Uczenie maszynowe w pythonie

Wyklad 1

Mateusz Serocki

Politechnika Gdańska

November 3, 2023

Agenda

1. Kilka słow o mnie
2. Reduckja wymiarowości
   - Wady
   - Zalety
3. Feature selection
   - Wariancja
   - Korelacja
   - Braki danych
   - Usuwanie zmiennych
4. Feature extraction
   - PCA
   - RandomForest
5. Uogólnione modele liniowe
   - Rodzina wykładnicza i funkcje wiażace
     - LASSO

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?

Principal Component Analysis

PCA

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

Przyklad numeryczny

Na wykładzie przejdziemy przez przykład numeryczny, podsumowanie dostepne pod Link

Przykład redukcji wymiarow

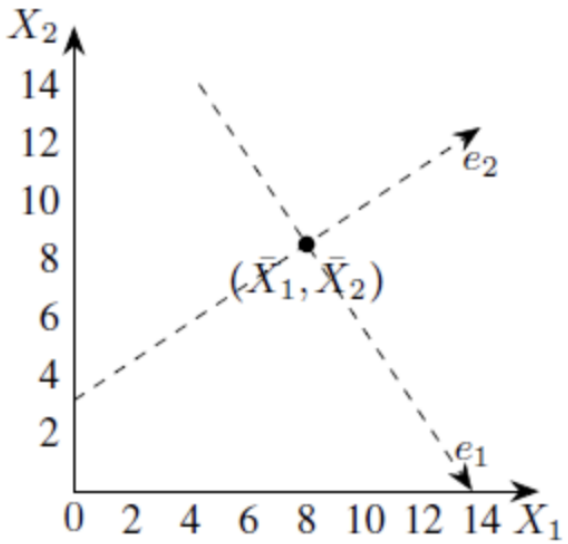

Figure: Rzut wektorów wlasnych

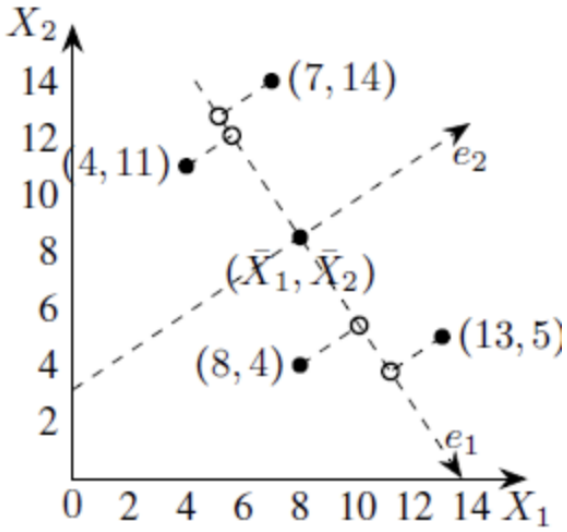

Figure: Mapowanie na podstawie wektorów wlasnych

Zadanie

Prosze uruchomic skrypty z poprzednich zajec, tj:

- column drop
- missing data
- pandas profiling
- pca

Zadanie

Prosze pobrac dowolny zbior danych ze strony kaggle.com, nastepnie napisac skrypt ktory, wczyta dane, usunie jedna z kolumn, sprawdzi brak danych, jezeli takie sie pojawia to zastapi braki danych wartoscia 0 lub srednia danej cechy. Opcjonalnie uruchomi pandas profiling. Wykona standaryzacje, a nastepnie PCA tylko dla podzbioru zmiennych.

Dla chetnych

Prosze uruchomic nastepnujacy skrypt Link

RandomForest/Drzewa decyzyjne

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

Algorytm omówiony na wykładzie, podsumowanie dostepne 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_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)
```

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 dostepne pod Link

Zadanie

Prosze 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.

W uogólnionym modelu liniowym zakłada sie, że każda rozważana zmienna losowa Y jest z rodziny wykładniczej (zob. *Definicja 1*). Należa do niej miedzy innymi rozkłady: normalny, Poissona, gamma, binarny, binominalny. Oznaczmy wartość oczekiwana zmiennej Y przez μ. Zakładamy, że μ zależy od nielosowych zmiennych niezależnych (objaśniajacych) $x = (x_1, x_2, ..., x_k)$, w nastepujacy sposób:

$$E(Y) = \mu = g^{-1}(x \circ \beta)$$

gdzie:

- $E(Y)$ – wartość oczekiwana Y;
- $x \circ \beta = \sum_{j=1}^{k} x_j \beta_j$ – model liniowy; liniowa kombinacja nieznanych parametrów β;
- g – funkcja wiażaca, dla której istnieje funkcja odwrotna g^{-1}.

Definicja wariancji dla zmiennej zależnej Y

Zakładamy, że wariancje zmiennej zależnej Y można wyrazić poprzez funkcje V:

$$Var(Y) = V(\mu) = V(g^{-1}(x \circ \beta)).$$

Nieznane parametry β sa szacowane metoda najwiekszej wiarygodności.

Funkcje rozkładu prawdopodobieństa dla zmiennej dyskretnej przedstawia *Definicja 1*, natomiast dla zmiennej ciagłej *Definicja 2*.

Definicja 1

Załóżmy, że $Y : S \to A$ $(A \subseteq \mathbb{R})$ jest dyskretna zmienna losowa zdefiniowana na przestrzeni próbki S. Funkcje rozkładu prawdopodobieństwa $f_Y : A \to [0; 1]$ dla zmiennej Y definiujemy jako:

$$f_Y(y) = P(Y = y) = P(\{s \in S : Y(s) = y),$$

która spełnia warunek:

$$\sum_{x \in A} f_Y(y) = 1.$$

Funkcja rozkładu prawdopodobieństwa P (gestościa) ciagłej zmiennej Y nazywamy nieujemna funkcje borelowska $f_Y : \mathbb{R}^N \to \mathbb{R}_+ \cup \{0\}$, taka że dla każdego zbioru borelowskiego $\mathcal{B} \subseteq \mathbb{R}^N$ zachodzi równość:

$$P(\mathcal{B}) = \int_{\mathcal{B}} f_Y(y)dy.$$

Jeżeli F_Y jest dystrybuanta zmiennej losowej Y, to:

$$F_Y(y) = \int_{-\infty}^{y} f_Y(u)du,$$

i jeżeli f_Y jest ciagła w y to:

$$f_Y(y) = \frac{\partial}{\partial y} F_Y(y),$$

która spełnia warunek:

$$\int_{\mathbb{R}^N} f_Y(y)dy = 1.$$

Rozkład należy do danej rodziny, jeżeli jesteśmy w stanie przedstawić jego

funkcje rozkładu prawdopodobieństwa za pomoca ogólnego wzoru tej rodziny.

Definition

Rodzina rozkładów prawdopodobieństwa nazywa sie rodzina wykładnicza, jeżeli każdy należacy do niej rozkład ma funkcje 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 dodatnia druga pochodna,
$c(y,\phi)$ – jest funkcja niezależna od parametru θ oraz $y \in \mathbf{R}$.

Definition

Funkcja prawdopodobieństwa rozkładu binarnego przedstawia sie nastepujaco:

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

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

$$E(Y) = \mu$$

$$Var(Y) = \mu(1 - \mu).$$

Rozkład binarny należy do rodziny wykładniczej.

Proof.

Funkcje prawdopodobieństwa rozkładu binarnego możemy zapisać za pomoca rodziny wykładniczej:

$$f(y) = \mu^y \cdot (1-\mu)^{1-y} =$$

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

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

$$= exp\left\{ y \cdot log\left(\frac{\mu}{1-\mu}\right) + log(1-\mu) \right\} =$$

$$= \left| \theta = log(\frac{\mu}{1-\mu}); \quad \mu = \frac{e^\theta}{1+e^\theta} \right| =$$

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

(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 nastepujacy sposób:

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

Wartość oczekiwana zmiennej Y to $E(Y) = \mu$. Model ten służy do przewidywania prawdopodobieństwa a posteriori[1] μ wystapienia sukcesu na podstawie danych (zmiennych niezależnych), korzystamy z niego w *Przykładzie 1*. Model regresji logistycznej, dla zmiennej objaśnianej o rozkładzie binarnym, dzieki użyciu funkcji łaczacej, zwraca wynik interpretowalny na całej przestrzeni liczb rzeczywistych. Wartość μ może zmieniać sie wraz ze zmiana wartości x, zatem zastepujemy μ przez $\mu(x)$, gdy chcemy opisać zależność od tej wartości. Dla ustalenia uwagi przyjmijmy logit za funkcje łaczaca:

$$g(\mu(x)) = \mathit{logit}(\mu(x)) = \beta_0 + \beta x \tag{2}$$

gdzie: β_0, β – współczynniki modelu.

[1] a posteriori – w filozofii termin oznaczajacy: po fakcie" lub w nastepstwie faktu"

Wyznaczajac funkcje odwrotna do g obliczamy, dla znanego x, prawdopodobieństwo a posteriori sukcesu:

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

Warto zauważyć, że:

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

Zatem:

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

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

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

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

Analiza różnic w modelach

Rozważmy różnice miedzy 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 katastrofa misji Challenger w 1986r. Tabela przedstawia temperature (F°) w czasie startu i informacje czy przynajmniej jedna uszczelka O-Rings rozszczelniła sie. Rozważmy model pierwszy z funkcja wiażaca $g(\mu) = \mu$ oraz model drugi z funkcja wiażaca $g(\mu) = logit(\mu) = log(\frac{\mu}{1-\mu})$.

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 ksiażki: Alan Agresti *An Introduction to Categorical Data Analysis*, Second Edition (tabela 4.10)

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

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	-0,057777778	0,0236471108
10	0,2822222222	0,2053729601	22	-0,095555556	0,0182774098
11	0,2444444444	0,1657427394	23	-0,17111111	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 należa do oczywistego przedziału prawdopodobieństwa [0;1], co niestety stawia pod znakiem zapytania ich interpretowalność. Taki problem nie wystepuje przy zastosowaniu logitowej funkcji wiażacej. Na pytanie dlaczego, odpowiemy przy okazji omawiania tej funkcji.

Zadanie

Prosze znalezc zbior danych binarnych na stronie kaggle.com, nastepnie uruchomic model logitowy do przewidywania prawdopodobienstwa, na podstawie prawdopodobienstwa prosze ustalic punkt odciecia, ktory zmaksymalizuje nam accuracy modelu. Gdzie accuracy modelu rozumiane jest poprzez sume(poprawnie sklasyifikowanych jedynek i 0) podzielone przez sume wszystkich przypadkow

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:

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

gdzie:
– L jest funkcja najwiekszej wiarygodności.

Wzór możemy zapisać w równoważnej formie Lagrange'a:

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