Analiza głównych składowych

Zadanie 1. Celem zadania jest zapoznanie się z metodą analizy głównych składowych (ang. Principal Component Analysis, PCA). Pracować będziemy na zbiorze `Plantdoc dataset`[1]. Jest to zbiór zdjęć przedstawiający choroby popularnych roślin uprawnych.

Preprocesing danych:

1. Wybierz ze zbioru `Plantdoc` podzbiór kilkudziesięciu (np. 60) zdjęć przedstawiających trzy choroby roślin jednego gatunku (np. 3 choroby ziemniaków lub 3 choroby pomidorów, po 20 zdjęć dla każdej choroby).

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 $224 \times 224 \times 3$.

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 oryginalej tablicy. Funkcja `np.flatten` zwraca kopię tablicy, co zwykle jest niepożądane.

   Każdy obraz powinien być teraz reprezentowany przez wektor o rozmiarze 50 176.

6. Przeprowadź centrowanie zbioru, czyli od każdego obrazu odemij średni obraz. Opcjonalnie możesz jeszcze wykonać dzielenie przez odchylenie standardowe.

   Uwaga. W zależności od sytuacji dzielenie przez odchylenie standardowe jest zalecane lub nie.

   W tym momencie wycentrowany zbiór zdjęć możemy reprezentować jako tablicę X_0 o wymiarze 60×50176.

[1] `https://github.com/pratikkayal/PlantDoc-Dataset`

Analiza głównych składowych

1. Wykonaj transformację PCA. Realizacja możliwa jest na kilka sposobów:
   (a) Poprzez użycie funkcji `sklearn.decomposition.PCA`. Nie mamy jednak głębszego wglądu w działanie metody.
   (b) Poprzez rozkład macierzy X_0 według wartości osobliwych, $X_0 = U\Sigma V^T$, a następnie wyliczenie macierzy $Z_0 = X_0 V$.
   (c) Poprzez wyliczenie wektorów i wartości własnych macierzy kowarianacji, tj. $\frac{1}{n} X_0^T X_0$. Wadą tego podejścia może być duże obciążenie pamięciowe, gdyż rozmiar macierzy kowariancji wynosi $m \times m$, gdzie m jest liczbą cech.
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 (ang. *principal axes*)? Zaprezentuj je posortowane według powiązanej wariancji.

   Zauważ, że wektory bazowe też są wektorami z oryginalnej przestrzeni. Ponieważ oryginalna przestrzeń zawierała zdjęcia, to znalezioną nową, lepszą bazę możemy również zwizualizować w postaci obrazów, tak jak średnią fotografię z poprzedniego punktu.
5. Zredukuj wymiarowość obserwacji do odpowiednio 3, 9 i 27 najważniejszych składowych, czyli cech w nowej bazie (ang. *principal components*). Jak wyglądają tak "odchudzone" z wymiarów zdjęcia? Żeby odpowiedzieć na to pytanie wykonaj poniższe kroki.
   (a) Wyzeruj wszystkie wartości składowych, poza wybranymi najważniejszymi składowymi.
   (b) Przetransformuj tak zmodyfikowane zdjęcia do oryginalnej bazy.
   (c) Do każdego zdjęcia dodaj średni wektor, odwracając wycentrowanie.
   (d) Przekształć wektor do kształtu zdjęcia i wyświetl.
6. Na koniec użyj PCA do zrzutowania zbioru na płaszczyznę.
   (a) Zredukuj wymiarowość do 2 najważniejszych aspektów danych. Nie zeruj odrzucanych cech, zamiast tego skróć wektory (obserwacje powinny stać się wektorami dwuwymiarowymi).
   (b) Użyj tych wektorów 2D jako współrzędnych na płaszczyźnie. Każdą obserwację zaznacz markerem, uzależniając kolor lub kształt markera od rodzaju choroby, którą przedstawiało dane zdjęcie.
7. Przedstaw wykres wariancji wyjaśnionej.