, z)$:

$$\frac{1}{\ell}\sum_{i=1}^{\ell} L(s_i, b_N(x_i)) \to \min_{b_N}; \quad s_i = \left.\frac{\partial L(y_i, z)}{\partial z}\right|_{z=b_{N-1}(x_i)}$$

Найдите все ошибки в этих формулах. Объясните, почему это ошибки.

3. Мы решаем задачу регрессии. Допустим, у всех объектов обучающей выборки целевые переменные положительные. Могут ли быть отрицательные прогнозы у решающего дерева, обученного на этой выборке (если оно обучается на MSE)? А у отдельных деревьев в составе случайного леса? А у отдельных деревьев в составе градиентного бустинга? Ответ обоснуйте.

4. В xgboost при регуляризации деревьев используется норма весов ответов в листьях. Почему точно такая же регуляризация не подойдёт для одиночного решающего дерева в произвольной задаче? Ответ обоснуйте.

Задача 3 (2.5 балла). Вам могут пригодиться следующие формулы при решении этого задания:

$$L_\delta(y, a) = \begin{cases} \frac{1}{2}(y-a)^2, & |y-a| < \delta \\ \delta\left(|y-a| - \frac{1}{2}\delta\right), & |y-a| \geqslant \delta \end{cases}$$

$$L(y, a) = \log\cosh(a-y)$$

Ответьте на вопросы о функциях потерь:

1. Решаем задачу бинарной классификации. Допустим, мы хотим построить модель с маленькой долей ошибок на обучающей выборке. Саму модель обучаем на логистическую функцию потерь. Почему обычно делается именно так вместо того, чтобы модель обучать непосредственно через минимизацию доли ошибок?

2. Продолжим предыдущий пункт. Может ли быть так, что при очень малой длине шага (при которой минимум функционала ошибки мы точно не перескакиваем) на обучающей выборке доля ошибок и логистическая ошибка вырастут на одном и том же шаге? А так, что доля ошибок вырастет, а логистическая ошибка упадёт? Ответы обоснуйте.

3. Предложите функцию потерь $L(y, a)$ для регрессии, одновременно удовлетворяющую следующим требованиям: (а) штрафует за завышение прогноза сильнее, чем за занижение, (б) устойчива к выбросам (то есть её значение растёт линейно с ростом $|y-a|$ при больших значениях $|y-a|$), (в) всюду дифференцируема. Кратко обоснуйте, почему предложенная функция потерь подходит под эти условия.

4. На лекциях разбирались L_1- и L_2-нормы для регуляризации. Какой эффект по сравнению с ними будет иметь L_3-регуляризатор $\sum_{j=1}^d |w_j|^3$? А L_{inf}-регуляризатор $\max_{j=1,\dots,d} |w_j|$? А L_0-регуляризатор $\sum_{j=1}^d [w_j \neq 0]$? Обоснуйте ответы. Какие сложности при обучении на эти три регуляризатора могут возникнуть?

Задача 4 (2.5 балла). Пусть выборка $X = (x_i, y_i)_{i=1}^\ell$ генерируется из распределения $p(x, y)$ такого, что $x_i \sim \mathcal{N}(0, \sigma^2)$, $y_i = \varepsilon_i x_i \; \forall i = 1, \dots, \ell$, где $\varepsilon_i \sim \text{Exp}(\lambda)$, и все ε_i и x_i независимы в совокупности. Найдите шум, смещение и разброс для модели $\mu(X)(x) = \frac{1}{\lambda\ell}\left(\sum_{i=1}^\ell x_i\right)\text{sign}(x)$.

Вам могут пригодиться статистики экспоненциального распределения: $\mathbb{E}[\varepsilon_i] = \frac{1}{\lambda}, \mathbb{D}[\varepsilon_i] = \frac{1}{\lambda^2}$.