Zintegrowany Program Rozwoju
Akademii Górniczo-Hutniczej w Krakowie
Nr umowy: **POWR.03.05.00-00-Z307/17**

Instrukcja do ćwiczeń laboratoryjnych

Nazwa przedmiotu	Eksploracja danych
Numer ćwiczenia	1
Temat ćwiczenia	Metody redukcji wymiaru: Principal Component Analysis (PCA)

Poziom studiów	II stopień
Kierunek	Informatyka
Forma studiów	Stacjonarne
Semestr	1

Wojciech Czech

Wydział Informatyki, Elektroniki i Telekomunikacji
Kraków, 2019

1. Cel ćwiczenia

— Praktyczne zapoznanie się z liniową metodą redukcji wymiaru PCA
— Przyswojenie pojęć: komponent wiodący, macierz kowariancji, wewnętrzna wymiarowość przestrzeni cech, *biplot*

2. Wprowadzenie do ćwiczenia

Redukcja wymiarowości przestrzeni cech jest są kluczową metodą eksploracji danych. Ma ona na celu wykrycie wewnętrznej wymiarowości danych, usunięcie redundancji i szumów oraz wyostrzenie cech najbardziej wartościowych z punktu widzenia dyskryminacji danych. Principal Component Analysis (PCA) jest bezparametryczną, liniową metodą redukcji wymiaru szeroko wykorzystywaną w analizie i przetwarzaniu danych w postaci wektorów cech.

3. Przykładowe dane

— `http://home.agh.edu.pl/~czech/vis-datasets/misc/nyt-frame.csv`
Zbiór danych o rozmiarze 101×4433 zawierający 101 wektorów cech reprezentujących artykuły New York Times w dwóch kategoriach: muzyka i sztuka. Wektory cech są znormalizowanymi i przeskalowanymi zgodnie z rankingiem IDF wektorami BoW. Rozmiar słownika wynosi 4433.
— `http://home.agh.edu.pl/~czech/vis-datasets/misc/04cars-data.csv`
Zbiór danych o rozmiarze 387×20 zawierający 386 wektorów cech reprezentujących własności samochodów produkowanych w roku 2004. Pierwsza kolumna zawiera model samochodu, 8 kolejnych kolumn - dane binarne opisujące typ samochodu, następne 11 kolumn dane numeryczne np. moc lub liczba cylindrów.
— `http://vis-www.cs.umass.edu/lfw/`
Zbiór danych Labeled Faces in the Wild (LFW) - benchmarkowy zbiór danych do rozpoznawania twarzy składający się z ponad 13000 zdjęć różnych osób.

4. Przydatne biblioteki i funkcje

1. Pandas: `https://pandas.pydata.org`
   — `csv_read()`
   — `DataFrame`
2. NumPy:
   — `random.choice`
   — `array`
   — `argsort`
3. SciKit Learn:
   — `PCA`
   — `preprocessing.scale`

5. Plan ćwiczenia: analiza zbioru danych New York Times

1. Załaduj zbiór danych NYT jako `DataFrame` (biblioteka Pandas)

```
import pandas as pd
df = pd.read_csv('./nyt-frame.csv', header = 0)
data = df.iloc[:,9:]
array = data.values
```

2. Wyświetl 20 losowych wybranych elementów słownika (nagłówek)

```
header = list(df.columns.values[9:])
sample_word = np.random.choice(header, 20, replace=False)
print(sample_word)
```

3. Dokonaj transformacji PCA pomijając kolumny, które nie reprezentują słów a jedynie metadane

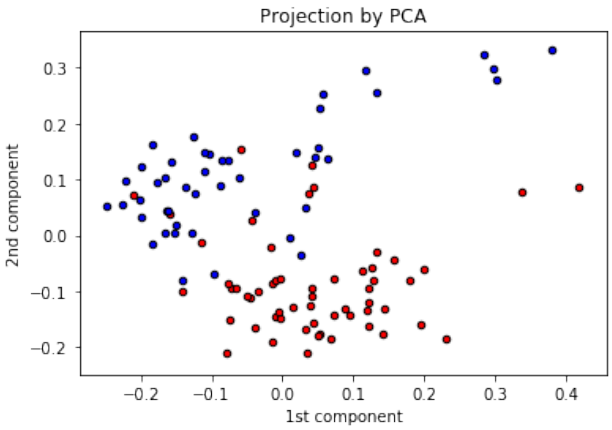
```
from sklearn.decomposition import PCA
pca = PCA()
X_pca = pca.fit_transform(array)
```

4. Dla pierwszego komponentu wiodącego wyświetl 15 elementów o największej wartości (wraz z nazwami kolumn - słowami). Jakim słowom odpowiadają największe wartości pierwszego komponentu wiodącego?
5. Dla pierwszego komponentu wiodącego wyświetl 15 elementów o najmniejszej wartości (wraz z nazwami kolumn - słowami). Jakim słowom odpowiadają najmniejsze wartości pierwszego komponentu wiodącego?
6. Powtórz eksperyment dla drugiego komponentu wiodącego. Skomentuj uzyskane wyniki.
7. Dokonaj wizualizacji wektorów cech zrzutowanych na 2(3) pierwsze komponenty wiodące. Zaznacz dwie klasy (art, music) oddzielnymi kolorami (patrz Rysunek 1).

```
import matplotlib.pyplot as plt
plt.figure()
plt.scatter(X_pca[np.array(reds), 0], X_pca[np.array(reds), 1], c="red")
plt.scatter(X_pca[np.array(blues), 0], X_pca[np.array(blues), 1], c="blue")
plt.title("Projection by PCA")
plt.xlabel("1st component")
plt.ylabel("2nd component")
plt.show()
```

8. Narysuj wykres zależności wartości wariancji od numeru kierunku wiodącego k. Jaka część wariancji zostaje zachowana po wykonaniu projekcji na pierwsze 10 komponentów wiodących?

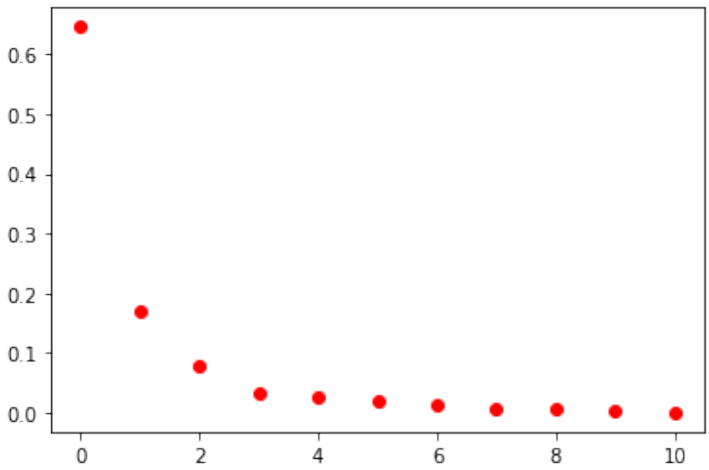
```
variance_ratio = pca.explained_variance_ratio_
plt.plot(variance_ratio, 'ro')
plt.show()
print(sum(variance_ratio[0:10]))
```



Rysunek 1: Rzutowanie na dwa pierwsze komponenty główne dla zbioru danych NYT.

6. Plan ćwiczenia: analiza zbioru danych 04Cars

1. Załaduj zbiór danych 04Cars jako `DataFrame` (biblioteka Pandas) i odfiltruj 11 ostatnich kolumn - zostaną one użyte jako wejście dla PCA.
2. Znormalizuj dane, zapewniając, że dla każdej cechy średnia arytmetyczne wynosi 0, a wariancja 1.
3. Dokonaj transformacji PCA dla wejściowego zbioru danych, a następnie przedstaw zależność bezwzględnej i względnej wartości wariancji od numeru kierunku wiodącego k. Jaka część wariancji zostaje zachowana przy redukcji wymiarowości do 2 i 3 (patrz Rysunek 2)?



Rysunek 2: Procent zachowanej wariancji na komponent główny dla zbioru danych 04Cars.

4. Wyświetl wartości elementów pierwszego i drugiego komponentu wiodącego wraz z odpowiadającą im nazwą cechy. Co oznaczają elementy o wartości bliskiej zeru? Z jakimi własnościami samochodu wiążą się:
   — Największe elementy pierwszego komponentu wiodącego
   — Najmniejsze elementy pierwszego komponentu wiodącego

— Największe elementy drugiego komponentu wiodącego
— Najmniejsze elementy drugiego komponentu wiodącego

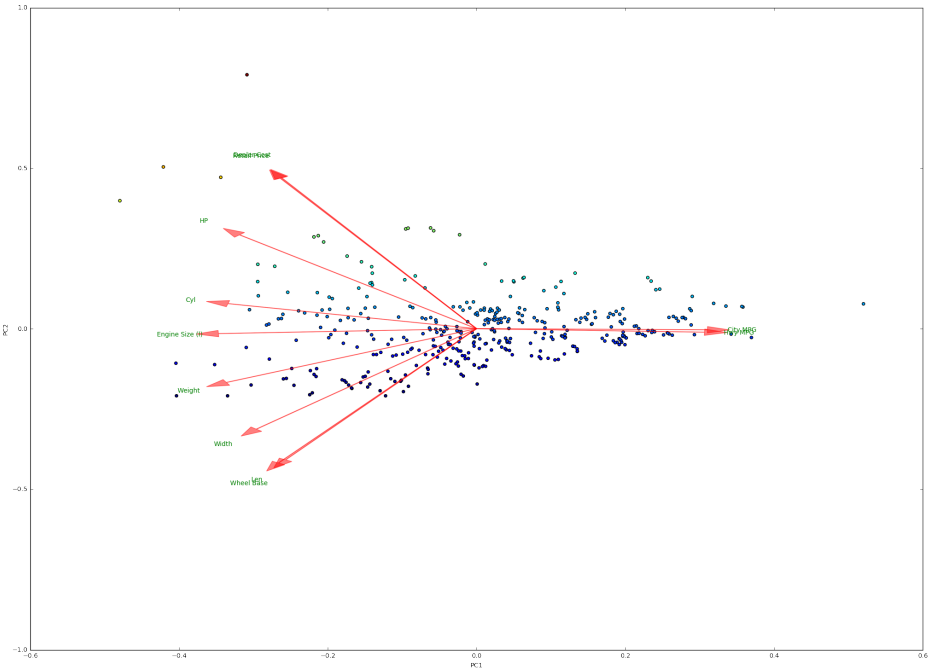
```
pc1 = pca.components_[0]
pc2 = pca.components_[1]
print('Attribute, PC1, PC2')
for i in range(0,pc1.shape[0]):
    print(attributes[i] + ':' + repr(pc1[i]) + ':' + repr(pc2[i]))
```

```
Attribute, PC1, PC2
Retail Price:-0.2637504434440343:0.46850869750253876
Dealer Cost:-0.2623186387530949:0.47014658513822577
Engine Size (l):-0.3470804920252009:-0.015347186463713367
Cyl:-0.33418875762863715:0.07803201087501883
HP:-0.3186022584840293:0.29221347613918247
City MPG:0.3104817267323128:-0.003365935761659622
Hwy MPG:0.30658863858044433:-0.010964460145349025
Weight:-0.336329366940488:-0.16746357154787023
Wheel Base:-0.2662100335710544:-0.4181771069592044
Len:-0.2567901876706823:-0.4084113806687549
Width:-0.29605459141706114:-0.31289135016250724
```

5. Dokonaj wizualizacji wektorów cech zrzutowanych na 2(3) pierwsze komponenty wiodące. Nanieś tekst z modelem samochodu na wykres. Jakie typy samochodów występują w poszczególnych częściach wykresu?
6. Na tym samym wykresie w 2D przedstaw zbiór danych o zredukowanym rozmiarze wraz z wizualicacją cech zrzutowanych na 2 pierwsze komponenty wiodące (*biplot*, patrz Rysunek 3). Co pokazuje tego typu wizualizacja?

7. Zadanie dodatkowe: budowa eigenfaces w oparciu o zbiór danych LFW

1. Korzystając z biblioteki SciKit Learn wczytaj podzbiór zbioru danych LFW składający się z osób reprezentowanych przez co najmniej 50 różnych zdjęć (`fetch_lfw_people()`). W ten sposób otrzymasz zbiór zdjęć 12 różnych osób.
2. Podziel otrzymany zbiór danych na część treningową i testową (0.7, 0.3)
3. Dla **zbioru treningowego,** oblicz PCA i opierając się na otrzymanych w ten sposób 100 pierwszych komponentach głównych dokonaj redukcji wymiarowości do 100 dla zbioru treningowego i testowego.
4. Korzystając ze zbioru treningowego o zredukowanej wymiarowości, wytrenuj dwa wybrane modele klasyfikacji (np. MLP, SVM) oraz oblicz dokładność klasyfikacji uzyskaną dla każdej z 12 klas na zbiorze testowym. Zamieść w raporcie uzyskane wartości: *precision*, *recall*, *f1-score*. Jak zmieniają się uzyskane wyniki w przypadku redukcji do 50 wymiarów? Wybierz optymalną wymiarowość w oparciu o *scree plot*.



Rysunek 3: *Biplot* dla zbioru danych 04Cars.

5. Wyświetl 20 pierwszych komponentów głównych (po uprzednim przeskalowaniu) w formie obrazu w skali szarości (*eigenfaces*). W jaki sposób można zinterpretować działanie *eigenfaces* w kontekście redukcji wymiarowości wcześniej nie widzianego zdjęcia (przykładu testowego)? Jakie cechy ekstrahują poszczególe *eigenfaces*?

8. Sposób oceny / uzyskania zaliczenia

Na uzyskanie zaliczenia z zajęć laboratoryjnych składa się:

— Zaliczenie wstępnej kartkówki z tematyki zajęć
— Wykonanie wszystkich zadań na laboratorium oraz przesłanie kodu za pomocą systemu UPeL

Ocena z zajęć laboratoryjnych (OL, w skali $2-5$) obliczana jest zgodnie ze wzorem:

$$OL = 0.5 * LA + 0.5 * LW,$$

gdzie:

— LA – ocena aktywności studenta podczas zajęć, wystawiana przez prowadzącego na podstawie zaangażowania studenta w realizację zadań oraz odpowiedzi ustnej na zadane pytania dotyczące realizowanego zadania;
— LW – ocena uzyskana za zadania wykonane na zajęciach (kod źródłowy) wystawiona przez prowadzącego na podstawie poprawności i kompletności zadania przesłanego na platformę UPeL.

9. Literatura

— *Data Mining: The Textbook*, Charu C. Aggarwal, Springer 2015.
— *Data Mining: Concepts and Techniques*, Jiawei Han, Micheline Kamber, Jian Pei, Elsevier 2012, Third Edition.
— *A tutorial on Principal Component Analysis: Derivation, Discussion and Singular Value Decomposition*, Jon Shlens, 2013.