Машинное обучение, DS-поток

Домашнее задание 5

Правила:

- Дедлайн **24 марта 23:59**. 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 5A"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `5A.N.ipynb` и `5A.N.pdf` , где `N` - ваш номер из таблицы с оценками.
- 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 - 6 баллов

Задача 1

Будем работать с датасетом **"bikes_rent.csv"**, в котором по дням записаны календарная информация и погодные условия, характеризующие автоматизированные пункты проката велосипедов, а также число прокатов в этот день. Последнее мы будем предсказывать; таким образом, мы будем решать задачу регрессии.

Данные предоставлены компанией capital bikeshare.

Для каждого дня проката известны следующие признаки (как они были указаны в источнике данных):

- `_season_` : 1 - весна, 2 - лето, 3 - осень, 4 - зима
- `_yr_` : 0 - 2011, 1 - 2012
- `_mnth_` : от 1 до 12
- `_holiday_` : 0 - нет праздника, 1 - есть праздник
- `_weekday_` : от 0 до 6
- `_workingday_` : 0 - нерабочий день, 1 - рабочий день
- `_weathersit_` : оценка благоприятности погоды от 1 (чистый, ясный день) до 4 (ливень, туман)
- `_temp_` : температура в Цельсиях
- `_atemp_` : температура по ощущениям в Цельсиях
- `_hum_` : влажность
- `_windspeed(mph)_` : скорость ветра в милях в час
- `_windspeed(ms)_` : скорость ветра в метрах в секунду
- `_cnt_` : количество арендованных велосипедов (это целевой признак, его мы будем предсказывать)

Считайте данные и разделите на обучение и тест.

In [ ]:

```
1
```

Посмотрите на графиках, как целевой признак зависит от остальных и поймите какой характер зависимости целевой переменной от остальных.

In [ ]:

```
1
```

Теперь посмотрите на среднее значение каждого признака. Что можно сказать? Какая тут проблема?

In [ ]:

```
1
```

Поняв проблему, исправьте ее.

In [ ]:

```
1
```

Обучите линейную регрессию на наших данных и посмотрите на веса признаков. Что в них не так? Почему так получилось? Какая здесь проблема и как ее можно решить?

In [ ]:

```
1
```

Решите проблему, обучите линейную модель и снова посмотрите на веса? Стало ли лучше?

In [ ]:

```
1
```

Обучите теперь `Lasso` и подберите оптимальный параметр `alpha` для него с помощью кросс-валидации. Метрика --- MSE. Возьмите `alpha` от 0 до 100. Для проведения кросс-валидации разделите выборку на 3 части и проведите 3 итерации для разных частей: для каждого `alpha` обучите `Lasso(alpha)` на двух частях и посмотрите на MSE на третей части. Визуализируйте 3 полученных графика завасимости MSE от `alpha` на разных данных. Сделайте выводы.

In [ ]:

```
1
```

Найдите оптимальное `alpha`

In [ ]:

```
1
```

Теперь проявите фантазию. Предобработайте данные, в том числе можно добавить какие-то признаки

или, наоборот, убрать. Рассмотрите разные модели: `Linear` , `Lasso` , `Ridge` . Выберите из них наилучшую, подобрав оптимальные параметры. Хочется получить хороший mse на тестовой выборке.

In [ ]:

```
1
```