Машинное обучение, DS-поток

Домашнее задание 4

Правила:

- Дедлайн **13 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 4"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `4.N.ipynb` и `4.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- Теоретические задачи необходимо оформить в 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 качестве основы, ничего не удаляя из него.
- Никакой код из данного задания при проверке запускаться не будет.

Баллы за задание:

- Задача 1 - 3 балла
- Задача 2 - 7 баллов
- Задача 3 - 3 балла

In [ ]:

```
import numpy as np
import pandas as pd
import scipy.stats as sps

from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import BaggingRegressor
from sklearn.linear_model import Ridge
from sklearn.datasets import fetch_california_housing

from sklearn.metrics import mean_squared_error as mse
from sklearn.model_selection import GridSearchCV
from sklearn.model_selection import train_test_split

import matplotlib.pyplot as plt
import seaborn as sns
sns.set(font_scale=1.5)
```

Задача 1

Пусть обучающая выборка $(X_1, Y_1), \ldots, (X_n, Y_n)$ такова, что

- объекты $X_1, \ldots, X_n$ одномерны и имеют распределение $\mathcal{N}(0, \sigma^2)$.

- отклик получается по правилу $Y_i = X_i^2 + \varepsilon_i$, где ε_i независимы, имеют нулевое среднее и не зависят от X_i.

Пусть также для объекта X отклик Y получен по аналогичному правилу, причем X и Y не зависят от обучающей выборки.

Для МНК-модели $\widehat{y}(x) = \widehat{\theta}x$ выпишите bias-variance разложение. Компонентну, отвечающую за разброс, разрешается не доводить до конца, как это было сделано на семинаре.

Задача 2

В этой задаче вам предлагается исследовать зависимость качества предсказаний модели случайного леса в зависимости от различных гиперпараметров на примере задаче регрессии. Используйте класс `RandomForestRegressor` библиотеки `sklearn`.

В качестве данных возьмём датасет `california_housing` из библиотеки `sklearn` о стоимости недвижимости в различных округах Калифорнии. Этот датасет состоит из 20640 записей и содержит следующие признаки для каждого округа: `MedInc`, `HouseAge`, `AveRooms`, `AveBedrms`, `Population`, `AveOccup`, `Latitude`, `Longitude`. `HouseAge` и `Population` - целочисленные признаки. Остальные признаки - вещественные.

Совет. При отладке кода используйте небольшую часть данных. Финальные вычисления проведите на полных данных. Для оценки времени работы используйте `tqdm` в циклах.

In [ ]:

```
housing = fetch_california_housing()
X, y = housing.data, housing.target
```

Разбейте данные на обучающую выборку и на валидацию, выделив на валидацию 25% данных.

In [ ]:

```
X_train, X_test, y_train, y_test = <...>
```

Посмотрите, как изменяется качество леса в зависимости от выбранных параметров. Для этого постройте графики зависимости MSE на тестовой выборке от количества деревьев (от 1 до 100) и от максимальной глубины дерева (от 3 до 25). Когда варьируете один из параметров, в качестве другого берите значение по умолчанию.

In [ ]:

```
<...>
```

Основываясь на полученных графиках, ответьте на следующие вопросы.

1. Какие закономерности можно увидеть на построенных графиках? Почему графики получились такими?
2. Как изменяется качество предсказаний с увеличением исследуемых параметров при достаточно больших значениях параметров?

3. В предыдущем задании вы на практике убедились, что решающее дерево начинает переобучаться при достаточно больших значениях максимальной глубины. Справедливо ли это утверждение для решающего леса? Поясните свой ответ, опираясь на своё знание статистики.

Ответ: <...>

Обучите случайный лес с параметрами по умолчанию и выведите mse на тестовой выборке. Проведите эксперимент 3 раза. Почему результаты отличаются?

In [ ]:

```
<...>
```

Ответ: <...>

Было бы неплохо определиться с тем, какое количество деревьев нужно использовать и какой максимальной глубины они будут. Подберите оптимальные значения `max_depth` и `n_estimators` с помощью кросс-валидации.

In [ ]:

```
<...>
```

Выведите найденные оптимальные параметры.

In [ ]:

```
<...>
```

Зафиксируем эти оптимальные значения параметров и в дальнейшим будем их использовать.

In [ ]:

```
max_depth = <...>
n_estimators = <...>
```

Оценим качество предсказаний обученного решающего леса.

In [ ]:

```
<...>
```

Исследуйте зависимость метрики `mse` от количества признаков, по которым происходит разбиение в вершине дерева. Поскольку количество признаков в датасете не очень большое (их 8), то можно перебрать все возможные варианты количества признаков, использующихся при разбиении вершин.

Не забывайте делать пояснения и выводы!

In [ ]:

```
<...>
```

Постройте график зависимости метрики mse на `test` и `train` в зависимости от числа признаков, использующихся при разбиении в каждой вершине.

In [ ]:

```
1  <...>
```

Почему график получился таким? Как зависит разнообразие деревьев от величины `n_features` ?

Ответ: <...>

Проведите эксперимент, в котором выясните, как изменится качество регрессии, если набор признаков, по которым происходит разбиение в каждой вершине определяется не заново в каждой вершине, а задан заранее. Поскольку результаты эксперимента могут сильно зависеть от того, какой набор признаков задан изначально, проведите несколько экспериментов для каждого значения `n_features` .

Для реализации данного эксперимента используйте класс беггинг-модели `sklearn.ensemble.BaggingRegressor` , у которого используйте следующие поля:

- `base_estimator` -- базовая модель, используйте `sklearn.tree.DecisionTreeRegressor`
- `max_features` -- количество признаков для каждой базовой модели
- `n_estimators` -- количество базовых моделей.

Постройте графики `mse` на обучающей и на валидационной выборке.

In [ ]:

```
1  <...>
```

Сравните результаты обычного случайного леса с только что построенным лесом.

Сделайте выводы. Объясните, чем плох такой подход пострения случайного леса. Какое преимущество мы получаем, когда выбираем случайное подмножество признаков в каждой вершине в обычном случайном лесу?

Вывод.

<...>

Поясните разницу между следующими конструкциями:

```
BaggingRegressor(base_estimator=DecisionTreeRegressor(),
max_features=n_features)
```

```
BaggingRegressor(base_estimator=DecisionTreeRegressor(max_features=n_features))
```

<...>

Задача 3

На лекции получена формула bias-variance разложения для беггинга. Проведите эксперимент, в котором выясните, насколько уменьшается разброс (variance-компонента) беггинг-модели на 100 базовых моделях по отношению к одной базовой модели. Используйте данные из предыдущей задачи. Рассмотрите беггинг на следующих базовых моделях:

- решающие деревья, можно использовать вариант случайного леса.
- ридж-регрессия.

Для решения задачи потребуется оценить корреляции предсказаний на тестовой выборке базовых моделей, входящих в состав беггинг-модели. Их можно получить с помощью поля `estimators_` у обученной беггинг-модели.

Насколько уменьшается разброс в каждом случае? Для каждого случая постройте также матрицу корреляций предсказаний базовых моделей и гистограмму по ним. Какую оценку коэффициента корреляции вы используете и почему?