

# PCA

Celem zadania jest zapoznanie się z metodą analizy głównych składowych (ang. Principal Component Analysis, PCA). Pracować będziemy na zbiorze Plantdoc dataset:

<https://github.com/pratikkayal/PlantDoc-Dataset>

Jest to zbiór zdjęć przedstawiający choroby popularnych roślin uprawnych.

Preprocessing danych:

1. Wybierz ze zbioru Plantdoc podzbiór kilkudziesięciu zdjęć przedstawiających trzy choroby roślin jednego gatunku (np. 3 choroby ziemniaków lub 3 choroby pomidorów).
2. Wczytaj zdjęcia do pamięci. Po wczytaniu każde zdjęcie będzie trójwymiarowym tensorem. Pomocna będzie biblioteka imageio lub Pillow.
3. Przeskaluj wszystkie zdjęcia do rozdzielczości  $224 \times 224$ , tak aby wszystkie obrazy miały ten sam rozmiar, równy  $224 \times 224 \times 3$ .
4. Skonwertuj obrazy do skali szarości, tak aby z trójwymiarowego tensora reprezentującego dane zdjęcie otrzymać tablicę dwuwymiarową.
5. Skonwertuj obrazy, będące teraz tablicami dwuwymiarowymi (macierzami) na wektory.

Pomocna będzie jedna z funkcji: `np.reshape`, `np.ravel`, `np.flatten`. Funkcje `np.reshape` i `np.ravel` zwracają, gdy tylko jest to możliwe, widok oryginalnej tablicy. Funkcja `np.flatten` zwraca kopię tablicy, co zwykle jest niepożądane.

Każdy obraz powinien być teraz reprezentowany przez wektor o rozmiarze 50176.

6. Przeprowadź standaryzację obrazów, czyli od każdego obrazu odemij średni obraz, a następnie podziel przez odchylenie standardowe.

*Uwaga.* W pewnych sytuacjach dzielenie przez odchylenie standardowe nie jest konieczne.

W tym momencie zbiór zdjęć możemy reprezentować jako tablicę  $X$  o wymiarze  $60 \times 50176$ .

Analiza głównych składowych

1. Wykonaj transformację PCA.

2. Jak wyglądała dla tego zbioru macierz kowariancji przed transformacją PCA, a jak po jej wykonaniu?
3. Jak wyglądało średnie zdjęcie, które odjęliśmy od pozostałych, by wycentrować zbiór?
4. Jak wyglądają znalezione nowe wektory bazowe (czyli główne składowe, ang. *principal components*)? Zaprezentuj je posortowane według powiązanej wariancji.  
  
Zauważ, że wektory bazowe też są wektorami z oryginalnej przestrzeni. A że oryginalna przestrzeń zawierała fotografie, to znalezionej "lepszą" bazę możemy również zwizualizować w postaci obrazów, tak jak średnią fotografię z poprzedniego punktu.
5. Zredukuj wymiarowość naszych obserwacji do odpowiednio 3, 9 i 27 najważniejszych cech. Jak wyglądają tak "odchudzone" z wymiarów fotografie? Żeby odpowiedzieć na to pytanie wykonaj poniższe kroki.
6. Przedstaw wykres wariancji wyjaśnionej.

Materiały:

1. <https://github.com/rasbt/machine-learning-book/tree/main/ch05>
2. <https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb>
3. <https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html>