Лабораториска Вежба 3

Ненадгледувано Учење

- Целта на оваа лабораториска вежба е да ги испробате техниките од ненадгледувано учење кои ги учевме на даденото податочно множество. За решавање на истата би требало да ви помогнале материјалите од лекциите за кластерирање и димензионалност.

Рокот на оваа лабораториска е се до бранењето на лабораториските кое најверојатно ќе биде во почеток на февруари.

Што точно треба да направите?

1. **Спуштете го податочното множество**.

   http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

   Број на податоци (N) = 1080, број на атрибути (D) = 82 (77). Целта е да се кластерираат податоците. Притоа, поради тоа што целта на лабораториската е ненадгледувано учење, треба да ги тргнете првиот и последните 4 атрибути, и да работите само со 77-те протеински експресии и сами да пробате да ги најдете кластерите кои ги сугерираат класите.

2. **Запознавање со податочното множество и претпроцесирање**. Визуелизирајте го податочното множество во 2 или 3 димензии користејќи PCA и t-SNE. Доколку сметате дека во новиот простор на PCA би добиле подобри резултати (преку Scree plot визуелизација) претворете ги податоците со PCA и намалете го бројот на компоненти на тоа што имплицира scree-plot-от, и продолжете да ги користите трансформираните податоци во кластерирањето.

3. **Кластерирање**.
   Искористете 2 од методите за кластерирање кои ги имаме учено и видете кои се резултатите кои се добиваат. Дали резултатите се слични?
   a) Gaussian Mixture Models
   b) K-Means
   c) Hierarchical Clustering
   d) DBSCAN

4. **Визуелизација на резултатите.**
   Визуелизирајте ги резултатите во 2 или 3 димензии со една боја на секој кластер, повторно користејќи некој од методите како PCA или t-SNE. (t-SNE е препорачливо).

5. **Документација:**

   a. Доколку користевте jupyter, на курсот поставете .ipynb или .html фајл од истиот.
   b. Доколку користевте стандардна работна околина, на курсот поставете .zip кој го содржи .py кодот и screenshots од резултатите и коментарите.