EE627A Final

Name:

Q1. a. Consider a matrix $\mathbf{X}$ contains two column vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$,

$$\mathbf{X} = [\mathbf{x}_1,\ \mathbf{x}_2]$$

Use the principal component analysis to find the first principal component column vector $\mathbf{y}$, which is a linear combination of $\mathbf{x}_1$ and $\mathbf{x}_2$, i.e.,

$$\mathbf{y}_1 = a\mathbf{x}_1 + b\mathbf{x}_2,$$

where $a = \frac{1}{\sqrt{2}}$ and $b = \frac{-1}{\sqrt{2}}$.
What are the linear combination factors $\{c,\ d\}$ for the second principle component $\mathbf{y}_2 = c\mathbf{x}_1 + d\mathbf{x}_2$.

b. Given 3 data points in 2-d space, (1, 1), (2, 2) and (-3, -3), what is the first principle component?

Q2. a What are the two major features with Hadoop?

b Explain the general data flows for MapReduce?

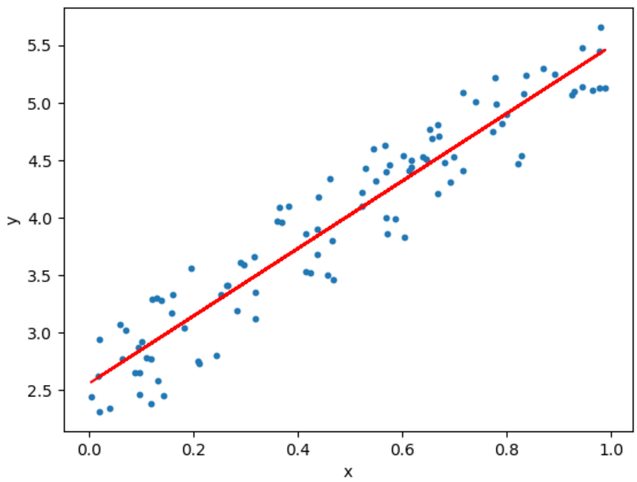
Q3. a What is RDD in Spark?

b What are the two types of operations with RDD?

c Explain why there is the lazy evaluation with RDD.

Q4. For a given rating matrix $\mathbf{R} \in \mathcal{R}^{N \times M}$, we can use matrix factorization to form $\mathbf{R} \approx \mathbf{P}\mathbf{Q}^T$, where $\mathbf{P} \in \mathcal{R}^{N \times K}$ and $\mathbf{Q} \in \mathcal{R}^{M \times K}$.

- For example, we have a user-rating matrix $\mathbf{R}$. How to deal with these empty elements during the matrix factorization? (No calculations needed. Just show the conceptual steps.)
- For these empty elements, how to use the matrix factorization to estimate them? (No calculations needed. Just show the conceptual steps.)



Q5. We have learned in class that, for a given $N \times K$ matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ \cdots, \mathbf{x}_K]$ and a given $N \times 1$ vector $\mathbf{y}$, if we like to find a linear combined vector $\mathbf{Xa} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_K\mathbf{x}_K$ to approximate $\mathbf{y}$, i.e.,

$$\mathbf{Xa} \approx \mathbf{y}$$

where $\mathbf{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_K \end{bmatrix}$ is a $K \times 1$ vector. The least-squares(LS) solution for this optimization problem is

$$\arg\min_{\mathbf{a}} \|\mathbf{Xa} - \mathbf{y}\|^2 = \mathbf{a}_{\text{LS}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

Now we have a set of observations (x_i, y_i), $i = 1, 2, \cdots, N$, we like to design a linear regression models using the above-mentioned classic Least Squares (LS) method.

$$y_i = ax_i + b, \quad i = 1, 2, \cdots, N$$

Derive your LS formula to calculate the parameters $\{a, b\}$ in this model.

(For example: in the above scatter plot, the linear regression is to find a straight line to fit the observations (x_i, y_i), where a is the slope and b is the intercept.)