INDIAN STATISTICAL INSTITUTE

Mid-Semester Examination : 2015–16

Course : Post Graduate Diploma in Business Analytics (First Year)

Subject : Computing for Data Sciences : BAISI–4 for PGDBA–I

Date : 11 September 2015 | Maximum Marks : 90 | Duration : 3 Hours

---

Problem A **[30]**

1. Define *norm* on the n-dimensional vector space $\mathbb{R}^n$. Given a norm $\rho(\cdot)$ on $\mathbb{R}^n$, define a related notion of *distance* between any two vectors in $\mathbb{R}^n$, and state its properties. [2 + 3]

2. Let the ℓ^p norm of a vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ in $\mathbb{R}^n$ be defined as $||\mathbf{x}||_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$. Comment on the significance of the ℓ^1 and ℓ^2 norms of $\mathbf{x}$ in $\mathbb{R}^n$, in terms of the geometrical depiction of the unit vectors in $\mathbb{R}^n$. Is there any relation between the ℓ^1 and ℓ^2 norms of $\mathbf{x}$ and the statistical properties of the set of real numbers $\{x_1, x_2, \ldots, x_n\}$? [5 + 5]

3. Let an *inner product* on $\mathbb{R}^n$ be defined as the *dot product* of two vectors: $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^n x_i y_i$, where $\mathbf{x} = [x_1, x_2, \ldots, x_n]^T$ and $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$. What is the geometrical significance of this inner product in $\mathbb{R}^n$? Is there any statistical significance of this inner product in connection with the sets of real numbers $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$? [2 + 3]

4. Suppose that you have an $n \times p$ matrix $\mathbf{X}$ representing a dataset, comprising of n independent observations along p features. Assume that the dataset is *centered*, that is, the mean of values along each column in $\mathbf{X}$ is zero. Comment on the statistical significance of the matrix $\mathbf{X}^T\mathbf{X}$ in terms of the features and observations in the dataset. [5]

5. What can you say about the dataset if the matrix $\mathbf{X}^T\mathbf{X}$ is diagonal? What can you say if the matrix $\mathbf{X}^T\mathbf{X}$ is block-diagonal, with k distinct blocks along the main diagonal? [2 + 3]

Problem B **[30]**

1. Describe the role of an $m \times n$ matrix $\mathbf{X}$ as a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$. Your description should include the conceptual notions of the fundamental subspaces – RowSpace, ColSpace and NullSpace of $\mathbf{X}$, as well as Rank of $\mathbf{X}$. [7]

2. Given the fundamental subspaces of an $m \times n$ matrix $\mathbf{X}$, how do you determine the following?
(a) Whether the matrix is a 1-to-1 linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$;
(b) Whether the matrix is an onto linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$;
(c) Whether the matrix is an invertible linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$. [3]

3. Suppose that the *full* Singular Value Decomposition of an $m \times n$ matrix $\mathbf{X}$ results in:

$$\mathbf{X} = \begin{bmatrix} | & & | & & | \\ \mathbf{u}_1 & \cdots & \mathbf{u}_r & \cdots & \mathbf{u}_m \\ | & & | & & | \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \begin{bmatrix} | & & | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_r & \cdots & \mathbf{v}_n \\ | & & | & & | \end{bmatrix}^T$$

Represent this decomposition as $\mathbf{X} = \mathbf{U\Sigma V}^T$, and comment on the dimension of each matrix in this representation. Discuss the connection of these matrices with the fundamental subspaces of $\mathbf{X}$. How can you determine the Rank of $\mathbf{X}$ given this SVD representation? [3 + 5 + 2]

4. As per the above representation of the SVD of $\mathbf{X}$, determine the dimension and rank of each of the matrices $\mathbf{Z}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where $1 \le i \le r$. Is there a way to reconstruct the original matrix $\mathbf{X}$ given the matrices $\mathbf{Z}_i$ for $1 \le i \le r$? [3 + 2]

5. Is there a way to reconstruct the original matrix $\mathbf{X}$ given the matrices $\mathbf{Z}_i$ for $1 \le i \le k$, where k is strictly less than r? If so, provide such a construction. If not, provide an *approximate* reconstruction of $\mathbf{X}$ using the available matrices $\mathbf{Z}_i$ for $1 \le i \le k$, and comment on the quality of such an approximation. [2 + 3]

Problem C [15]

Represent a book in the form of an $m \times n$ matrix $\mathbf{B}$, where m is the total number of sentences in the book and n is the total number of distinct words in the book, such that the entry $\mathbf{B}[i,j]$ in this matrix represents the frequency of occurrence of the j-th word W_j in the i-th sentence S_i.

Importance of the words and sentences are denoted by *scores*. The score u_i of S_i is equal to the sum of scores of the words in it, weighted by the frequencies of occurrence. The score v_j of W_j is equal to the sum of scores of the sentences it is contained in, weighted by the frequencies of occurrence.

$$u_i = \sum_{j=1}^{n} \mathbf{B}[i,j] \cdot v_j \quad \text{for } i = 1, 2, \ldots, m \qquad\qquad v_j = \sum_{i=1}^{m} \mathbf{B}[i,j] \cdot u_i \quad \text{for } j = 1, 2, \ldots, n$$

Devise an efficient strategy to identify 10 *keywords* (i.e., the most important words) from the book.

Problem D [15]

Suppose that you have a dataset where m individuals have reviewed a collection of n movies, and have provided scores (between 0 to 9, say) for each one. Suppose that I have also watched and reviewed some (not all) of these n movies, and you know my scores. Devise a strategy to suggest movies for me, from within the same set of the n movies, which I have not watched, but I may like.

Answer ALL questions, respecting the order of sub-questions. Problems C and D will be considered for bonus marks.