Big data computing - 2020/2021

Homework 1

Due date: Thursday, December 10th, 11.59pm

You must hand your homework by the due date and time by an email to the instructor (becchetti@diag.uniroma1.it) that will contain as attachment a pdf with a short (1-2 pages) report on your answers to the theory questions.

Important: the subject of your email should be: [BD] [Last_name First_name] HW1

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 $[\theta_2, \theta_1)$.

Task. Assume your signature matrix is going to have m rows. 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.

Assignment 2

- Assume that A is a *square invertible*, n-dimensional matrix, with SVD $A = U\Sigma V^T = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$. Show that the inverse of A is $B = \sum_{i=1}^{n} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^T$.

  Hint: recall the properties of the matrices U and V.

- Suppose again that A is square and has SVD $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, but this time A is *not necessarily invertible*. Let again $B = \sum_{i=1}^{r} \frac{1}{\sigma_i} \mathbf{v}_i \mathbf{u}_i^T$. Show that $BA\mathbf{x} = \mathbf{x}$, for every vector $\mathbf{x}$ that can be expressed as a linear combination of the right singular vectors of A. I.e., we consider vectors of the form $\mathbf{x} = \sum_{i=1}^{r} \alpha_i \mathbf{v}_i$ (we say that $\mathbf{x}$ is in the *span* of the right singular vectors of A). B is called the *pseudo-inverse* of A and can play the role of A^{-1} in many applications.