Zaawansowane metody uczenia maszynowego

Wyklad 1

Mateusz Serocki

Politechnika Gdańska

October 18, 2023

Agenda

Kilka słow o mnie

- Aktualnie starszy inżynier uczenia maszynowego w NIKE
- Main skillset: Python/AWS
- 5.5 roku doświadczenia zawodowego
- Main topics: Regresja, Klasyfikacja, Prognoza, MLOps
- LinkedIn: https://www.linkedin.com/in/mateuszserocki/
- Email: Mateusz.Serockiii@gmail.com

Reduckja wymiarowości

Definicja

Redukcja wymiarowości to transformacja danych z przestrzeni wielowymiarowej do przestrzeni niskowymiarowej, tak aby reprezentacja niskowymiarowa zachowała pewne znaczace właściwości oryginalnych danych, idealnie zbliżone do ich wewnetrznego wymiaru.

Interpretacja słowna

Innymi slowym staramy sie zmniejszyć wymiarowość naszego zbioru danych przy równoczesnym zachowaniu maksymalnej ilości informacji jaka z tego zbioru pochodzi.

Przykład redukcji wymiarow

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	target
0	5.1	3.5	1.4	0.2	0.0
1	4.9	3.0	1.4	0.2	0.0
2	4.7	3.2	1.3	0.2	0.0
3	4.6	3.1	1.5	0.2	0.0
4	5.0	3.6	1.4	0.2	0.0
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	2.0
146	6.3	2.5	5.0	1.9	2.0
147	6.5	3.0	5.2	2.0	2.0
148	6.2	3.4	5.4	2.3	2.0
149	5.9	3.0	5.1	1.8	2.0

Figure: Full dataset

	feature1	feature2	target
0	5.80	5.20	0.0
1	5.60	5.00	0.0
2	5.35	4.80	0.0
3	5.35	4.70	0.0
4	5.70	5.10	0.0
...	...	...	...
145	9.30	7.85	2.0
146	8.80	7.25	2.0
147	9.10	7.50	2.0
148	8.90	7.35	2.0
149	8.45	6.80	2.0

150 rows × 3 columns

Figure: Reduced dataset

Wady redukcji wymiarów

Redukcja wymiarów ma też negatywne skutki takie jak:

- Zmniejszenie ilości informacji
- Ryzyko usuniecia potencjalnie informatywnej zmiennej
- Calkowita lub cześciowa utrata interpretowalności zmiennych
- Kolejny krok potrzebny do przygotowania danych

Zalety redukcji wymiarów

Pozytywne aspekty redukcji wymiarów to:

- Redukcja szumu pochodzacego ze zmiennych
- Brak zmiennych które nie maja uzasadnionego wpływu na wynik
- Mniejsza moc obliczeniowa potrzebna do utworzenia modelu
- Mozemy zrezygnowac z pobierania niektorych zmiennych jezeli na tym etapie sadzimy ze sa nieistotne

Wariancja

Teoria

W teorii prawdopodobieństwa i statystyce wariancja jest kwadratem odchylenia od średniej zmiennej losowej. Wariancje czesto definiuje sie także jako kwadrat odchylenia standardowego.

Zastosowanie

W praktyce wariancja odzwierciedla zmienność danej cechy, niska wariancja może nieść za soba niska informatywność. Warto roważyć usuniecie takiej cechy jeżeli nasz zbiór jest bardzo duży. Koniecznie musimy sprawdzić wpływ tego usuniecia na wynik modelu.

Korelacja

Typy korelacji

- Pearson - standardowa korelacja liniowa
- Spearman - korelacja rankingowa
- Kendal - korelacja rankingowa dla grup z powtarzajacymi sie wartosciami
- more...

Przykład liczenia korelacji Pearsona

Link

Teoria dot. korelacji spearmana

Link

Teoria dot. korelacji Kendalla

Link

Biblioteka python do analizy zbioru danych

Pandas profiler

Biblioteka służy do szybkiej analizy danych, pomaga nam w zauważeniu potencjalnych problemów wenatrz naszych danych, oraz może służyć do wstepnej analizy naszego zbioru

```
import pandas as pd
data = pd.read_csv('path_to_your_data')
profile = data.profile_report(
  title = "Pandas Profiling Report",
  correlations = {
    "pearson": {"calculate": True},
    "spearman": {"calculate": True},
    "kendall": {"calculate": True}
  },
)
profile.to_file("pandas_profiled.html")
```

Braki danych

Sposoby na radzenie sobie z brakiem danych

- Usuniecie
- Inputacja
- Podstawienie

Przyklad 1

Przykładem algorytmu który radzi sobie z brakami danych jest XGBoost który wykorzystuje średnia wartość danej zmiennej zamiast wartości NULL (brak)

Przyklad 2

Przykładem algorytmu który nie radzi sobie z brakami danych jest regresja liniowa (i pochodne) które w przypadku napotkania wartości null zwróca bład processowania

Przykład analizy braku danych

```
import sklearn.datasets
import pandas as pd
import numpy as np

iris = sklearn.datasets.load_iris(as_frame=True)
data = pd.DataFrame(data= np.c_[iris['data'],
                                iris['target']],
      columns= iris['feature_names'] + ['target'])
data.isnull().sum()
```

Usuwanie zmiennych

Code example

Przykład usuniecia kolumny za pomoca python - pandas

```
import pandas as pd

data = pd.read_csv('path_to_your_data')

new_data = data.drop('column_name', axis=1)

# or

data.drop('column_name', axis=1, inplace=True)
```

Standaryzacja zmiennych

Zapamietaj

Każdy algorytm który działa na podstawie liczenia odległości miedzy różnymi zmiennymi wymaga od nas wykonanai standaryzacji

Typy standaryzacji

- MinMaxScaler
- StandardScaler

Przykład

Na wykładzie.

Pytanie z *

Czy w przypadku gdy wszystkie nasze cechy posiadaja rozkład binarny, standaryzacja jest wymagana?

PCA

Principal component analysis

Analiza głównych składowych (PCA) to popularna technika analizy dużych zbiorów danych zawierajacych duża liczbe wymiarów/cech na obserwacje, zwiekszajaca interpretowalność danych przy jednoczesnym zachowaniu maksymalnej ilości informacji i umożliwiajaca wizualizacje danych wielowymiarowych

Przykład redukcji wymiarow

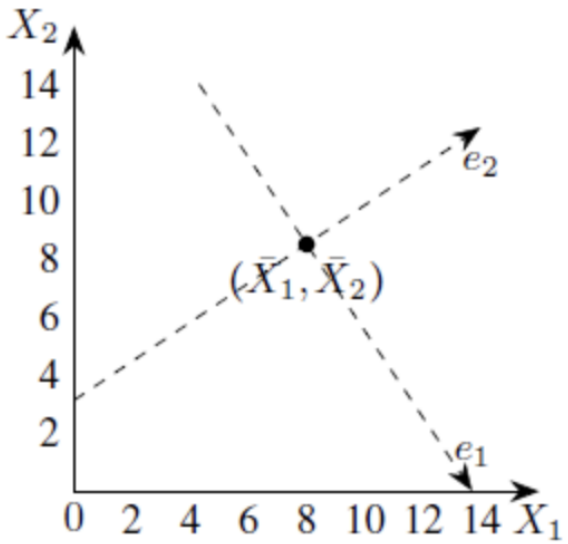

Figure: Rzut wektorów wlasnych

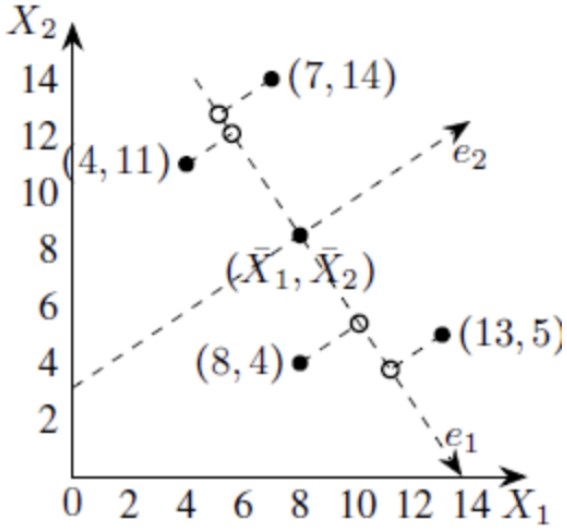

Figure: Mapowanie na podstawie wektorów wlasnych

RandomForest

Opis algorytmu

Losowe lasy lub losowe lasy decyzyjne to metoda uczenia sie zespołowego do klasyfikacji, regresji i innych zadań, która polega na konstruowaniu wielu drzew decyzyjnych w czasie szkolenia.

Zastosowanie

- Regresja
- Klasyfikacja

Dokładne dzialanie algorytmu

Na wykladzie.

RandomForest

Code example

```
import sklearn.datasets
import sklearn.ensemble as ens
import pandas as pd
iris = sklearn.datasets.load_iris(as_frame=True)
X_clf = pd.DataFrame(data=iris.data, columns=iris.feature_name
y_clf = iris.target
model_clf = ens.RandomForestClassifier()
model_clf.fit(X=X_clf, y=y_clf)
model_clf.predict(X_clf)
boston = sklearn.datasets.load_boston()
X_reg = pd.DataFrame(data=boston.data, columns=boston.feature_
y_reg = boston.target
model_reg = ens.RandomForestRegressor()
model_reg.fit(X=X_reg, y=y_reg)
model_reg.predict(X_reg)
```

Least Absolute Shrinkage and Selection Operator

Opis algorytmu

LASSO jest metoda analizy regresji, która przeprowadza zarówno selekcje zmiennych oraz ich regularyzacje w celu zwiekszenia dokładności przewidywania i interpretowalności modelu statystycznego, który generuje.

Historia

Została wprowadzona przez Roberta Tibshiraniego w 1996r. Warto w tym miejscu zaznaczyć, że w naszej pracy korzystamy ze zmodyfikowanej metody LASSO. Pierwotnie autor użył metody najmniejszych kwadratów do estymacji parametrów $\hat{\beta}$, natomiast my wykorzystujemy logarytm najwiekszej wiarygodności. Metoda ta została zmodyfikowana w 2006 r.

LASSO

Niech **X** oznacza macierz predyktorów, a **y** wektor odpowiedzi. Dla danego parametru t współczynniki $\hat{\beta}^{LASSO} = (\hat{\beta}_1, \ldots, \hat{\beta}_k)$ sa rozwiazaniem poniższej optymalizacji z ograniczeniem:

$$min\{-L(\mu; \mathbf{y})\} \quad dla \quad \sum_{j=1}^{k} |\beta_j| \leq t \tag{1}$$

gdzie:
– L jest funkcja najwiekszej wiarygodności.

Wzór możemy zapisać w równoważnej formie Lagrange'a:

$$min\Big\{ - L(\mu; \mathbf{y}) + \lambda \cdot \sum_{j=1}^{k} ||\beta_j|| \Big\} \tag{2}$$