POLITECHNIKA POZNAŃSKA

Platformy Programowania: Laboratorium nr 9

Uczenie maszynowe z scikit-learn

Studia stacjonarne

2020-2021

dr inż. Jarosław Bąk
dr inż. Michał Ciesielczyk
mgr inż. Michał Blinkiewicz

Wprowadzenie

Wymagania

- Python (3.5 lub nowszy) – do pobrania np. tutaj: WinPython.
- Zintegrowane środowisko programistyczne dla Pythona, np.:
  - IDLE (dołączone do dystrybucji WinPython),
  - Spyder (dołączone do dystrybucji WinPython), lub
  - JetBrains PyCharm.

Materiały

- 10 Minutes to pandas
- An introduction to machine learning with scikit-learn

Instrukcja

Twoim dzisiejszym zadaniem będzie przewidywanie cen nieruchomości z wykorzystaniem algorytmu regresji liniowej na zbiorze danych Boston Housing Dataset.

Zbiór danych zawiera ceny nieruchomości z różnych lokalizacji w Bostonie. Każdy rekord obok samej ceny nieruchomości posiada 13 dodatkowych atrybutów takich jak średnia liczba pokoi, poziom przestępczości, czy wiek mieszkańców. Pełen opis zbioru danych jest dostępny pod adresem `https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.names`.

Ponieważ zbiór danych jest częścią biblioteki scikit-learn, można go zaimportować bezpośrednio z wykorzystaniem funkcji `load_boston`.

Zaimportuj do swojego skryptu następujące moduły:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

Jeśli korzystasz ze środowiska Jupyter Notebook, możesz również dodać instrukcję:

```
%matplotlib inline
```

dzięki której generowane wykresy będą wyświetlane pod kodem źródłowym.

Zadania

Zadanie 1 (2 pkt)

Wczytaj zbiór danych Boston Housing DataSet z biblioteki scikit-learn:

```
from sklearn.datasets import load_boston
boston_dataset = load_boston()
```

Załadowane dane znajdują się wewnątrz struktury przypominającej słownik. Wyświetl nazwy poszczególnych kluczy:

```
print(boston_dataset.keys())
```

Surowe dane znajdują się w polu `data`, natomiast nazwy poszczególnych kolumn w polu `feature_names`. Utwórz na ich podstawie obiekt typu `pandas.DataFrame`, a następnie wyświetl 10 pierwszych i 10 ostatnich rekordów.

Zauważ, że w utworzonym obiekcie reprezentującym dane brakuje cen stanowiących cel predykcji. Aby to zmienić, utwórz nową kolumnę o nazwie `'MEDV'` z danymi z pola `target` z wczytanego wcześniej zbioru danych (`boston_dataset`).

Wskazówka Do pobierania pierwszych lub ostatnich rekordów z obiektu skorzystaj odpowiednio z funkcji `DataFrame.head` i `DataFrame.tail`.

Zadanie 2 (1 pkt)

Po wczytaniu plików z danymi, nadszedł czas na ich analizę. Skorzystaj z funkcji `DataFrame.info` i odpowiedz na następujące pytania:

a) Ile jest próbek/obserwacji w obu zbiorach?
b) Jakiego typu są dane w poszczególnych kolumnach?
c) Czy w zbiorze danych są kolumny zawierające puste (brakujące) wartości?

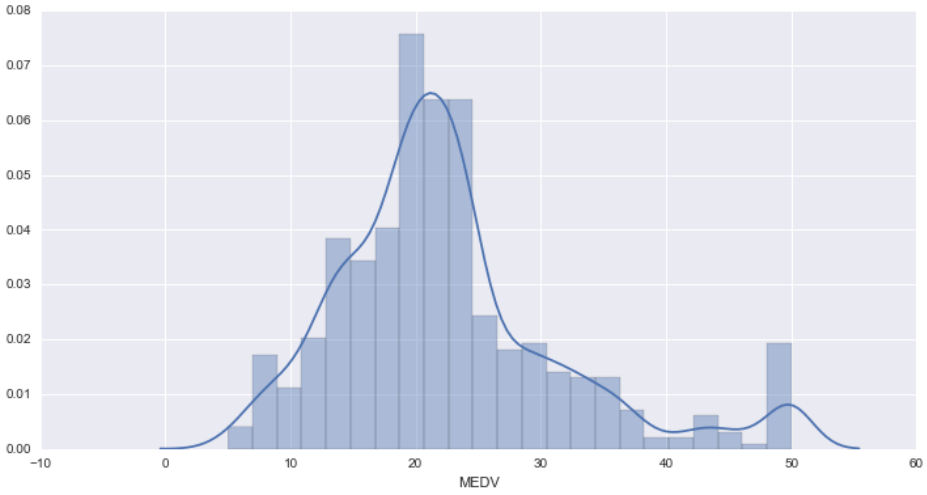
Zadanie 3 (1 pkt)

Wyświetl podstawowe statystki dla kolumn numerycznych korzystając z funkcji `DataFrame.describe` i odpowiedz na poniższe pytania:

a) Ile wynosi średni współczynnik przestępczości i jakie jest jego odchylenie standardowe (kolumna *CRIM*)?
b) Jaka jest maksymalna i minimalna cena nieruchomości (kolumna *MEDV*)?
c) Jaka jest mediana osób o niższym statusie społecznym (kolumna *LSTAT*)?

Zadanie 4 (1 pkt)

Wyświetl histogram przedstawiający rozkład wartości w kolumnie zawierającej ceny nieruchomości. Na przykład:



Wskazówka 1 Skorzystaj z funkcji `seaborn.distplot`.

Wskazówka 2 Jeśli nie korzystasz z interaktywnego interpretera Python aby wyświetlić wykres konieczne może być wywołanie funkcji `plt.show()`.

Zadanie 5 (2 pkt)

Przeanalizuj korelacje pomiędzy poszczególnymi atrybutami w danych. W tym celu wyznacz macierz korelacji korzystając z funkcji `DataFrame.corr` . Następnie wyświetl wyznaczoną macierz na wykresie typu heatmap (skorzystaj z `seaborn.heatmap`) razem z wartościami zaokrąglonymi do 2 miejsc po przecinku. Odpowiedz na poniższe pytania:

a) Które atrybuty są mocno skorelowane z ceną (kolumna *MEDV*)
b) Który atrybut można uznać za niepowiązany z ceną (kolumna *MEDV*)
c) Czy w danych istnieją atrybuty, których współczynnik korelacji wynosi więcej niż 0.9?

Wyświetl, przy pomocy wykresów punktowych, zależność pomiędzy cenami nieruchomości a wartościami kolumny:

a) dodatnio skorelowanej z cenami,
b) ujemnie skorelowanej z cenami, oraz
c) najmniej skorelowanej z cenami.

Jak myślisz, które z tych cech będą bardziej przydatne podczas predykcji cen?

Wskazówka Do generowania wykresów punktowych możesz skorzystać z `seaborn.regplot` .

Dodatkowe informacje:

- Współczynnik korelacji

Zadanie 6 (1 pkt)

Przygotuj dane do uczenia maszynowego:

a) macierz obserwacji `X`, oraz

b) tablicę etykiet docelowych `y`.

Do macierzy `X` wybierz z danych wszystkie kolumny, które uważasz, że mogą być przydatne podczas predykcji cen (oprócz samych cen oczywiście). Do zmiennej `y` przypisz kolumnę z cenami.

Podziel wszystkie dane na część uczącą oraz testową z wykorzystaniem funkcji `sklearn.model_selection.train_test_split` . Odsetek danych przeznaczonych do testowania modelu możesz ustawić na 0.2 (parametr `test_size`).

Wskazówka Aby wybrać podzbiór kolumn z obiektu `DataFrame` możesz skorzystać ze składni `__getitem__` (korzystając z operatora `[]`), np.: `df[['a','b']]`

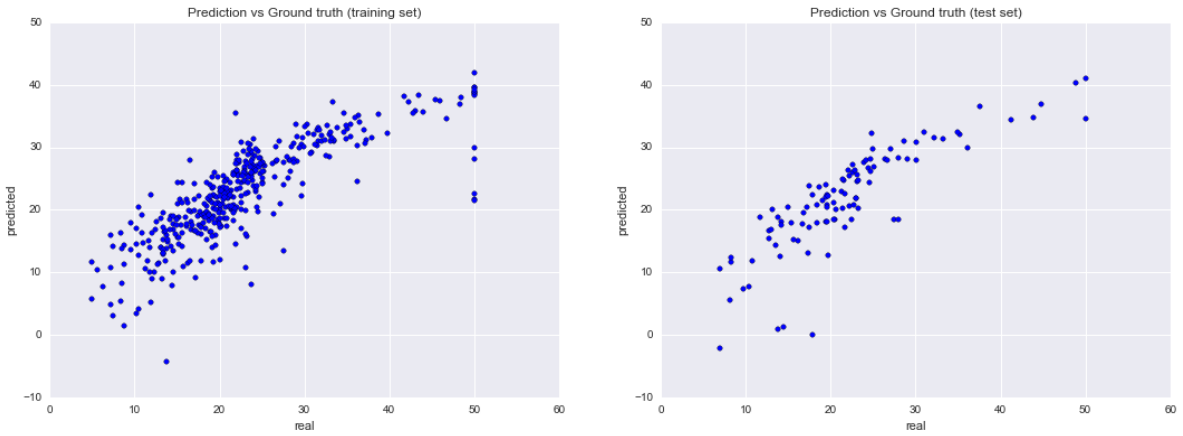
Zadanie 7 (2 pkt)

Utwórz model `LinearRegression` i naucz go danymi uczącymi (tj. zbiorem treningowym). Zwróć uwagę, że wszystkie parametry algorytmu `LinearRegression` są ustawiane w konstruktorze (każdy ma jednak wartości domyślne, z których możesz skorzystać).

Przy pomocy nauczonego modelu spróbuj odtworzyć wartości ze zbioru treningowego oraz przewidzieć wartości ze zbioru testowego. Sprawdź uzyskane wyniki wizualizując dane na wykresach punktowych przedstawiających zależność:

a) wartości odtworzonych od cen rzeczywistych ze zbioru treningowego,

b) predykcji od cen rzeczywistych ze zbioru testowego.



Rysunek 1: Wykresy przedstawiające zależność predykcji od wartości rzeczywistych, wygenerowane przy pomocy `pyplot.scatter` .

Dodatkowe informacje:

- An introduction to machine learning with scikit-learn: Learning and predicting

Zadanie 8 (2 pkt)

Policz metryki RMSE oraz MAE dla:

a) wartości odtworzonych ze zbioru treningowego, oraz
b) predykcji na zbiorze testowym

Skorzystaj z gotowych funkcji z biblioteki scikit-learn: `sklearn.metrics.mean_squared_error` oraz `sklearn.metrics.mean_absolute_error` .

Dodatkowe informacje:

- Model evaluation: quantifying the quality of predictions. Regression metrics

Na następne zajęcia

- LaTeX- system składu tekstu
- Nie za krótkie wprowadzenie do systemu LaTeX2ε