Computing for Data Sciences

Mid-Sem : Hints, Answers and Pointers

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]

    Answer: A *norm* on the n-dimensional vector space $\mathbb{R}^n$ is a function $\rho : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties.

    (a) $\rho(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, with $\rho(\mathbf{x}) = 0$ if and only if $\mathbf{x} = \mathbf{0}$.
    (b) $\rho(a \cdot \mathbf{x}) = |a| \cdot \rho(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^n$ and all $a \in \mathbb{R}$.
    (c) $\rho(\mathbf{x} + \mathbf{y}) \leq \rho(\mathbf{x}) + \rho(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

    Answer: A *distance* defined in relation to ρ on $\mathbb{R}^n$ is a function $\delta : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that $\delta(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{x} - \mathbf{y})$. The distance δ satisfies the following properties.

    (a) $\delta(\mathbf{x}, \mathbf{y}) \geq 0$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, with $\delta(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$.
    (b) $\delta(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{y}, \mathbf{x})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
    (c) $\delta(\mathbf{x}, \mathbf{y}) \leq \delta(\mathbf{x}, \mathbf{z}) + \delta(\mathbf{z}, \mathbf{y})$ for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.

    Ref: Lecture 1, 2

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 $||\mathbf{x}||_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$. 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]

    Pointer: For the geometrical significance, note that just the depiction of unit vectors is required. Thus, one may discuss the geometrical appearance of the unit sphere in each of these norms, and that should be enough.

    Ref: Lecture 2

    Answer: The ℓ^1 norm $||\mathbf{x}||_1 = \sum_{i=1}^n |x_i|$ is identical to the Mean Deviation (scaled by n) of the set of real numbers $\{x_1, x_2, \ldots, x_n\}$, in case the mean of these numbers is zero, that is, in case these numbers are *centered*. Similarly, the ℓ^2 norm $||\mathbf{x}||_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$ is identical to the Standard Deviation (scaled by n) of the set of real numbers $\{x_1, x_2, \ldots, x_n\}$, in case the numbers are *centered*.

    Ref: Lecture 8

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]

   Pointer: For the geometric significance, one may discuss the dot product in terms of the projection of the vector $\mathbf{x}$ on $\mathbf{y}$, or the other way round. In addition, one may write the geometric form of the dot product $\mathbf{x} \cdot \mathbf{y}$ as $||\mathbf{x}||_2 ||\mathbf{y}||_2 \cos\theta$, where θ denotes the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$.

   Ref: Lecture 3

   Answer: The *dot product* of two vectors $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i$ is identical to the Covariance (scaled by n) between the sets of real numbers $\{x_1, x_2, \ldots, x_n\}$ and $\{y_1, y_2, \ldots, y_n\}$, in case both the sets are *centered*, i.e., the mean of each set is zero.

   Ref: Lecture 8

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]

   Pointer: Note that $\mathbf{X}$ may be considered as a set of columns $\mathbf{F}_i$, for $i = 1, 2, \ldots, p$, representing the p features in the dataset, with each feature properly *centered*. Thus in the matrix $\mathbf{X}^T\mathbf{X}$, the diagonal terms $\mathbf{X}^T\mathbf{X}[i,i] = \mathbf{F}_i \cdot \mathbf{F}_i = ||\mathbf{F}_i||_2^2$ denote the Variance (scaled) of the individual features $\mathbf{F}_i$, and the other terms $\mathbf{X}^T\mathbf{X}[i,j] = \mathbf{F}_i \cdot \mathbf{F}_j$ denote the Covariance (scaled) between features $\mathbf{F}_i$ and $\mathbf{F}_j$, where $i, j = 1, 2, \ldots, p$. This is why the matrix $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ is called the *covariance matrix* of $\mathbf{X}$.

   Ref: Lecture 8

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]

   Pointer: If the matrix $\mathbf{X}^T\mathbf{X}$ is diagonal, then covariance $\mathbf{F}_i \cdot \mathbf{F}_j$ between features $\mathbf{F}_i$ and $\mathbf{F}_j$ is zero for all $i, j = 1, 2, \ldots, p$. This means that the features are uncorrelated.

   Ref: Lecture 8

   Pointer: If the matrix $\mathbf{X}^T\mathbf{X}$ is block-diagonal, then the covariance $\mathbf{F}_i \cdot \mathbf{F}_j$ between features $\mathbf{F}_i$ and $\mathbf{F}_j$ is zero *between* blocks, but non-zero *within* blocks. In other words, the features across any two blocks are uncorrelated, but the features within a specific block are correlated. Such a structure of $\mathbf{X}^T\mathbf{X}$, with k distinct blocks along the main diagonal, automatically denotes the presence of k clusters within the features, and a strong tendency of correlation within each block (cluster) indicates that there may exist a good k-dimensional approximation of the original dataset $\mathbf{X}$.

   Ref: Lecture 8, 9

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]

   Pointer: This is straight from the Lectures 3 and 4. One may provide a concise summary of the action of a matrix $\mathbf{X}$ as a linear operator $X : \mathbb{R}^n \to \mathbb{R}^m$, and discuss the notions of each of the *fundamental subspaces*. Additionally, one may extend the discussion to the Singular Value Decomposition of $\mathbf{X}$, and provide the basis for each subspace as discussed above. No proof is required, but a figure would be great.

   Ref: Lecture 3, 4

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]

   Pointer: The matrix $\mathbf{X}$ is a 1-to-1 linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ if and only if $m \geq n$ and NullSpace of $\mathbf{X}$ is zero-dimensional; more precisely, a singleton set $\{\mathbf{0}\}$ in $\mathbb{R}^n$. The matrix $\mathbf{X}$ is an onto linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ if and only if $m \leq n$ and NullSpace of $\mathbf{X}^T$ is zero-dimensional; more precisely, a singleton set $\{\mathbf{0}\}$ in $\mathbb{R}^m$. The matrix $\mathbf{X}$ is an invertible linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ if and only if $m = n$ and both NullSpace of $\mathbf{X}$ and NullSpace of $\mathbf{X}^T$ are zero-dimensional. One may of course answer the same questions using the RowSpace, ColSpace or Rank of $\mathbf{X}$.

   Ref: Lecture 3, 4

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 $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, 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]

   Answer: Dimension of $\mathbf{U}$ is $m \times m$, as there are exactly m vectors $\mathbf{u}_i$, each belonging to the range $\mathbb{R}^m$. Dimension of $\mathbf{V}$ is $n \times n$, as there are exactly n vectors $\mathbf{v}_i$, each belonging to the domain $\mathbb{R}^n$. Thus, the dimension of the central matrix $\mathbf{\Sigma}$ must be $m \times n$, where the top-left $r \times r$ square block of $\mathbf{\Sigma}$ is a diagonal matrix with entries $\{\sigma_1, \ldots, \sigma_r\}$, and the remaining entries of $\mathbf{\Sigma}$ are all zero.

Pointer: The matrix $\mathbf{\Sigma}$ with its top-left $r \times r$ diagonal block indicates that the Rank of $\mathbf{X}$ is r. The set of first r vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ in $\mathbf{U}$ constitute an orthonormal basis for the ColSpace of $\mathbf{X}$. The remaining $(m-r)$ vectors $\{\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m\}$ in $\mathbf{U}$ constitute an orthonormal basis for the NullSpace of $\mathbf{X}^T$. The set of first r vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ in $\mathbf{V}$ constitute an orthonormal basis for the RowSpace of $\mathbf{X}$. The remaining $(n-r)$ vectors $\{\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n\}$ in $\mathbf{V}$ constitute an orthonormal basis for the NullSpace of $\mathbf{X}$. Thus the full SVD of a matrix $\mathbf{X}$ completely defines its fundamental subspaces.

Pointer: Given this SVD representation of $\mathbf{X}$, the Rank of $\mathbf{X}$ is exactly equal to the number of non-zero entries in $\mathbf{\Sigma}$, that is, r. This is in fact the dimension of the RowSpace as well as the ColSpace of $\mathbf{X}$, as defined by $\mathbf{U}$ and $\mathbf{V}$ in the SVD.

Ref: Lecture 4, 5

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]

Pointer: Note that $\mathbf{u}_i \mathbf{v}_i^T$ is the *outer product* of the $m \times 1$ vector $\mathbf{u}_i$ with the $n \times 1$ vector $\mathbf{v}_i$, and σ_i is just a real number (scalar). Thus, the dimension of $\mathbf{Z}_i = \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is $m \times n$, same as that of $\mathbf{X}$, for each $1 \le i \le r$. The rank of each $\mathbf{Z}_i$ is 1, as the SVD of $\mathbf{Z}_i$ will return the decomposition $\mathbf{Z}_i = [\mathbf{u}_i]_{m\times 1}[\sigma_i]_{1\times 1}[\mathbf{v}_i]_{1\times n}^T$, with a single non-zero element σ_i in the matrix $\mathbf{\Sigma}$.

Pointer: The SVD representation of $\mathbf{X}$ directly provides a relation between $\mathbf{X}$ and the matrices $\mathbf{Z}_i$, as follows: $\mathbf{X} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = \sum_{i=1}^{r} \mathbf{Z}_i$. This is the required way to reconstruct $\mathbf{X}$, as we know all $\mathbf{Z}_i$ for $1 \le i \le r$. Note that $\mathbf{X}$ is a rank r matrix, as obtained from its SVD above, and each $\mathbf{Z}_i$ is a rank 1 matrix, as proved above. Thus, it is quite interesting to note that the proposed reconstruction in effect takes the sum of r rank 1 matrices $\mathbf{Z}_i$ to reconstruct a single rank k matrix $\mathbf{X}$.

Ref: Lecture 9, 10

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]

Answer: Note that $\mathbf{X}$ is a rank r matrix and each $\mathbf{Z}_i$ is a rank 1 matrix. Thus, if we are provided with k number of rank 1 matrices $\mathbf{Z}_i$, for $1 \le i \le k$, where $k < r$, it is *not possible* to completely reconstruct the rank r matrix $\mathbf{X}$.

Pointer: One may construct a rank k matrix $\mathbf{X}_k = \sum_{i=1}^{k} \mathbf{Z}_i = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, and consider $\mathbf{X}_k$ as an approximation of $\mathbf{X}$. Now one may judge the quality of this approximation quantitatively using a proper notion of *distance* between the matrices $\mathbf{X}_k$ and $\mathbf{X}$. The notion of such a distance may be formulated from the Frobenius norm of matrices: $||\mathbf{X}||_F = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{X}[i, j]^2$, and one would find that $\mathbf{X}_k$ is the *best* rank k approximation of the rank r matrix $\mathbf{X}$ in terms of the Frobenius norm.

Ref: Lecture 9, 10

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.

> **Hint:** Summations are simply a specific way to represent matrix-vector multiplications. Look past the individual sums mentioned in the problem to observe the broad picture. If you construct the vectors $\mathbf{u} = [u_1, u_2, \ldots, u_m]^T$ and $\mathbf{v} = [v_1, v_2, \ldots, v_n]^T$, then the above problem translates to the following set of matrix equations: $\mathbf{u} = \mathbf{B}\mathbf{v}$ and $\mathbf{v} = \mathbf{B}^T\mathbf{u}$. These two equations denote a cyclic mutual dependence between $\mathbf{u}$ and $\mathbf{v}$, and if we put one equation into another, we obtain $\mathbf{u} = (\mathbf{B}\mathbf{B}^T)\mathbf{u}$ and $\mathbf{v} = (\mathbf{B}^T\mathbf{B})\mathbf{v}$. Looks familiar?

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.

> **Hint:** To identify which movies I might like to watch, the goal is to identify the cluster of individuals to whom I am 'closest' regarding the taste for movies. Suppose that you know that I have watched p movies out of the n, and you have my rating vector, which is of dimension $1 \times p$. You can take only these specific p columns from the complete rating matrix of dimension $m \times n$ obtained from the reviewers, and try to find a cluster of reviewers who are 'closest' to me. Once you find such a cluster of reviewers, you may simply suggest me the movies from the $(n - p)$ ones that I have not watched, following the order (highest to lowest) in which this specific cluster of reviewers have rated them.
>
> In addition, one may partition the n movies in the collection into k genres, which are more or less uncorrelated – Action, Drama, Comedy, Romance, Noir, etc. One may now treat these genres as features in the dataset (instead of the movies). It may be observed that an individual likes movies of the same type(s) in general, that is, the preferences are generally based on genres. Thus, it is likely that you can reduce the huge dimension n (of movies in the collection) to a somewhat manageable p (the number of genres).