TDA 231 Machine Learning: Homework 5 Solution Sketch

Aristide Tossou

March 13, 2017

1 Practical problems

Problem 1.1 [k-Means Implementation, 8 points]

(a) (3 points)
- Make sure the centroid are updated correctly and the points updated to their closest centroid. A correct implementation grants half the grade. Further grading points can be gained by handling the notes below.
- The initial centroid should be generated randomly from the points already present in the dataset. Special care must be taken such as to avoid duplicate centroids. It means the random samples should be generated without replacement. This could be done for example in Matlab using the function *randperm* or *datasample* with the parameter *Replace* set to false.
- The stopping condition of the main loop should avoid comparing exactly a *double* type. In particular, the students are expected to compare if the points assignment (of *integer* type) changed. If the centroids location are directly compared, a small epsilon difference should be allowed.
- The code should be commented and variables name intuitive and easy to understand. Care should be made so as to avoid having an array variable changes its size at each iteration.

(b) (1 point)
- The plot should have a legend or a short explanation.
- The assignment change from the second iteration should be generated from the same run of the implemented kmeans. It means, it is not allowed to run two different instances of the kmeans, one with maximum iteration to 2 and the other till convergence

(c) (2 points)
- Special care must be taken so as to implement correctly the gaussian kernel as well as the distance of each point to the centroids.
- The code should be commented and variables name intuitive and easy to understand. Care should be made so as to avoid having an array variable change its size at each iteration.

(d) (2 points)
- Here it is expected that the plot of the RBF kernel kmeans should correctly clustered the points. Two co-centric circles are expected. The inner circle should have points in the same color A. The outer circle points should have all the same color B (color A different than color B)

Problem 1.2 [k-Means Analysis, 12 points]

(a) (2 points)
- For each cluster, the 10 closest words should be reported in the report for a total of 100 words. Failing to do that will grant half points and only if the implementation looks reasonable.

(b) (8 points)
- The implementation of N_0 should be straight forward: Run the built-in matlab kmeans. The cluster where the word *cavalry* belong to, should be found. Once it is found, one should retrieve all words belonging to that cluster. Let's call N the number of words belonging to the cluster of *cavalry* (including *cavalry*). N_0 will be equal to all pairs of such words which is: $N_0 = \frac{N \cdot (N-1)}{2}$

- Here one should run another instance of kmeans. A correct algorithm to implement N_1 should loop through all pairs of words found in the previous step. For each pair of words w_1 and w_2, one should increment N_1 by 1 if both w_1 and w_2 belong to the same cluster (regardless of what that cluster is).

  It is important here to note that we don't just pick the words being in the same cluster as *cavalry* anymore. Words could belong to any of the 10 clusters.
- The correct fraction usually belong approximately to the interval [0.53, 0.61]. It is important to note that the fraction should never be greater than 1 in a correct implementation as N_1 should always be lower than N_0.
- For the conclusion about clustering, this indicates that the clustering depends quite a lot on the initialization of the cluster centroids. This in turn means that some words likely belong equally well in several different topics.
- 3 points are granted for the conclusion about clustering, 2 points for a correct implementation of N_0, 3 points for a correct implementation of N_1.
- Coding style or comments not required for the implementation here. But submitting the code is required.
- (2 points) – Students are expected to combine the usage of the tsne function as well as kmeans to come up with a nice plot.