Consider we are using PCA to compress face images using top K eigenvectors and then we do the reconstruction. Then

(A) **[Ans]** Compression (for face images) is lossy

(B) Compression (for face images) is lossless

(C) **[Ans]** Reconstruction will be bad for non-face images (say buildings)

(D) Reconstruction will be good for non-face images (say buildings)

(E) None of these

Consider we are dong PCA to go from R^2 data to R^1. Consider each point is denoted by (X_i, Y_i). Then in which of these situations will PCA work reasonably well:

(A) **[Ans]** $Y_i = X_i + 10$

(B) **[Ans]** $Y_i = X_i + 10 + \epsilon_i$ where $\epsilon_i \sim N(0, 1)$

(C) $X_i^2 + Y_i^2 = 10$

(D) $X_i^2 + Y_i^2 <= 10$

(E) None of these

Consider we have data in R^2. Then the linear regression line and the PCA line

(A) will always be the same

(B) will never be the same

(C) **[Ans]** can sometimes be the same

(D) None of these

We want to do PCA using gradient descent. Then the update rule is Assume that Σ is the covariance matrix, η is the learning rate.

(A) $u_{k+1} = \eta \Sigma u_k$

(B) **[Ans]** $u_{k+1} = (I + \eta \Sigma) u_k$

(C) $u_{k+1} = (I - \eta \Sigma) u_k$

(D) None of these

PCA solves this problem:

$$\max_{u} u^T \Sigma u - \lambda(u^T u - 1)$$

where Σ is the covariance matrix. Which of the following are true regarding PCA

(A) **[Ans]** λ is the variance captured by the eigen vector u

(B) **[Ans]** Sum of variances captured by all eigenvectors is $\text{tr}(\Sigma)$

(C) If all data points are on a line then at least one of the eigenvalues is 1

(D) **[Ans]** If all data points are on a line then at least one of the eigenvalues is 0

Let $X = UDV^T$. Then

(A) Columns of U are eigenvectors of X^TX

(B) **[Ans]** Columns of V a re eigenvectors of X^TX

(C) Rows of U are eigenvectors of X^TX

(D) Rows of V are eigenvectors of X^TX

(E) None of these

Let $X = UDV^T$. Then

(A) **[Ans]** Columns of U are eigenvectors of XX^T

(B) Columns of V are eigenvectors of XX^T

(C) Rows of U are eigenvectors of XX^T

(D) Rows of V are eigenvectors of XX^T

(E) None of these

Consider X to be a square matrix of size $n \times n$ and $X = UDV^T$.

(A) **[Ans]** Both X^TX and XX^T have the same eigenvalues

(B) Both X^TX and XX^T have the same eigenvectors

(C) X, XX^TX and XX^T have the same eigenvalues

(D) **[Ans]** D^2 contains the eigenvalues of X^TX on its diagonal

(E) D contains the eigenvalues of X^TX on its diagonal

(F) None of these

Consider X to be a square matrix of size $n \times n$ and $X = UDV^T$. Then:

(A) **[Ans]** If rank(X) $= n$, D has all non-zero entries in diagonal.

(B) If rank(X) $= k$, D has k zeros in diagonal

(C) **[Ans]** If rank(X) $= k$, D has $n - k$ zeros in diagonal

(D) **[Ans]** if rank(X) $= n$ but $|A|$ is a very small number then, D takes the form $D = diag(d_1, d_2, .., \epsilon)$ where ϵ is a very small number

(E) None of these

Suppose you want to apply PCA to your data X which is in 2D and you decompose X as UDV^T. Then,

(A) PCA can be useful if all elements of D are equal

(B) **[Ans]** PCA can be useful if all elements of D are not equal

(C) **[Ans]** D is not full-rank if all points in X lie on a straight line

(D) V is not full-rank if all points in X lie on a straight line

(E) D is not full-rank if all points in X lie on a circle

(F) None of these

Given a set of 2D points X on a line that makes 45 degree to the x-axis:

$$X = \{[1,1]^T, [2,2]^T, [3,3]^, [4,4]^T, [5,5]^T\}$$

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) **[Ans]** $\lambda_2 = 0$

(B) $\lambda_1 = \lambda_2$

(C) $\lambda_1 = -1$

(D) **[Ans]** Σ is singular

(E) none of the above

Given a set of 2D points X on a line that makes 45 degree to the x-axis:

$$X = \{[-2, 2]^T, [-3, 3]^,[-4, 4]^T, [-5, 5]^T[-6, 6]^T\}$$

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) **[Ans]** $\lambda_2 = 0$

(B) $\lambda_1 = \lambda_2$

(C) $\lambda_1 = -1$

(D) **[Ans]** Σ is singular

(E) none of the above

Given a set of 2D points X on the vertical line $x_1 = 5$,

$$X = \{[5,1]^T, [5,2]^T, [5,3]^T, [5,4]^T, [5,5]^T\}$$

We now add an additional point $[4,3]^T$ to X.

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) **[Ans]** $\lambda_1 \geq \lambda_2$

(B) $\mathbf{u}_1$ and $\mathbf{u}_2$ are nearly orthogonal, but not perfectly orthogonal.

(C) Σ is singular

(D) **[Ans]** Σ is diagonal

(E) None of the above.

Given a set of 2D points X on the vertical line $x_2 = 5$,

$$X = \{[1,5]^T, [2,5]^T, [3,5]^T, [4,5]^T, [5,5]^T\}$$

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) **[Ans]** $\lambda_1 \geq \lambda_2$

(B) **[Ans]** μ is on the same line.

(C) **[Ans]** Σ is singular

(D) **[Ans]** Σ is diagonal

(E) None of the above.

Set X has 10 points. 5 of them are on a line that makes 45 degrees with the x_1 axis and another 5 from on a line that makes 135 degrees with the x_1 axis.

We compute the covariance matrix, and its eigen values and eigen vectors. Then:

(A) $\lambda_1 = \lambda_2 \neq 0$

(B) Σ is singular

(C) Σ is diagonal

(D) μ is on either of these lines.

(E) **[Ans]** None of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathcal{L}_1 = \mathcal{L}_2$ for all $\mathbf{w}$, then

(A) **[Ans]** A is a diagonal matrix

(B) $A_{ij} = \alpha_i \cdot \alpha_j$

(C) **[Ans]** $A_{ii} = \alpha_i$ else zero

(D) $A_{ii} = \frac{1}{\alpha_i}$ else zero

(E) none of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 1$ for all i, and $\lambda_1 = \lambda_2 = 1$, then

(A) **[Ans]** Both the loss functions are identical i.e., $\mathcal{L}_1 = \mathcal{L}_2$

(B) **[Ans]** The optima of the first objective $\mathbf{w}_1^*$ is same as the optima of $\mathcal{L}_2$, i.e., $\mathbf{w}_2^*$

(C) **[Ans]** At the optima, value of the losses are same. i.e., $\mathcal{L}_1^* = \mathcal{L}_2^*$

(D) $\mathcal{L}_1$ is a scalar and $\mathcal{L}_2$ is a vector

(E) none of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 2$ for all i, and $\lambda_1 = \lambda_2 = 0$, then

(A) Both the loss functions are identical i.e., $\mathcal{L}_1 = \mathcal{L}_2$

(B) **[Ans]** The optima of the first objective $\mathbf{w}_1^*$ is same as the optima of $\mathcal{L}_2$, i.e., $\mathbf{w}_2^*$

(C) At the optima, value of the losses are same. $\mathcal{L}_1^* = \mathcal{L}_2^*$

(D) $\mathcal{L}_1$ is a scalar and $\mathcal{L}_2$ is a vector

(E) none of the above

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A[Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

If $\mathbf{A} = I$, $\alpha_i = 1$ for all i, and $\lambda_1 \neq \lambda_2 \neq 0$, then

(A) The optimal parameters $\mathbf{w}^*$ is independent of λ_i.

(B) The larger the lambda, the better the solution.

(C) The smaller the lambda, the better the

(D) When lambda is nonzero (positive), loss will increase (since $g(w)$ is also positive in practice), better to use $\lambda = 0$.

(E) **[Ans]** None of the above.

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

See $\mathcal{L}_2$ closely,

(A) **[Ans]** When A is a diagonal matrix, this is equivalent to weighing each sample independently.

(B) When A is not a diagonal matrix, this loss does not make any sense. Don't use.

(C) **[Ans]** When A is PD, we can do cholesky decomposition of A as LL^T and an equivalent formulation is possible in $\mathcal{L}_1$ is each sample getting transformed as $\mathbf{L}^T\mathbf{x}_i$ (as in LMNN/Metric Learning)

(D) **[Ans]** When A is a rank deficient matrix, an equivalent formulatiion is possible in $\mathcal{L}_1$ with a dimensionality reduction (this could be proved with eigen decomposition).

(E) None of the above

Consider a vocabulary of size d. One hot representation of a word i is "1" at the location (index) corresponding to that word and zero else where.

Given a document that contains P words, $\mathbf{w_1}, \ldots, \mathbf{w_P}$, we compute

$$\mathbf{x} = \sum_{i=1}^{P} \mathbf{w}_i$$

Then,

(A) **[Ans]** $\mathbf{x}$ is the histogram of the words, with x_i as the frequency of i th word.

(B) **[Ans]** $\mathbf{x}$ is in R^d independent of the number of words in the document.

(C) $\mathbf{x}$ is in R^P independent of the vocabulary size.

(D) **[Ans]** $\sum_i x_i$ is P (x_i is the i th element of $\mathbf{x}$)

Consider a document is represented by a histogram of the words in the document. **h** i.e., h_i is the number of occurrence of the i th word in the document.

We define a linguistic operation: Paraphrasing (P1). P1 is defined as permuting sentences in a document and rewriting a sentence by permuting the words.

(A) **[Ans] h** is invariant to the P1

(B) **h** is not invariant to the P1

(C) **h** is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

(D) **[Ans]** a Euclidean distance computed over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a".

Consider a document is represented by a histogram of the words in the document **h** i.e., h_i is the number of occurrence of the i th word in the document.

We define a linguistic operation: Paraphrasing (P2). P2 is defined as replacing a set of words by their synonyms.

(A) **h** is invariant to the P2

(B) **[Ans] h** is not invariant to the P2

(C) **h** is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

(D) a Euclidean distance computed over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under in which order the vocabulary is constructed (eg. "a to z" or "z to a"

A professor suspected that students while submiting home works are doing the paraphasing operations i.e., both P1 and P2. This resulted in failure of some similarity tests.

Professor designs a $d \times d$ word similarity matrix $\mathbf{S}$ such that $\mathbf{S}_{ij} = \mathbf{S}_{ji} = 1$ if words i and j are synonyms and zero else. (Note: d is the size of vocabulary).

Now to compare two documents, professor multiplies the histogram representations by $\mathbf{S}$.

$$\mathbf{h}'_i = \mathbf{S}\mathbf{h}_i$$

(Note: $\mathbf{h}'_i$ is the new representation. Also, note, after multiplying with the $\mathbf{S}$, the dimension does not change)

(A) **[Ans]** the new representation is invariant under the operation$P1$ and $P2$. (i.e., All the plagiarism now will be detected.)

(B) the new representation is not invariant for $P2$ and it does not help.

(C) the new representation helps for detecting people who have paraphrased with P2. But now it

(D) the idea is worth, but then $\mathbf{S}$ should not have made symmetric. with only one of $\mathbf{S}_{ij}$ or $\mathbf{S}_{ji}$ as 1. The method could have worked as expected.

We want to compare two documents i and j which are represented as histogram (popular known as bag of words) of words h_i and h_j.

Here is what four students argued:

(A) **[Ans]** histograms should be normalized by dividing by the number of words in the document so that the comparison operation becomes "some what invariant" to another linguistic operation: "summarization".

(B) **[Ans]** Cosine distance is a popular distance to compare two documents using this representation.

(C) **[Ans]** we should remove the stop words (common words in the language) from the sentence so that the comparison will be more useful. Two documents have the same number of 'the' does not mean any useful similarity between them.

If $A = UDV^T$, then A^TA is

(A) **[Ans]** VD^2V^T

(B) UD^2U^T

(C) **[Ans]** A square matrix

(D) is always full rank

(E) none of the above

Consider a matrix A of size $m \times n$. Rank of A is (choose one one most correct answer)

(A) **[Ans]** $\leq \min(m, n)$

(B) $\leq \max(m, n)$

(C) $\geq \min(m, n)$

(D) $\geq \max(m, n)$

(E) $\frac{m+n}{2}$

A and B are two independent events such that $P(\overline{A}) = 0.4$ and $P(A \cap B) = 0.2$ Then $P(A \cap \overline{B})$ is equal to

(A) **[Ans]** 0.4

(B) 0.2

(C) 0.6

(D) 0.8

(E) None of the above

If $\mathbf{A}$ is a $n \times n$ matrix, with every pair of columns orthogonal i.e., $\mathbf{a_i} \cdot \mathbf{a_j} = \mathbf{0} \ \forall i, j$ and $||\mathbf{a_i}|| = 1$. Then:

(A) **[Ans]** $\mathbf{A^{-1}} = \mathbf{A}^T$.

(B) **[Ans]** $\mathbf{AA}^T = \mathbf{I}$

(C) $\mathbf{AA}^T$ has only one 1 in every column and all others zero.

(D) $\mathbf{A}^{-1}$ has only one 1 in every column and all others zero.

(E) none of the above

Product of Eigen values of a real square matrix is:

(A) **[Ans]** Determinant

(B) Rank

(C) Trace

(D) non-Negative

(E) None of the above

$X \sim N(0,1)$, $Y \sim N(1,1)$ and $Z = X + Y$. Then,

(A) $Z \sim N(0,2)$

(B) $Z \sim N(0,1)$

(C) $Z \sim N(1,1)$

(D) **[Ans]** $Z \sim N(1,2)$

(E) None of the above