# Машинное обучение, DS-поток

## Домашнее задание 5

**Правила:**

- Дедлайн **20 марта 02:00**. После дедлайна работы не принимаются кроме случаев наличия уважительной причины.
- Выполненную работу нужно отправить на почту `mipt.stats@yandex.ru` , указав тему письма `"[ml] Фамилия Имя - задание 5"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `5.N.ipynb` и `5.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- Теоретические задачи необходимо оформить в tex/markdown или же прислать фотку в правильной ориентации рукописного решения, **где все четко видно**.
- Решения, размещенные на каких-либо интернет-ресурсах не принимаются. Кроме того, публикация решения в открытом доступе может быть приравнена к предоставлении возможности списать.
- Для выполнения задания используйте этот ноутбук в качествие основы, ничего не удаляя из него.
- Никакой код из данного задания при проверке запускаться не будет.

**Баллы за задание:**

- Задача 1 - 1 балл
- Задача 2 - 1 балл
- Задача 3 - 3 балла
- Задача 4 - 1 балл
- Задача 5 - 7 баллов

## Теория

Рассмотрим задачу бинарной классификации, причем $\mathscr{Y} = \{+1, -1\}$. Пусть так же $\widehat{y}$ --- некоторый классификатор, предсказывающий степень принадлежности классу. При этом решающее правило имеет вид $f(x) = \text{sign}\left(\widehat{y}(x)\right)$. Рассмотрим логистическую функцию потерь:

$$\mathcal{L}(y, z) = \log\left(1 + \exp(-yz)\right)$$

### Задача 1

Покажите, что задача минимизации функционала ошибки $Q(\widehat{y}) = \sum_{i=1}^{n} \mathcal{L}\left(Y_i, \widehat{y}(x_i)\right)$ для логистической функции потерь эквивалентна максимизации по $y$ функции правдоподобия в предположении $Y_i \sim Bern(\sigma(y(x)))$.

### Задача 2

Рассмотрим градиентный бустинг с логистической функцией потерь. Выпишите для градиентного спуска формулу для вектора сдвигов и задачу поиска новой базовой модели.

### Задача 3

Предположим, модель градиентного бустинга $\widehat{y}_{t-1}$ уже построена.

1. Выпишите вид функционала ошибки $Q(\widehat{y}_t) = \sum\limits_{i=1}^{n} \mathcal{L}\left(Y_i, \widehat{y}_t(x_i)\right)$ для логистической функции потерь. Одинаковый ли вклад вносят разные объекты в ошибку?
2. Выпишите формулу для вектора сдвигов. Как она выражается через отклики на объектах обучающей выборки? Одинаковый ли вклад вносят разные объекты в формирование вектора сдвигов?
3. На лекции было показано, что для экспоненциальной функции есть проблема: базовый классификатор может настраиваться только на шумовые объекты. Наблюдается ли такая проблема у логистической функции потерь?

### Задача 4

Рассмотрим градиентный бустинг над решающими деревьями. После построения дерева выполняется делать перенастройку в листьях дерева.

1. Выпишите оптимизационную задачу для коэффициентов $\gamma_{tj}$ --- новых ответов в листьях.
2. Решите полученную задачу сделав один шаг метода Ньютона из начального приближения $\gamma_{tj} = 0$, что соответствует отсутствию базовой модели $b_t$.

## Практика

### Задача 5

*Внимание!* Перед выполнением задачи прочитайте полностью условие. В задаче используются смеси различных моделей с разными гиперпараметрами. Подумайте над тем, какой гиперпараметр как подбирать и на каком множестве. Не забудьте, что на тестовой выборке, по которой делаются итоговые выводы, ничего не должно обучаться.

**1.** Повторите исследование, проведенное в задаче 2 предыдущего домашнего задания, используя градиентный бустинг из `sklearn`. Сравните полученные результаты со случайным лесом. Детали:

- В качестве основы можно использовать как свое решение предыдущего задания, так и выложенное на Вики. В большинстве случаев нужно только заменить `RandomForestRegressor` на `GradientBoostingRegressor`.
- У градиентного бустинга есть также важный гиперпараметр `learning_rate`. Поясните его смысл и проведите аналогичные исследования.
- При сравнении методов по одинаковым свойствам желательно рисовать результаты на одном графике.
- Обратите внимание на метод `staged_predict` у `GradientBoostingRegressor`. Он позволяет получить "кумулятивные" предсказания, то есть по первым $t$ деревьям по всем значениям $t$.
- При кросс-валидации проводите достаточное количество итераций рандомизированного поиска (при $\geqslant 2$ параметров) на большой сетке параметров. Даже если долго обучается.

**2.** Выберите самый значимый признак согласно `feature_importances_` и визуализируйте работу первых 10 деревьев на графиках зависимости таргета от этого признака. Пример графиков смотрите в лекции.

**3.** Обучите градиентный бустинг на решающих деревьях, у которого в качестве инициализирующей модели используется линейная регрессия. Для этого используйте класс `GradientBoostingRegressor`, которому при инициализации в качестве параметра `init` передайте

модель ридж-регрессии Ridge, которая должна быть инициализирована, но необучена. Подберите оптимальные гиперпараметры такой композиции. Как вы будете подбирать гиперпараметр ридж-регрессии? Улучшилось ли качество модели на тестовой выборке?

**4.** Рассмотрим модели смеси градиентного бустинга $\widehat{y}_{gb}$ и случайного леса $\widehat{y}_{rf}$ в виде

$$\widehat{y}(x) = w\widehat{y}_{gb}(x) + (1 - w)\widehat{y}_{rf}(x),$$

где $w \in [0, 1]$ --- коэффициент усреднения. Подберите оптимальное значение гиперпараметра $w$. Удалось ли добиться улучшения качества на тестовой выборке?