Eksploracja danych

Laboratorium 1 – Dane

Marcin Pietrzykowski

ver. 1.00

Cel

Celem laboratorium jest przebadania podstawowych metod selekcji danych i ich wizualizacji.

Zadania

1. Wczytaj następujące zbiory danych:
   (a) `iris` oraz rozdziel je na cześć wejściową (X) i decyzje (Y).

   ```
   from sklearn import datasets
   iris = datasets.load_iris()
   X = iris.data
   Y = iris.target
   ```

   (b) `zoo`

   ```
   import pandas
   data = pd.read_csv('zoo.csv')
   ```

   (c) `autos` - dla tego zbioru danych należy wykonać jedynie kroki od 2 - 4. *Dalsze kroki są dla osób chętnych i wymagają uprzedniej binaryzacji danych jakościowych.*
2. Określić typy poszczególnych atrybutów
3. Podać częstości atrybutów dyskretnych oraz średnie (`statistics.mean`) i odchylenia standardowe (`statistics.stdev`) atrybutów ciągłych
4. Wykreślić histogramy (`matplotlib.pyplot.hist`) i wykresy pudełkowe (`matplotlib.pyplot.boxplot` albo `pandas.DataFrame.boxplot`) dla wybranych atrybutów ciągłych

5. Dokonać transformacji wykorzystując metodą PCA `sklearn.decomposition.pca.PCA`.

6. Wybrać liczbę składowych głównych, która wyjaśnia 90% zmienność zmiennych oryginalnych (`pca.explained_variance_ratio_`).

7. Zwizualizować dane w rzucie na dwie i trzy pierwsze składowe główne (na osobnych rysunkach), zaznaczając każdą z klas decyzyjnych innym kolorem.

8. Porównać jaka jest jakość wizualizacji w porównaniu z losową dwu lub trzy wymiarową projekcją danych oryginalnych.