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.

Preprocesing 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 224×224, tak aby wszystkie obrazy miały ten sam rozmiar, równy $224\times224\times3$.
4. Skonwertuj obrazy do skali szarości, tak aby z trójwymiarowego tensorareprezentującego 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 oryginalej 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 60×50176.

Analiza głównych składowych

1. Wykonaj transformację PCA.

2. Jak wyglądała dla tego zbioru macierz kowariancji przed transformacjąPCA, 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 znalezioną ”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:

   a Wyzeruj wartości wszystkich cech poza tą wybraną garstką.

   b Przetransformuj tak zmodyfikowane obserwacje ponownie do oryginalnej bazy (może być konieczne użycie odwrotności macierzy przejścia lub odpowiedniej metody z bibliotecznej implementacji).

   c Dodaj do każdej z nich średni wektor (odwracając wycentrowanie).

   d Przekształć wektor ponownie do kształtu fotografii i wyświetlmy.

   e W praktyce wygląda to tak: robimy PCA (konwersja: fotografia → wektor cech w nowej bazie), usuwamy zbędne cechy (zerując pozostałe), robimy odwrotność PCA (konwersja: zmodyfikowany wektor cech w nowej bazie → fotografia).

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