

Uczenie maszynowe w pythonie

Wykład 1

Mateusz Serocki

Politechnika Gdańska

November 17, 2023

Agenda

- 1 Kilka słów o mnie
- 2 Redukcja wymiarowości
 - Wady
 - Zalety
- 3 Feature selection
 - Wariancja
 - Korelacja
 - Braki danych
 - Usuwanie zmiennych
- 4 Feature extraction
 - PCA
 - RandomForest
- 5 XGBoost
- 6 Uogólnione modele liniowe
 - LASSO

Kilka słów 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

Redukcja wymiarowości

Definicja

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

Interpretacja słowna

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

Przykład redukcji wymiarów

| | 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 usunięcia potencjalnie informatywnej zmiennej
- Całkowita lub częściowa utrata interpretowalności zmiennych
- Kolejny krok potrzebny do przygotowania danych

Zalety redukcji wymiarów

Pozytywne aspekty redukcji wymiarów to:

- Redukcja szumu pochodzącego ze zmiennych
- Brak zmiennych które nie mają uzasadnionego wpływu na wynik
- Mniejsza moc obliczeniowa potrzebna do utworzenia modelu
- Możemy zrezygnować z pobierania niektórych zmiennych jeżeli na tym etapie sadzimy że są nieistotne

Wariancja

Teoria

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

Zastosowanie

W praktyce wariancja odzwierciedla zmienność danej cechy, niska wariancja może nieść za sobą niską informatywność. Warto rozważyć usunięcie takiej cechy jeżeli nasz zbiór jest bardzo duży. Koniecznie musimy sprawdzić wpływ tego usunięcia na wynik modelu.

Korelacja

Typy korelacji

- Pearson - standardowa korelacja liniowa
- Spearman - korelacja rankingowa
- Kendal - korelacja rankingowa dla grup z powtarzającymi się wartościami
- 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 wewnątrz naszych danych, oraz może służyć do wstępnej 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

Przykład 1

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

Przykład 2

Przykładem algorytmu który nie radzi sobie z brakami danych jest regresja liniowa (i pochodne) które w przypadku napotkania wartości null zwróca błąd 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 usunięcia kolumny za pomocą 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

Zapamiętaj

Każdy algorytm który działa na podstawie liczenia odległości między różnymi zmiennymi wymaga od nas wykonania standaryzacji

Typy standaryzacji

- MinMaxScaler
- StandardScaler

Przykład

Na wykładzie.

Pytanie z *

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

Principal Component Analysis

PCA

Analiza głównych składowych (PCA) to popularna technika analizy dużych zbiorów danych zawierających dużą liczbę wymiarów/cech na obserwacje, zwiększająca interpretowalność danych przy jednoczesnym zachowaniu maksymalnej ilości informacji i umożliwiającą wizualizację danych wielowymiarowych

Przykład numeryczny

Na wykładzie przejdziemy przez przykład numeryczny, podsumowanie dostępne pod [Link](#)

Przykład redukcji wymiarów

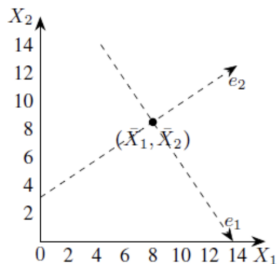


Figure: Rzut wektorów własnych

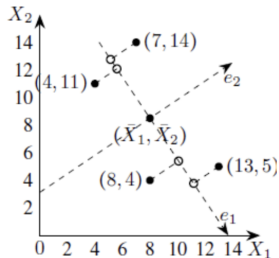


Figure: Mapowanie na podstawie wektorów własnych

Zadanie

Proszę uruchomic skrypty z poprzednich zajęć, tj:

- column drop
- missing data
- pandas profiling
- pca

Zadanie

Proszę pobrać dowolny zbiór danych ze strony [kaggle.com](https://www.kaggle.com), następnie napisać skrypt który, wczyta dane, usunie jedną z kolumn, sprawdzi brak danych, jeżeli takie się pojawia to zastąpi braki danych wartością 0 lub średnią danej cechy. Opcjonalnie uruchomi pandas profiling. Wykona standaryzację, a następnie PCA tylko dla podzbioru zmiennych.

Dla chetnych

Proszę uruchomic następujący skrypt [Link](#)

RandomForest/Drzewa decyzyjne

Opis algorytmu

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

Zastosowanie

- Regresja
- Klasyfikacja

Dokładne działanie algorytmu

Algorytm omówiony na wykładzie, podsumowanie dostępne pod [Link](#)

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_names)
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_names)
y_reg = boston.target

model_reg = ens.RandomForestRegressor()
model_reg.fit(X=X_reg, y=y_reg)
model_reg.predict(X_reg)
```

RandomForest

Feature Importance

- Gini Importance / Mean Decrease in Impurity (MDI)
- Permutation Importance or Mean Decrease in Accuracy (MDA)

Omówienie metod

Metody omówione na wykładzie, podsumowanie dostępne pod [Link](#)

Zadanie

Proszę uruchomic skrypt randomforest, nastepnie policzyc MDI oraz MDA importances porownac wyniki ze soba na podstawie bar plotu (moga byc dwa osobne wykresy).

Wartosci powinny byc posortowane malejaco.

Opis

XGBoost jest zbiorowym, bazującym na drzewach, algorytmem uczenia maszynowego, wykorzystującym strukturę wzmacniającą gradient

Przykład działania algorytmu

Omwówiony na wykładzie, [Link](#)

Zadanie

Proszę uruchomic skrypty xgboost clf i xgb reg (lecture two), policzyć FI, wyświetlić posortowane, skrypt uruchomic również w notebooku

Definition

Rodzina rozkładów prawdopodobieństwa nazywa się rodziną wykładniczą, jeżeli każdy należący do niej rozkład ma funkcję rozkładu prawdopodobieństwa postaci

$$f(y|\theta, \phi) = \exp\left\{\frac{y \cdot \theta - b(\theta)}{\phi} + c(y, \phi)\right\},$$

gdzie:

θ – parametr kanoniczny ($\theta \in \mathbf{R}$),

ϕ – parametr dyspersji ($\phi \in \mathbf{R}_+$),

$b(\theta)$ – jest funkcja dwukrotnie różniczkowalna z dodatnią drugą pochodną,

$c(y, \phi)$ – jest funkcja niezależna od parametru θ oraz $y \in \mathbf{R}$.

Definition

Funkcja prawdopodobieństwa rozkładu binarnego przedstawia się następująco:

$$f(y) = \begin{cases} \mu & \text{gdy } y = 1 \\ 1 - \mu & \text{gdy } y = 0. \end{cases}$$

Wartość oczekiwana oraz wariancja dla zmiennej losowej Y o rozkładzie binarnym wynosi:

$$E(Y) = \mu$$

$$\text{Var}(Y) = \mu(1 - \mu).$$

Rozkład binarny należy do rodziny wykładniczej.

Proof.

Funkcje prawdopodobieństwa rozkładu binarnego możemy zapisać za pomocą rodziny wykładniczej:

$$\begin{aligned} f(y) &= \mu^y \cdot (1 - \mu)^{1-y} = \\ &= \exp\left\{y \cdot \log \mu + (1 - y) \cdot \log(1 - \mu)\right\} = \\ &= \exp\left\{y \cdot \log \mu + \log(1 - \mu) - y \cdot \log(1 - \mu)\right\} = \\ &= \exp\left\{y \cdot \log\left(\frac{\mu}{1-\mu}\right) + \log(1 - \mu)\right\} = \\ &= \left| \theta = \log\left(\frac{\mu}{1-\mu}\right); \quad \mu = \frac{e^\theta}{1+e^\theta} \right| = \\ &= \exp\left\{y \cdot \theta + \log\left(\frac{1}{1+e^\theta}\right)\right\} = \end{aligned}$$

(1)

Proof.

$$= \exp \left\{ y \cdot \theta - \log(1 + e^{\theta}) \right\}$$

gdzie:

$$\begin{cases} \theta := \log\left(\frac{\mu}{1-\mu}\right) \\ \phi := 1 \\ b(\theta) := \log(1 + e^{\theta}) \\ c(y, \phi) := 0 \end{cases}$$

Zatem rozkład binarny należy do rodziny wykładniczej. □

Binarnej regresji logistycznej używamy do budowania modelu w przypadku gdy rozkład zmiennej zależnej Y jest określony funkcja rozkładu prawdopodobieństwa w następujący sposób:

$$P(Y = 1) = \mu \quad \text{oraz} \quad P(Y = 0) = 1 - \mu$$

Wartość oczekiwana zmiennej Y to $E(Y) = \mu$. Model ten służy do przewidywania prawdopodobieństwa a posteriori¹ μ wystąpienia sukcesu na podstawie danych (zmiennych niezależnych), korzystamy z niego w *Przykładzie 1*. Model regresji logistycznej, dla zmiennej objaśnianej o rozkładzie binarnym, dzięki użyciu funkcji łączącej, zwraca wynik interpretowalny na całej przestrzeni liczb rzeczywistych. Wartość μ może zmieniać się wraz ze zmianą wartości x , zatem zastępujemy μ przez $\mu(x)$, gdy chcemy opisać zależność od tej wartości. Dla ustalenia uwagi przyjmijmy logit za funkcję łączącą:

$$g(\mu(x)) = \text{logit}(\mu(x)) = \beta_0 + \beta x \quad (2)$$

gdzie: β_0, β – współczynniki modelu.

¹a posteriori – w filozofii termin oznaczający: po fakcie” lub w następstwie faktu”

Wyznaczając funkcję odwrotną do g obliczamy, dla znanego x , prawdopodobieństwo a posteriori sukcesu:

$$g^{-1}(x) = \mu(x) = \frac{\exp(\beta_0 + \beta x)}{1 + \exp(\beta_0 + \beta x)} \quad (3)$$

Warto zauważyć, że:

$$\text{szansa}(x) = \frac{\mu(x)}{1 - \mu(x)} = e^{\beta_0 + \beta x}$$

Zatem:

$$\frac{\text{szansa}(x+1)}{\text{szansa}(x)} = \frac{e^{\beta_0 + \beta(x+1)}}{e^{\beta_0 + \beta x}} = e^{\beta}$$

Wiec szansa wzrasta e^{β} razy przy wzroście wartości x o 1. Wówczas $\mu(x)$ zazwyczaj rośnie bądź maleje w sposób ciągły wraz ze wzrostem x . Jej monotoniczność zależy od znaku współczynnika β . Relacja ta została przedstawiona na rysunku przedstawionym na wykładzie.

Jeśli istnieje wiele zmiennych objaśniających, równość (2) rozszerzamy do postaci:

$$\text{logit}(\mu(x_1, x_2, \dots, x_k)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k = \beta_0 + \sum_{i=1}^k \beta_i x_i$$

Analiza różnic w modelach

Rozważmy różnice między analizowaniem wyniku regresji liniowej, a regresji logistycznej. Przykład omówimy na tablicy.

Przykład

Rozpoczniemy do aplikacyjnego charakteru uogólnionych modeli liniowych (GLM). W tabeli 1 przedstawiamy dane z 23 lotów kosmicznych przed katastrofą misji Challenger w 1986r. Tabela przedstawia temperaturę (F°) w czasie startu i informacje czy przynajmniej jedna uszczelka O-Rings rozszczelniła się. Rozważmy model pierwszy z funkcją wiążącą $g(\mu) = \mu$ oraz model drugi z funkcją wiążącą $g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

| Lp. | Temperatura | Usterka | Lp. | Temperatura | Usterka |
|-----|-------------|---------|-----|-------------|---------|
| 1 | 53 | 1 | 13 | 70 | 0 |
| 2 | 57 | 1 | 14 | 70 | 1 |
| 3 | 58 | 1 | 15 | 72 | 0 |
| 4 | 63 | 1 | 16 | 73 | 0 |
| 5 | 66 | 0 | 17 | 75 | 0 |
| 6 | 67 | 0 | 18 | 75 | 1 |
| 7 | 67 | 0 | 19 | 76 | 0 |
| 8 | 67 | 0 | 20 | 76 | 0 |
| 9 | 68 | 0 | 21 | 78 | 0 |
| 10 | 69 | 0 | 22 | 79 | 0 |
| 11 | 70 | 0 | 23 | 81 | 0 |
| 12 | 70 | 0 | | | |

Table: Dane z książki: Alan Agresti *An Introduction to Categorical Data Analysis*, Second Edition (tabela 4.10)

Uwaga: Usterka (1=wystąpiła, 0=nie wystąpiła)

Model pierwszy:

$$E(Usterka) = 2,888889 + temp \cdot (-0,037778)$$

Model drugi:

$$E(Usterka) = \frac{e^{16,798079 + temp \cdot (-0,263060)}}{1 + e^{16,798079 + temp \cdot (-0,263060)}}$$

| Lp. | Model 1 | Model 2 | Lp. | Model 1 | Model 2 |
|-----|--------------|--------------|-----|---------------|--------------|
| 1 | 0,8866666667 | 0,9456234303 | 13 | 0,2444444444 | 0,1657427394 |
| 2 | 0,7355555556 | 0,8585953369 | 14 | 0,2444444444 | 0,1657427394 |
| 3 | 0,6977777778 | 0,8235537071 | 15 | 0,1688888889 | 0,1050600495 |
| 4 | 0,5088888889 | 0,5560912516 | 16 | 0,1311111111 | 0,0827706285 |
| 5 | 0,3955555556 | 0,3626534324 | 17 | 0,0555555556 | 0,0506228136 |
| 6 | 0,3577777778 | 0,3042955086 | 18 | 0,0555555556 | 0,0506228136 |
| 7 | 0,3577777778 | 0,3042955086 | 19 | 0,0177777778 | 0,0393745994 |
| 8 | 0,3577777778 | 0,3042955086 | 20 | 0,0177777778 | 0,0393745994 |
| 9 | 0,32 | 0,2516210163 | 21 | -0,0577777778 | 0,0236471108 |
| 10 | 0,2822222222 | 0,2053729601 | 22 | -0,0955555556 | 0,0182774098 |
| 11 | 0,2444444444 | 0,1657427394 | 23 | -0,1711111111 | 0,0108813664 |
| 12 | 0,2444444444 | 0,1657427394 | | | |

Table: Wyestymowane prawdopodobieństwo dla dwóch modeli

Jak widać w tabeli 2, niektóre z naszych wyników dla modelu pierwszego nie należą do oczywistego przedziału prawdopodobieństwa $[0;1]$, co niestety stawia pod znakiem zapytania ich interpretowalność. Taki problem nie występuje przy zastosowaniu logitowej funkcji wiążacej. Na pytanie dlaczego, odpowiemy przy okazji omawiania tej funkcji.

Zadanie

Proszę znaleźć zbiór danych binarnych na stronie [kaggle.com](https://www.kaggle.com), następnie uruchomić model logitowy do przewidywania prawdopodobieństwa, na podstawie prawdopodobieństwa proszę ustalić punkt odciecia, który zmaksymalizuje nam accuracy modelu. Gdzie accuracy modelu rozumiane jest poprzez $\frac{\text{sume}(\text{poprawnie sklasyfikowanych jedynek i 0})}{\text{sume wszystkich przypadkow}}$

Least Absolute Shrinkage and Selection Operator

Opis algorytmu

LASSO jest metoda analizy regresji, która przeprowadza zarówno selekcję zmiennych oraz ich regularyzację w celu zwiększenia 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 największej wiarygodności. Metoda ta została zmodyfikowana w 2006 r.

LASSO

Niech \mathbf{X} oznacza macierz predyktorów, a \mathbf{y} wektor odpowiedzi. Dla danego parametru t współczynniki $\hat{\beta}^{\text{LASSO}} = (\hat{\beta}_1, \dots, \hat{\beta}_k)$ są rozwiązaniem poniższej optymalizacji z ograniczeniem:

$$\min\{-L(\mu; \mathbf{y})\} \quad \text{dla} \quad \sum_{j=1}^k |\beta_j| \leq t \quad (4)$$

gdzie:

– L jest funkcja największej wiarygodności.

Wzór możemy zapisać w równoważnej formie Lagrange'a:

$$\min\left\{-L(\mu; \mathbf{y}) + \lambda \cdot \sum_{j=1}^k \|\beta_j\|\right\} \quad (5)$$