Laboratorium: Redukcja wymiarów

1 Cel/Zakres

- Redukcja liczby wymiarów.
- Ocena efektów redukcji wymiarów.

2 Przygotowanie danych

Dane są dwa poniższe wielowymiarowe zbiory danych.

```
from sklearn import datasets
data_breast_cancer = datasets.load_breast_cancer()
```

```
from sklearn.datasets import load_iris
data_iris = load_iris()
```

3 Ćwiczenie

1. Przeprowadź analizę PCA, tak aby tak zredukować liczbę wymiarów dla każdego z w/w zbiorów. Nowa przestrzeń ma pokrywać przynajmniej 90% różnorodności (zmienności) danych i ma mieć jak najmniej wymiarów.
2. Ćwiczenia przeprowadź najpierw na oryginalnych danych, a później na danych przeskalowanych. Porównaj wyniki.

   W podanych zbiorach są istotnie różne zakresy dla poszczególnych cech. Aby je przeskalować, by były porównywalne, użyj `StandardScaler()`. Klasa `PCA()` centruje dane automatycznie, ale ich nie skaluje!
3. Utwórz listę z współczynnikami zmienności nowych wymiarów (dla danych przeskalowanych). W przypadku `data_breast_cancer` listę zapisz w pliku Pickle o nazwie `pca_bc.pkl`

   3 pkt.

   W przypadku `data_iris` listę zapisz w pliku Pickle o nazwie `pca_ir.pkl`

   3 pkt.
4. Dla danych przeskalowanych utwórz listę indeksów cech oryginalnych wymiarów, w kolejności od cechy, która ma największy udział w nowych cechach, do tej, która ma najmniejszy.

Podpowiedź: zob. atrybut `components_` klasy `PCA`, użyj wartości w `explained_variance_ratio_` jako wagę istotności udziałów starych cech w nowych cechach, czyli pomnóż `components_` przez `explained_variance_ratio`. W otrzymanej macierzy oblicz wartość bezwzględne dla każdej wartości (bo mogą być zarówno dodatnie jak i ujemne). Im większa wartość tym większy udział starego wymiaru w nowym. Posortuj wartości rosnąco, znajdź odpowiadające im indeksy starych cech i zapisz jako listę bez powtórzeń. Przydatne funkcje: `numpy.argsort()`, `ndarray.flatten()`. Usunięcie powtórzeń z listy możesz zrealizować np. tak: `list(dict.fromkeys([20,19,5,20,6]))`.

W przypadku `data_breast_cancer` listę zapisz w pliku Pickle o nazwie `idx_bc.pkl`

3 pkt.

W przypadku `data_iris` listę zapisz w pliku Pickle o nazwie `idx_ir.pkl`

3 pkt.

4 Prześlij raport

Prześlij plik o nazwie `lab08/lab08.py` realizujący ww. ćwiczenia.

Sprawdzane będzie, czy skrypt Pythona tworzy wszystkie wymagane pliki oraz czy ich zawartość jest poprawna.