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?

a. $y_1 \rightarrow y_2$

$y_1 = aX_1 + bX_2$

$y_2 = cX_1 + dX_2$

$\because$ y_1 is perpendicular to y_2

$\therefore$ $ac + bd = 0 \Rightarrow \frac{1}{\sqrt{2}}c - \frac{1}{\sqrt{2}}d = 0$

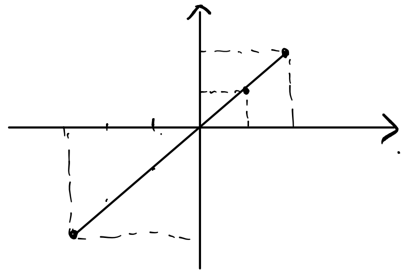
$\therefore$ $c = d$

$[c, d]^T[c, d] = 1 \Rightarrow c^2 + d^2 = 1$

$\therefore$ $c = d = \pm\frac{1}{\sqrt{2}}$

$y_2 = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2$ or $y_2 = -\frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2$

C. (1, 1) (2, 2) (-3, -3)



Obviously. $A = B$

And $\Sigma w^2 = 1 \Rightarrow A^2 + B^2 = 1$

$\therefore$ $A = B = \pm\frac{1}{\sqrt{2}}$ $\quad Var(y) = \begin{bmatrix}A\\B\end{bmatrix}^T \cdot 1 \begin{bmatrix}A\\B\end{bmatrix} = 1$

$\therefore$ $y = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2$ or $y = -\frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2$

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

b Explain the general data flows for MapReduce?

a. I. Hadoop Distributed File System

HDFs provide a storage layer for hadoop. It's hadoop's own rack-aware file system. which is a UNIX-based data storage Layer. HDFS is the partition of data and computation across many hosts, and the execution of application compueations in parallel, close to their data.

II. MapReduce

MapReduce is the heart of hadoop. It's a programming model for processing large datasets distributed on a large cluster. Its programming paradigm allows performing massive data processing across thousand of servers configured. with hadoop clusters.

b. I. Preloading data in HDFS.

II. Running MapReduceng by calling Driver.

III. Reading of input data by the Mappers, which results in the splitting of the data execution of Mapper custom logic and the generation of intermediate key-value pairs.

IV. Executing Combiner and the shuffle phase to optimize the overall Hadoop MapReduce processing.

VI. Sorting and provide of intermedrate key-value pairs to the Reduce phase The Redue phase is then executed. Reducers take these partitioned key-value pairs and aggregate them based on Reducing logic.

VII. The final output data is stored at HDFS.

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.

a. RDD is simply a distributed collection of elements.
-'s the core concept in Spark.

b. Transformations and Actions.

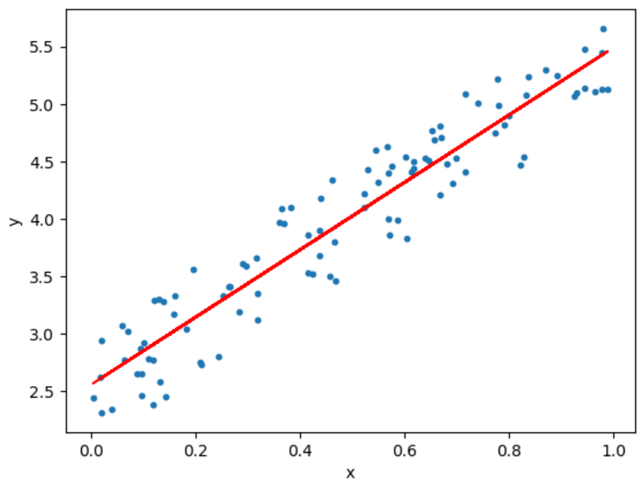
c. Spark internally records metadata that this operation has been requested. Rather than thinking of an RDD as containing specific data, it is best to think of each RDD as consisting of instructions on how to compute the data that we build up through transformations.

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.)

a. For these empty elements in matrix R. We should let them to be zero. And in the follow steps, we don't have to calculate the error between etimated rating and the real rating. for these element.

b. First randomly generate matrix $P_{N \times K}$ and $Q_{M \times K}$. then caculate all error between estimated rating and the real rating except those empty element. After that, we have to update P and Q matrix by formulation. Those two formulation will make the $R \approx PQ^T$ (only for those not empty elements). The same time, we can use P and Q to calculate all those empty element.



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.)

For $y_i = a x_i + b$.

We need $\min \sum (y_i - \hat{y}_i)^2$ which $\hat{y}_i$ is the estimated value. y_i is the real value.

Let $\varphi = \sum (y_i - \hat{y}_i)^2$

$= \sum (a x_i + b - y_i)^2$

To minimize φ.

$$\frac{\partial \varphi}{\partial a} = \sum 2x_i (b + a x_i - y_i) \Rightarrow (\sum x_i) b + (\sum x_i^2) a = \sum (x_i \cdot y_i) \quad ①$$

$$\frac{\partial \varphi}{\partial b} = \sum 2 (b + a x_i - y_i) \Rightarrow nb + (\sum x_i) a = \sum y_i \quad ②$$

$$① \Rightarrow (\sum x_i) b + (\sum x_i^2) a = \sum (x_i \cdot y_i) \Rightarrow a = n\sum(x_i \cdot y_i) - \frac{\sum x_i \cdot \sum y_i}{n \sum x_i^2 - \sum x_i \cdot \sum x_i}$$

$$② \Rightarrow n a_0 + (\sum x_i) a_1 = \sum y_i \Rightarrow b = \frac{1}{n} \cdot (\sum y_i - a \sum x_i)$$

$$\therefore \quad a = n \cdot \sum (x_i \cdot y_i) - \frac{\sum x_i \cdot \sum y_i}{n \sum x_i^2 - \sum x_i \cdot \sum x_i}$$

$$b = \frac{1}{n} (\sum y_i - a \sum x_i)$$

or

$$a = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

in which $\bar{x}$ is the mean of x_i, $\bar{y}$ is the mean of y_i.

$$b = \bar{y} - a\bar{x}$$