Domanda **1**

Risposta non ancora data

Punteggio max.: 31

Find ***clusters*** for this dataset, according to the directions below.

The solution must be produced as a Python Notebook.

The notebook must include appropriate comments and must operate as follows:

	Task	Point(s)
1	Load the data, the attributes are all qualitative and there is no label column, show the shape of the data and for each column show the frequencies of each distinct value (hint: you can use the numpy function **unique(x, return_counts = True)**	4
2	Do the appropriate pre-processing in order to use the sklearn algorithms on this dataset; the values are qualitative and must be considered as *nominal*	4
3	As an external *background knowledge*, we are told that for this dataset a requirement for a good clustering scheme is to have clusters with *low deviation in sizes*, e.g. a scheme with cluster sizes (330, 670) is less acceptable than one with (333, 333, 334). In order to obtain this, we want to compute, for each clustering scheme with **n_clusters** clusters and represented by the labels in **y**, a *size deviation index* with the formula **np.sqrt(np.unique(y, return_counts = True)[1].var())/n_clusters** For varying number of clusters fit KMeans and compute the inertia, the silhouette index and the above-mentioned *size deviation index*.	4
4	In this dataset the *elbow method* will show an almost "vanishing" elbow for inertia, and the silhouette is totally non-effective. Make two plots, one with *inertia* and *silhouette*, another with *inertia* and *size deviation index*, then decide the best number of clusters and refit KMeans using that value	4
5	Fit another clustering method of your choice, trying to reproduce the same number of clusters you have chosen in the previous step	5
6	Compare the results of the two clustering showing the result of `sklearn.metrics.cluster.pair_confusion_matrix` and `sklearn.metrics.adjusted_rand_score`	4
Q	Quality of the code: Include appropriate comments with reference to the numbered requirements Useless cells, pieces of code and non-required output will be penalized Remove the code you use for testing and inspecting the variables during the development Naming style of variables must be uniform and in English Bad indentation and messy code will be penalized	6

Additional directions, the assignments not compliant with the rules below will not be considered

1. The notebook name must be ***emailusername.ipynb*** in lowercase letters

a. E.G. if your email is mario.rossi45@studio.unibo.it the notebook filename will be mario.rossi45.ipynb

2. The first cell must contain as a comment the student first name, last name and email

3. The notebook must directly access the data in the same folder of the notebook (i.e. the filename must not be preceded by a path)

4. Upload the notebook only to eol, any other way of submitting the notebook will be ignored

Cooperative work will be **heavily sanctioned**

The candidate can freely access any kind of materials

Dimensione massima dei file: 100MB, numero massimo di file: 1

File

Per caricare file, trascinali e rilasciali qui.

Vai a...

Machine Learning Theory ►