

Tutorial 2 Feature selection and extraction

1. Opisać najważniejsze procedury selekcji cech. Jako przykład przedstawić działanie odpowiednich procedur scikit-learn eliminujących cechy na podstawie wybranego zbioru UCI dataset (np. Wisconsin_cancer, i jeszcze jakiś jeden inny, nie zapomnieć o normalizacji). Przedstawić uporządkowany histogram wag poszczególnych cech.
2. Zbadać jakość klasyfikatorów k-nn (dla różnego $k=1,3,5$) dla pełnego zestawu N cech, dla mniejszego zestawu m cech wybranych w sposób losowy i dla wyselekcjonowanego zestawu m najlepszych cech. Przy pomocy studentów wypełnić tabelkę: ($k=1,3,5$; $m=N,5,2$) dla tych dwóch zbiorów.
3. Opisać Pythonową procedurę liczenia macierzy kowariancji. Pokazać, jak tworzy się „hot map” dla takiej macierzy. Co można z niej odczytać.
4. Opisać działanie transformacji PCA. Opisać Pythonowe procedury realizujące transformację PCA. Pokazać jak wybierane są nowe cechy, kombinacją jakich oryginalnych cech są nowe cechy. Przedstawić (podobnie jak w przypadku selekcji cech) uporządkowany histogram wag (wartości własnych) poszczególnych nowych cech. Porównać do tego, stworzonego wcześniej, dla oryginalnych cech.
5. Zbadać jakość klasyfikatorów k-nn (dla różnego $k=1,3,5$) dla pełnego zestawu N „starych” cech oraz N nowych cech. A także dla mniejszego zestawu m najlepszych „starych” cech i dla wyselekcjonowanego zestawu m „nowych” najlepszych cech (przy pomocy PCA). Przy pomocy studentów wypełnić tabelkę: ($k=1,3,5$; $m=N,5,2$) dla tych dwóch zbiorów.
6. Pokazać wizualizację zbioru danych przy pomocy PCA w 2 wymiarach. Na rysunku zaznaczyć przykłady źle sklasyfikowane i tzw. Outliery (znaleźć procedurę do poszukiwania outlierów!)
7. ZADANIE: W zbiorach MNIST, NMIST, CIFAR-10 . Do dalszych eksperymentów wybrać ilość przykładów nie większy niż $M=2 \times N$!!! (N - ilość cech).
 - a. znaleźć NAJMNIEJ INFORMATYWNE cechy (piksele). Zobrazować je na rysunku, wielkością odpowiadającemu klasyfikowanemu obrazkom.
 - b. Dokonać klasyfikacji k-nn na pełnym zbiorze i zbiorze bez m najmniej informatywnych cech.
 - c. Przetransformować zbiory przy pomocy PCA z N -D do N -D. Jak wyglądają (obrazki) wektory własne odpowiadające największym wartościom własnym. Sprawdzić, czy poprawił się wynik klasyfikacji. Dokonać wizualizacji 2-D przy pomocy PCA.
 - d. Usunąć m najmniej informatywnych cech PCA. Jak wygląda wynik klasyfikacji.
 - e. Wybrać m NAJLEPSZYCH cech PCA. Jak wygląda teraz wynik klasyfikacji.

- f. Wartość m w przypadku wyboru najgorszych cech ma być duże (dla $N=784$ jakieś $m=500$), w przypadku wyboru najlepszych małe ($m=10-20$)
- g. Dokonać klasyfikacji z PCA i bez PCA (na pełnym zbiorze cech i zadanym małym M), ale zwiększając ilość przykładów przy pomocy augmentacji (imgaug).
- h. Wnioski (co lepsze augmentacja czy inżynieria cech?)