Fundamentals of Machine Learning - 2022

Practice 3 - Unsupervised learning

September 12 2022

This practice is intended for your own exercise and **does not** need to be turned in.

1. Questions

1. What is clustering and in which cases do you think it is convenient to use it?
2. What are strenghts and weaknesses of the k-mean algorithm?
3. What is label propagation? How does it relate to semi-supervised learning?
4. What is anomaly detection and novelty detection? You may need to do some research (bibliography/internet) for this one.
5. What criterium can you use to set the number of dimensions when using PCA?
6. For which particular distance is classical MDS equivalent to PCA?

2. Problems

1. Let's revisit the 'Housing' dataset from practice 2. Use the clustering methods given in class to generate a new categorical feature associated to the distance to the shore. The aim of the exercise is to infer a number of classes and compare it to the one present in the dataset ('ocean_proximity').

   For this, you will need to compute the distance from scratch, and then apply clustering methods (try K-Means, DBScan and Hierarchical Agglomerative Clustering). Compare your results with the classes present in 'ocean_proximity'. Check the effect of adding other features to the clustering process and pay attention to how each of the three methods manage the number of clusters differently.

2. This problem uses a dataset of images of faces (file `'faces_dict.p'` in the `./Datasets/` folder). There are 10 different photos of the faces of 40 subjects and, in each photo, there are properties that change (closed or open eyes, wearing or not wearing glasses, etc). All the images were taken against a dark homogeneous background with the subjects in an upright and frontal position. These are 64 x 64 pixel images (key 'images' in the dictionary object), and each image is flattened to a 1D vector of size 4096 (key 'data' in the dictionary object). The value of the element correspond to a grey level of that pixel. The goal is to train a model that can predict the person at the image.

a) Load the pickle object `'faces_dict.p'`, which contains a dictionary with vectors representing face images as well as integers with the associated target.

b) You already know: separate your dataset. Think about whether it would be convenient to use the stratified sampling (why?).

c) Cluster the images using K-Means, for different k values. Try to quantify how good the clustering is (e.g. using the intertia metric) so as to choose the best k. Then check inside your clusters, do they make some sense?

d) Now, we'll compare classifying with and without cluster information. First, train a classifier to predict which person is represented in each picture (no cluster information) and see which validation error you get.

e) Then, use K-Means as a dimensionality reduction tool and train a classifier on the reduced set (i.e., using the distances to the centroids). What performance do you reach compared to the image-only information?

3. We further explore the faces dataset with two visualization methods, t-SNE and UMAP. Preprocess the dataset by applying PCA and reducing the number of dimensions to 30. Next, compare the outcome of applying t-SNE and UMAP. Comment on the effect of the different tuning parameters (you can revisit the associated lecture) and compare the results qualitatively.