Big data computing - 2019/2020

Homework 1

Due date: Sunday, November 10th, 11.59pm

Assignment 1

Suppose you are using Locality Sensitive Hashing for Jaccard similarity. I.e., each data point can be seen as a set of integer values. Assume, for the purpose of the analysis, that the hash functions you are using behave like an ideal, min-wise independent family. You are given the following constraints:

- You have two thresholds θ_1 and θ_2, with $\theta_1 > \theta_2$. Two sets X and Y are considered "similar" (i.e., they are a *true positive pair*), whenever $Jaccard(X, Y) \geq \theta_1$, while they are not similar (i.e., a *true negative pair*), whenever $Jaccard(X, Y) < \theta_2$.
- Given a true positive pair (X, Y), you want the probability of considering them as a negative pair (false negative probability) to be at most a value p_1.
- Given a true negative pair (X, Y), you want the probability of considering it as a positive pair (false positive probability) to be at most p_2.
- We are not interested in the probability of misclassifying pairs with Jaccard similarity in the interval $[\theta_2, \theta_1)$.

Assume your signature matrix is going to have m rows.

1. Work out the equations that give the relationships between the parameters that describe the requirements you are given and the numbers r and b of rows and bands, so as to achieve false negative and false positive probabilities p_1 and p_2 respectively.
2. Assuming $\theta_1 = 0.7$, $\theta_2 = 0.5$, $p_1 = p_2 = 0.01$, try to identify the minimum value of m that allows to meet these requirements. Given the strong non-linearity of the equations you are working with, you are advised to proceed numerically, by trial and error.

Assignment 2

The goal of this assignment is to highlight issues that arise in unsupervised classification (clustering) when the number of dimensions of the feature space is high (*the curse of dimensionality*), a topic we discussed in class. In the simplified scenario we consider here, we have a set of points that might be effectively clustered if they were first projected onto the right subspace. Unfortunately, finding the right subspace onto which to project is hard in general.

- To begin, consider n points in $\mathbb{R}^d$, each belonging to one of two possible clusters C_1 and C_2. Let $a, b \in [0, 1]$, with $|a - b| > 0$. A point $\mathbf{x} \in C_1$ is generated as follows:

$$\mathbf{x}_j = \begin{cases} a, \ j = i, \\ \text{distributed u.a.r. in } [0, 1], j \neq i \end{cases}$$

  Likewise, a point $\mathbf{y} \in C_2$ is generated in the same way, with the only difference that, this time $\mathbf{y}_i = b$. In words, the values along all dimensions but the i-th are distributed independently and uniformly at random in $[0, 1]$ for both points in C_1 and C_2, while the value on the i-th dimension is *deterministically* a for points in C_1 and b for points in C_2.

Provide an answer (with convincing proofs) to the following questions:

1. What is the expected distance between two points from the same cluster?
2. What is the expected distance between two points from different clusters?
3. What do the results from points 1 and 2 suggest as regards the possibility that a clustering algorithm will effective identify the two clusters C_1 and C_2? Please elaborate.
4. Assume next that you know the dimension i. What would be an effective strategy to cluster the points in this case (assuming $|a - b|$ is "not too small")? Assume you have a good clustering algorithm (such as k-means++) to use as a subroutine.

- Discuss why the strategy you identified in point 4 above may not be feasible in general.