Предобработка данных в Pandas

Данное задание направлено на знакомство с инструментарием, который пригодится в дальнейших практических заданиях.

Вы научитесь:

- работать с данными используя язык Python и пакет Pandas
- делать предобработку данных
- находить простые закономерности в данных

Введение

Сейчас Python является одним из наиболее распространенных языков программирования. Одним из его преимуществ является большое количество пакетов, решающих самые разные задачи. В нашем курсе мы рекомендуем использовать библиотеки Pandas, NumPy и SciPy, которые существенно упрощают чтение, хранение и обработку данных. В дальнейших работах вы также познакомитесь с пакетом Scikit-Learn, в котором реализованы многие алгоритмы машинного обучения.

Начало работы

Для того, чтобы начать работать с данными, необходимо сначала загрузить их из файла. В данном задании мы будем работать с данными в формате CSV, предназначенном для хранения табличных данных: столбцы разделяются запятой, первая строка содержит имена столбцов.

Пример загрузки данных в Pandas:

```
import pandas
data = pandas.read_csv('titanic.csv', index_col='PassengerId')
```

Данные будут загружены в виде DataFrame, с помощью которого можно удобно работать с ними. В данном случае параметр

```
index_col='PassengerId'
```

означает, что колонка PassengerId задает нумерацию строк данного датафрейма.

Для того, чтобы посмотреть что представляют из себя данные, можно воспользоваться несколькими способами:

более привычным с точки зрения Python (если индекс указывается только один, то производится выбор строк):

```
data[:10]
```

или же воспользоваться методом датафрейма:

```
data.head()
```

Один из способов доступа к столбцам датафрейма — использовать квадратные скобки и название столбца:

```
data['Pclass']
```

Для подсчета некоторых статистик (количества, среднее, максимум, минимум) можно также использовать методы датафрейма:

```
data['Pclass'].value_counts()
```

Более подробно со списком методов датафрейма можно познакомиться в документации.

Материалы

Набор данных взят с сайта Kaggle: Titanic: Machine Learning from Disaster.

Инструкция по выполнению

Загрузите датасет titanic.csv и, используя описанные выше способы работы с данными, найдите ответы на вопросы:

1. Какое количество мужчин и женщин ехало на корабле? В качестве ответа приведите два числа через пробел.

2. Какой части пассажиров удалось выжить? Посчитайте долю выживших пассажиров. Ответ приведите в процентах (знак процента не нужен).

3. Какую долю пассажиры первого класса составляли среди всех пассажиров? Ответ приведите в процентах (знак процента не нужен).

4. Какого возраста были пассажиры? Посчитайте среднее и медиану возраста пассажиров. В качестве ответа приведите два числа через пробел.

5. Коррелируют ли число братьев/сестер с числом родителей/детей? Посчитайте корреляцию Пирсона между признаками SibSp и Parch.

6. Какое самое популярное женское имя на корабле? Извлеките из полного имени пассажира (колонка Name) его личное имя (First Name). Это задание — типичный пример того, с чем сталкивается специалист по анализу данных. Данные очень разнородные и шумные, но из них требуется извлечь необходимую информацию. Попробуйте вручную разобрать несколько значений столбца Name и выработать правило для извлечения имен, а также разделения их на женские и мужские.

При необходимости округляйте ответ до двух знаков после запятой.

Ответ на каждое задание — текстовый файл, содержащий ответ в первой строчке. Обратите внимание, что отправляемые файлы не должны содержать пустую строку в конце. Данный нюанс является ограничением платформы Coursera. Мы работаем над тем, чтобы убрать это ограничение.