Машинное обучение, DS-поток

Домашнее задание 1

Правила:

- Дедлайн **21 декабря 10:00**. 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 1"` . Квадратные скобки обязательны. Если письмо дошло, придет ответ от автоответчика.
- Прислать нужно ноутбук и его pdf-версию (без архивов). Названия файлов должны быть такими: `1.N.ipynb` и `1.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 - 5 баллов
- Задача 2 - 5 баллов
- Задача 3 - 10 баллов

Задача 1.

Пусть $\widehat{\theta}$ --- оценка коэффициентов линейной модели в методе ридж-регрессии.

- Посчитайте $MSE_{\widehat{\theta}}(\theta) = \mathsf{E}_{\theta}\left(\widehat{\theta} - \theta\right)^T \left(\widehat{\theta} - \theta\right)$.
- Покажите, что в отличии от МНК вектор оценок отклика $\widehat{Y}$ на обучающей выборке в методе ридж-регрессии не перпендикулярен остаткам модели $\widehat{e} = Y - \widehat{Y}$.

Задача 2.

Выведите итерационную формулу пересчета коэффициентов модели с помощью формулы решения через проксимальный оператор для случая ридж-регрессии. Какой вы можете видеть эффект при изменении параметра регуляризации и в чем его отличие от лассо-регрессии?

Задача 3.

Вам предлагается изучить и сравнить свойства линейных регрессионных моделей: обычной и с регуляризациями -- Lasso, Ridge, Elastic Net.
При выполнении задания воспользуйтесь готовыми реализациями методов в sklearn.

Скачайте данные `cost of living 2018` (https://dasl.datadescription.com/datafile/cost-of-living-2018/?_sfm_cases=539+541), в которых используйте следующие столбцы:

- `Cost of Living Index` --- является относительным показателем цен на потребительские товары, включая продукты, рестораны, транспорт и коммунальные услуги. Cost of Living Index не включает расходы на проживание, такие как аренда или ипотека. Если город имеет индекс стоимости жизни 120, это означает, что Numbeo оценивает его на 20% дороже, чем Нью-Йорк.
- `Rent Index` --- это оценка цен на аренду квартир в городе по сравнению с Нью-Йорком. Если индекс арендной платы равен 80, Numbeo оценивает, что цена аренды в этом городе в среднем на 20% меньше, чем цена в Нью-Йорке.
- `Cost of Living Plus Rent Index` --- это оценка цен на потребительские товары, включая арендную плату, по сравнению с Нью-Йорком.
- `Restaurant Price Index` --- сравнение цен на блюда и напитки в ресторанах и барах по сравнению с Нью-Йорк.
- `Local Purchasing Power Index` --- показывает относительную покупательную способность при покупке товаров и услуг в данном городе за среднюю заработную плату в этом городе. Если внутренняя покупательная способность составляет 40, это означает, что жители этого города со средней зарплатой могут позволить себе покупать в среднем на 60% меньше товаров и услуг, чем жители Нью-Йорка со средней зарплатой по Нью-Йорку.
- `Groceries Index` --- это оценка цен на продукты в городе по сравнению с Нью-Йорком. Для расчета этого раздела Number использует веса товаров в разделе "Рынки" для каждого города.

In [ ]:

```
data = pd.read_csv('cost-of-living-2018.txt', sep='\t')
data = data[[
    'Cost of Living Index',
    'Rent Index',
    'Cost of Living Plus Rent Index',
    'Restaurant Price Index',
    'Local Purchasing Power Index',
    'Groceries Index'
]]
data.head()
```

1. Задача заключается в построении предсказания `Groceries Index` по известным значениям остальных параметров. Разделите данные на признаки X и таргет y.

In [ ]:

```

```

Разбейте данные на обучающую и тестирующие выборки в соотношении 7:3 с помощью `train_test_split` из sklearn. Далее везде вплоть до сравнения моделей используйте обучающую выборку.

In [ ]:

```

```

Методы с регуляризацией требуют страндартизацию признаков. Поясните, почему это необходимо.

<...>

Примените стандартизацию к данным обучающей выборке, используя класс `StandardScaler`

(https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler).

In [ ]:

```
1
```

2. Исследуйте зависимость значений коэффициентов от параметра регуляризации `alpha` для Ridge, Lasso, Elastic регрессии. Для Elastic также исследуйте зависимость от параметра `l1_ratio` . Нарисуйте графики, используя код с семинара. Сделайте предположение об оптимальном значении параметров.

In [ ]:

```
1
```

Расчитайте индекс обусловленности для случая линейной регрессии. Можно ли сделать вывод о мультиколлинеарности данных?
Нарисуйте график зависимость индекса обусловленности от параметра регуляризации для Ridge-регрессии.

In [ ]:

```
1
```

3. С помощью кросс-валидации определите наилучшие параметры для Ridge, Lasso, Elastic моделей. В качестве метрики качества используйте среднеквадратичную ошибку (MSE).

In [ ]:

```
1
```

На тестовой части данных сравните качество моделей с оптимальными параметрами. Какая модель дала лучший результат?

In [ ]:

```
1
```

4. Исследуйте остатки модели Ridge-регрессии. Можно ли говорить о гомоскедастичности. Если нет, попытайтесь несложными преобразованиями признаков и отклика визуально прийти к гомоскедастичности.

In [ ]:

```
1
```

С помощью модели Ridge-регрессии постройте предсказательный интревал для наблюдаемого отклика уровня доверия 0.95. Какой смысл имеет этот интервал? В чем его отличие от доверительного интервала? Посчитате долю точек выходящих за предсказательный интревал.

In [ ]:

```
1
```

5. Сделайте общий вывод по задаче.

In [ ]:

```
1
```