Quiz 1

Name: **Roll Number**.....

- Answer precisely in the space given. No overwriting. Make assumptions, if really ambiguous.
- Questions 1-10: 1 Points; Questions 11-12: 2.5 Points. Total 15 Points. ∘

---

1. Consider a simple linear classifier, with bias. $\mathbf{w} = [2.5, 0.5, 1.5]^T$. This is tested on the following 5 samples. $(\mathbf{x}_i, y_i)$

$$([0,0]^T, +), ([+1,+1]^T, +), ([-1,+1]^T, +), ([-1,-1]^T, -), ([+1,-1]^T, -)$$

What is the accuracy of this classifier ? Ans: ———-%.

Solution:

We have a simple linear classifier characterized by the weight vector

$$w = \begin{bmatrix} 2.5 \\ 0.5 \\ 1.5 \end{bmatrix}.$$

Here, the first two components correspond to the coefficients for x_1 and x_2, and the third component (1.5) acts as the bias term.

Hence, for a 2D input point $\mathbf{x} = (x_1, x_2)$, the classifier predicts

$$\hat{y} = \text{sign}(2.5\,x_1 \;+\; 0.5\,x_2 \;+\; 1.5).$$

We test this classifier on the following 5 labeled samples $(\mathbf{x}_i, y_i)$:

$$\begin{aligned} (0,0), &\quad y = +1, \\ (1,1), &\quad y = +1, \\ (-1,+1), &\quad y = +1, \\ (-1,-1), &\quad y = -1, \\ (1,-1), &\quad y = -1. \end{aligned}$$

Step-by-step classification:

(a) $\mathbf{x}_1 = (0,0), \quad y_1 = +1.$

$$2.5 \cdot 0 \;+\; 0.5 \cdot 0 \;+\; 1.5 = 1.5 \quad \Longrightarrow \quad \text{sign}(1.5) = +1.$$

This matches the true label +1. *Correct.*

(b) $\mathbf{x}_2 = (1,1), \quad y_2 = +1.$

$$2.5 \cdot 1 \;+\; 0.5 \cdot 1 \;+\; 1.5 = 2.5 + 0.5 + 1.5 = 4.5 \quad \Longrightarrow \quad \text{sign}(4.5) = +1.$$

This matches the true label +1. *Correct.*

(c) $\mathbf{x}_3 = (-1,+1), \quad y_3 = +1.$

$$2.5 \cdot (-1) \;+\; 0.5 \cdot (+1) \;+\; 1.5 = -2.5 + 0.5 + 1.5 = -0.5 \quad \Longrightarrow \quad \text{sign}(-0.5) = -1.$$

This does *not* match the true label +1. *Incorrect.*

(d) $\mathbf{x}_4 = (-1, -1), \quad y_4 = -1.$

$$2.5 \cdot (-1) \; + \; 0.5 \cdot (-1) \; + \; 1.5 = -2.5 - 0.5 + 1.5 = -1.5 \quad \Longrightarrow \quad \text{sign}(-1.5) = -1.$$

This matches the true label -1. *Correct.*

(e) $\mathbf{x}_5 = (1, -1), \quad y_5 = -1.$

$$2.5 \cdot 1 \; + \; 0.5 \cdot (-1) \; + \; 1.5 = 2.5 - 0.5 + 1.5 = 3.5 \quad \Longrightarrow \quad \text{sign}(3.5) = +1.$$

This does *not* match the true label -1. *Incorrect.*

Accuracy Calculation:

Out of the 5 samples, the classifier is correct on 3 of them (Samples 1, 2, and 4) and incorrect on 2 (Samples 3 and 5). Thus, the accuracy is

$$\frac{\text{Number of correct predictions}}{\text{Total samples}} \; = \; \frac{3}{5} \; = \; 0.6 \; = \; 60\%.$$

The accuracy of this classifier on the given samples is 60%.

---

2. A 3NN classifier is built over four training examples, and tested on the same test set of Q1.

$$([+2.0, 0.5]^T, +), (-3.0, +0.5]^T, -), ([0.0, 3.0]^T, +), ([0.0, -3.0]^T, -)$$

What is the accuracy of this classifier ? Ans: ———-%.

Solution:

Training Data

$$\textbf{TP1} : \left([2.0, 0.5]^T, +\right) \quad \textbf{TP2} : \left([-3.0, 0.5]^T, -\right)$$
$$\textbf{TP3} : \left([0.0, 3.0]^T, +\right) \quad \textbf{TP4} : \left([0.0, -3.0]^T, -\right)$$

Test Data (from Q1)

$$\textbf{TS1} : \left([0, 0]^T, +\right) \quad \textbf{TS2} : \left([1, 1]^T, +\right) \quad \textbf{TS3} : \left([-1, 1]^T, +\right)$$
$$\textbf{TS4} : \left([-1, -1]^T, -\right) \quad \textbf{TS5} : \left([1, -1]^T, -\right)$$

Calculations for Each Test Sample

(a) **Test Sample 1 (TS1:** $[0, 0]^T, +$**)**

Distance to TP1: $(2.0 - 0)^2 + (0.5 - 0)^2 = 4.25$

Distance to TP2: $(-3.0 - 0)^2 + (0.5 - 0)^2 = 9.25$

Distance to TP3: $(0.0 - 0)^2 + (3.0 - 0)^2 = 9$

Distance to TP4: $(0.0 - 0)^2 + (-3.0 - 0)^2 = 9$

3 Nearest Neighbors: TP1 (+), TP3 (+), TP4 (−)
Majority Vote: + (Correct).

(b) **Test Sample 2 (TS2: $[1, 1]^T$, +)**

$$\text{Distance to TP1: } (2.0 - 1)^2 + (0.5 - 1)^2 = 1.25$$
$$\text{Distance to TP2: } (-3.0 - 1)^2 + (0.5 - 1)^2 = 16.25$$
$$\text{Distance to TP3: } (0.0 - 1)^2 + (3.0 - 1)^2 = 5$$
$$\text{Distance to TP4: } (0.0 - 1)^2 + (-3.0 - 1)^2 = 17$$

3 Nearest Neighbors: TP1 (+), TP3 (+), TP2 (−)
Majority Vote: + (Correct).

(c) **Test Sample 3 (TS3: $[-1, 1]^T$, +)**

$$\text{Distance to TP1: } (2.0 + 1)^2 + (0.5 - 1)^2 = 9.25$$
$$\text{Distance to TP2: } (-3.0 + 1)^2 + (0.5 - 1)^2 = 4.25$$
$$\text{Distance to TP3: } (0.0 + 1)^2 + (3.0 - 1)^2 = 5$$
$$\text{Distance to TP4: } (0.0 + 1)^2 + (-3.0 - 1)^2 = 17$$

3 Nearest Neighbors: TP2 (−), TP3 (+), TP1 (+)
Majority Vote: + (Correct).

(d) **Test Sample 4 (TS4: $[-1, -1]^T$, −)**

$$\text{Distance to TP1: } (2.0 + 1)^2 + (0.5 + 1)^2 = 11.25$$
$$\text{Distance to TP2: } (-3.0 + 1)^2 + (0.5 + 1)^2 = 6.25$$
$$\text{Distance to TP3: } (0.0 + 1)^2 + (3.0 + 1)^2 = 17$$
$$\text{Distance to TP4: } (0.0 + 1)^2 + (-3.0 + 1)^2 = 5$$

3 Nearest Neighbors: TP4 (−), TP2 (−), TP1 (+)
Majority Vote: − (Correct).

(e) **Test Sample 5 (TS5: $[1, -1]^T$, −)**

$$\text{Distance to TP1: } (2.0 - 1)^2 + (0.5 + 1)^2 = 3.25$$
$$\text{Distance to TP2: } (-3.0 - 1)^2 + (0.5 + 1)^2 = 18.25$$
$$\text{Distance to TP3: } (0.0 - 1)^2 + (3.0 + 1)^2 = 17$$
$$\text{Distance to TP4: } (0.0 - 1)^2 + (-3.0 + 1)^2 = 5$$

3 Nearest Neighbors: TP1 (+), TP4 (−), TP3 (+)
Majority Vote: + (Incorrect).

Accuracy Calculation

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Test Samples}} \times 100 = \frac{4}{5} \times 100 = \boxed{80\%}$$

3. Posterior probability of two classes (i.e., $P(\omega_1|\mathbf{x})$ and $P(\omega_2|\mathbf{x})$) for two classes in a binary classification problem (with $\mathbf{x} \in R^1$)are known to be normal distributions $\mathcal{N}(20, 25)$ and $\mathcal{N}(30, 5)$ respectively. Given a test sample of 25.1, predict the class. Ans: ————-

Solution:

Given the posterior probabilities for two classes ω_1 and ω_2 modeled as normal distributions:

$$P(x|\omega_1) \sim \mathcal{N}(20, 25)$$
$$P(x|\omega_2) \sim \mathcal{N}(30, 5)$$

For the test sample $x = 25.1$, we compute the likelihoods:

1. Likelihood for Class ω_1:

$$P(x|\omega_1) = \frac{1}{5\sqrt{2\pi}} e^{-\frac{(25.1-20)^2}{2 \cdot 25}} = \frac{1}{5\sqrt{2\pi}} e^{-\frac{26.01}{50}} \approx 0.0474$$

2. Likelihood for Class ω_2:

$$P(x|\omega_2) = \frac{1}{\sqrt{5}\sqrt{2\pi}} e^{-\frac{(25.1-30)^2}{2 \cdot 5}} = \frac{1}{\sqrt{5}\sqrt{2\pi}} e^{-\frac{24.01}{10}} \approx 0.0162$$

Decision:

Since $P(x|\omega_1) > P(x|\omega_2)$, the test sample $x = 25.1$ is classified as $\boxed{\omega_1}$.

Intuitive Explanation:

Imagine two bell curves on a number line:

- **Class** ω_1 is centered at 20 and is wide (large spread due to higher varience), meaning it accepts a broader range of values around 20.
- **Class** ω_2 is centered at 30 but is narrow (small spread due to lower varience), meaning it strongly prefers values close to 30.

The test sample 25.1 is closer to 30 (distance = 4.9) than to 20 (distance = 5.1). However:

- ω_2's narrow spread penalizes values far from 30, even if they're slightly closer.
- ω_1's wide spread is more forgiving, making 25.1 relatively likely despite being farther from 20.

Result:
Even though 25.1 is slightly closer to ω_2's mean, ω_1's flexibility makes it the better fit.

Answer: ω_1

	Predicted +ve	Predicted -ve
Actual +ve	A	B
Actual -ve	C	D

4. Given the following confusion matrix.

   Write an expression for precision. Ans ————————

   Write an expression for accuracy. Ans ————————

   Expressions for:

   Precision:

$$\text{Precision} = \frac{A}{A + C}$$

   Recall:

$$\text{Recall} = \frac{A}{A + B}$$

   Accuracy:

$$\text{Accuracy} = \frac{A + D}{A + B + C + D}$$

---

5. An $N \times N$ matrix A is composed of consecutive N^2 integers starting from K. (i.e., $K, K + 1, \ldots, K + (N^2 - 1)$)

   Rank of A is independent of N. (True or False?) ———

   Rank of A is independent of K. (True or False?) ———

   Matrix Form. A natural way to arrange these consecutive numbers in A is row by row. Hence

$$A_{ij} \;=\; K \;+\; (i-1)\,N \;+\; (\,j-1\,), \quad \text{for } 1 \le i, j \le N.$$

   Thus the first row is $(K,\ K+1,\ K+2, \ldots)$, the second row continues $(K+N,\ K+N+1, \ldots)$, etc.

   1) Independence from K.

   Consider what happens if we replace K by any other value K'. One can see that *adding the same constant* to every entry of a matrix does not change its rank, because rank depends only on linear dependence/independence of rows (or columns).

   Concretely, you can perform elementary row (or column) operations such as subtracting one row from another; any constant K cancels out. Hence the rank is unaffected by shifting all entries by K.

   Therefore, *the rank of A is independent of K.*

   2) Dependence on N.

   Case $N = 1$. Then A is the 1×1 matrix $[\,K\,]$. Its rank is 1 if $K \neq 0$, and 0 if $K = 0$.

Case $N \geq 2$. A simple row-difference argument shows that the rank is always 2 for all $N \geq 2$. Indeed, if you subtract the first row from the second, the second row from the third, etc., all those differences become constant rows (and quickly collapse to at most 2 nonzero rows). One can check that at least two rows remain linearly independent, but no more than two. Hence for every $N \geq 2$, the rank is 2, regardless of K.

Putting these observations together:

$$\operatorname{rank}(A) = \begin{cases} 1 & \text{if } N = 1 \text{ and } K \neq 0, \\ 0 & \text{if } N = 1 \text{ and } K = 0, \\ 2 & \text{if } N \geq 2. \end{cases}$$

Thus, as N changes from 1 to 2, the rank jumps from (possibly) 1 to 2. For $N \geq 2$, it remains constantly 2. In other words, the rank is *not the same for all* N (it differs between $N = 1$ and $N \geq 2$).

Hence, *the rank of* A *is* ***not*** *independent of* N.

---

6. Bag I contain 10 white and 5 black balls. Bag II contains 15 white and 5 black balls.

   A ball is drawn at random from one of the bags, and it is found to be white. What is the probability that it was drawn from Bag I.

   Ans ————-

   Solution:

   We have two bags:

$$\begin{aligned} &\text{Bag I: 10 white, 5 black} \quad (\text{total } 15), \\ &\text{Bag II: 15 white, 5 black} \quad (\text{total } 20). \end{aligned}$$

   Assume each bag is chosen with probability $\frac{1}{2}$ (since the problem says "a ball is drawn at random from one of the bags" without further specification).

   Let W = event that the drawn ball is white, I = event that Bag I is chosen.

   We want $P(I \mid W)$. By Bayes' Theorem:

$$P(I \mid W) = \frac{P(W \mid I)\, P(I)}{P(W)}.$$

   Step 1: Compute $P(W \mid I)$:

$$P(W \mid I) = \frac{\text{number of white balls in Bag I}}{\text{total in Bag I}} = \frac{10}{15} = \frac{2}{3}.$$

Step 2: Compute $P(W \mid \text{II})$:

$$P(W \mid \text{II}) = \frac{15}{20} = \frac{3}{4}.$$

Step 3: Since each bag is equally likely,

$$P(I) = \frac{1}{2}, \quad P(\text{II}) = \frac{1}{2}.$$

Step 4: Compute $P(W)$ using the Law of Total Probability:

$$P(W) = P(W \mid I)\,P(I) \;+\; P(W \mid \text{II})\,P(\text{II}) = \left(\frac{2}{3}\right)\left(\frac{1}{2}\right) + \left(\frac{3}{4}\right)\left(\frac{1}{2}\right) = \frac{1}{3} + \frac{3}{8} = \frac{8}{24} + \frac{9}{24} = \frac{17}{24}.$$

Step 5: Apply Bayes' Theorem:

$$P(I \mid W) = \frac{P(W \mid I)\,P(I)}{P(W)} = \frac{\frac{2}{3} \cdot \frac{1}{2}}{\frac{17}{24}} = \frac{\frac{1}{3}}{\frac{17}{24}} = \frac{1}{3} \times \frac{24}{17} = \frac{8}{17}.$$

The probability that the white ball was drawn from Bag I is $\frac{8}{17}$.

---

7. A set of samples were pre-processed by a simple linear transformation $\mathbf{x}' = \mathbf{A}\mathbf{x}$ Let d_{ij} is the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ and d'_{ij} is the distance between $\mathbf{x}'_i$ and $\mathbf{x}'_j$

   (a) When $\mathbf{A}$ is a permutation matrix (i.e., every row and column has only one '1' and all other elements being '0'; Note: $\mathbf{A}$ need not be identity). Then, $d_{ij} = d'_{ij}$ for all i, j. True or False? ———-

   (b) When $\mathbf{A}$ is $\rho\mathbf{I}$, a simple linear classifier $sign(\mathbf{w}^T\mathbf{x})$ will not report any change in accuracy after the transformation. True or False? ———

Solution:

Let $\{x_i\}_{i=1}^n$ be a set of samples. We apply a linear transformation

$$x'_i \;=\; A\,x_i,$$

where A is a matrix. Denote the Euclidean distance between x_i and x_j in the original space by

$$d_{ij} \;=\; \|x_i - x_j\|,$$

and in the transformed space by

$$d'_{ij} \;=\; \|x'_i - x'_j\| \;=\; \|A\,x_i - A\,x_j\|.$$

(a) A is a permutation matrix

A permutation matrix A has exactly one entry 1 in each row and each column, with all other entries 0. Such matrices are orthonormal, satisfying

$$A^\top A = I.$$

Hence,

$$\|A(x_i - x_j)\|^2 = (x_i - x_j)^\top A^\top A(x_i - x_j) = (x_i - x_j)^\top (x_i - x_j) = \|x_i - x_j\|^2.$$

Taking square roots, we get

$$\|A\,x_i - A\,x_j\| = \|x_i - x_j\| \quad \Longrightarrow \quad d'_{ij} = d_{ij}.$$

Answer for (a): This statement is *True* because permutation matrices preserve Euclidean distances.

(b) $A = p\,I$

Suppose A is a scalar multiple of the identity, i.e. $A = pI$ for some scalar p. Then

$$x'_i = p\,I\,x_i = p\,x_i.$$

Consider a linear classifier of the form $\text{sign}(w^\top x)$. Under the transformation,

$$w^\top x' = w^\top (p\,x) = p\,(w^\top x).$$

- If $p > 0$, then
$$\text{sign}(w^\top x') = \text{sign}(p\,w^\top x) = \text{sign}(w^\top x).$$
Thus the predicted labels remain the same, and the accuracy does not change.
- If $p < 0$, then
$$\text{sign}(w^\top x') = \text{sign}(p\,w^\top x) = \text{sign}(-(|p|)\,w^\top x) = -\,\text{sign}(w^\top x),$$
which flips the sign.

If $p > 0$, multiplying by p does not change the sign. Hence the predicted label $\text{sign}(p\,(\mathbf{w}^\top \mathbf{x}))$ coincides with $\text{sign}(\mathbf{w}^\top \mathbf{x})$. Thus, *no classification decisions change*, and the accuracy is unaffected. But if $p < 0$, multiplying by p does change the sign. Hence the predicted label $\text{sign}(p\,(\mathbf{w}^\top \mathbf{x}))$ will change

Answer: False (because change in accuracy depend on p sign if it's positive then there is no change in accuracy and it it's negative then there will be change in accuracy).

(a) is True and (b) is False

8. Consider a vocabulary of size d. One hot representation of a word i, $\mathbf{w}_i$, 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) $\mathbf{x}$ is the histogram of the words, with its ith element x_i as the frequency of i th word.

(b) $\mathbf{x}$ 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) $\sum_i x_i$ is P (x_i is the i th element of $\mathbf{x}$)

Which of the above statements are True? ————————

Statement Analysis:

(a) "x is the histogram of the words, with its i-th element x_i as the frequency of the i-th word."

When we sum one-hot vectors, each position i in x accumulates the number of times the i-th word appeared in the document. This is the definition of a histogram. Therefore, this statement is **TRUE**.

(b) "x is in $\mathbb{R}^d$ independent of the number of words in the document."

Each w_i is a d-dimensional vector. The sum x is also a d-dimensional vector. The number of words P affects the magnitude of the elements in x, but not the dimensionality. Therefore, this statement is **TRUE**.

(c) "x is in $\mathbb{R}^P$ independent of the vocabulary size."

As explained in (b), x is in $\mathbb{R}^d$, where d is the vocabulary size. The dimension of x depends on the vocabulary size, not the number of words P. Therefore, this statement is **FALSE**.

(d) "$\sum_i x_i$ is P (x_i is the i-th element of x)."

Each word contributes exactly one '1' to its one-hot representation. When we sum P words, the total sum of the elements in x will be equal to the number of words, P. Therefore, this statement is **TRUE**.

Consider a simple example:

- Vocabulary size $d = 4$ (words: "cat", "dog", "bird", "fish")
- Document with $P = 3$ words: ["cat", "dog", "cat"]

One-hot representations would be:

$$\begin{aligned} \text{"cat"} &= [1, 0, 0, 0] \\ \text{"dog"} &= [0, 1, 0, 0] \\ \text{"bird"} &= [0, 0, 1, 0] \\ \text{"fish"} &= [0, 0, 0, 1] \end{aligned}$$

2. Computing x

For our example document ["cat", "dog", "cat"]:

$$\mathbf{x} = [1, 0, 0, 0] + [0, 1, 0, 0] + [1, 0, 0, 0]$$
$$\mathbf{x} = [2, 1, 0, 0]$$

3. Analysis of Each Statement

(a) x is the histogram of the words, with its ith element x_i as the frequency of ith word.

- In our example, $\mathbf{x} = [2, 1, 0, 0]$ means:
  - $x_1 = 2$ (word "cat" appears twice)
  - $x_2 = 1$ (word "dog" appears once)
  - $x_3 = 0$ (word "bird" appears zero times)
  - $x_4 = 0$ (word "fish" appears zero times)
- This exactly matches the definition of a histogram counting frequencies
- Therefore, **TRUE**

(b) x is in $\mathbb{R}^d$ independent of the number of words in the document.

- Each one-hot vector $\mathbf{w}_i$ has dimension d (vocabulary size)
- When we sum vectors, the result has the same dimension as the vectors
- For example:
  - If $P = 3$ words: $\mathbf{x} \in \mathbb{R}^4$
  - If $P = 100$ words: $\mathbf{x} \in \mathbb{R}^4$
- Dimension remains d regardless of P
- Therefore, **TRUE**

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

- This claims $\mathbf{x}$ has dimension P (number of words)
- But we proved in (b) that $\mathbf{x}$ has dimension d
- For example:
  - If vocabulary size $d = 1000$, $\mathbf{x} \in \mathbb{R}^{1000}$
  - The number of words P doesn't affect this
- Therefore, **FALSE**

(d) $\sum_i x_i$ is P (x_i is the ith element of x)

- Each word vector $\mathbf{w}_i$ has exactly one '1' and rest zeros
- When we sum P such vectors:
  - Each vector contributes exactly one '1'
  - Total sum must be P

- In our example:
  - $\mathbf{x} = [2, 1, 0, 0]$
  - $\sum_i x_i = 2 + 1 + 0 + 0 = 3 = P$
- Therefore, **TRUE**

(a), (b) and (d) are True

Problem 8 of other set:

Consider a document is represented by a histogram of the words in the document $\mathbf{h}$ 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 respective synonym.

(a) $\mathbf{h}$ is invariant to the P2

(b) $\mathbf{h}$ is not invariant to the P2

(c) $\mathbf{h}$ 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")

Which of the above statements are True? ————————

Solution:

(a) "h is invariant to P2"

False. If we replace some words in the document by synonyms (i.e. different tokens in the vocabulary), the counts for those original words go down and the counts for the new synonym words go up. Hence the histogram entries $\mathbf{h}$ change. Therefore $\mathbf{h}$ *does* change under paraphrasing.

(b) "h is *not* invariant to P2"

True. As argued above, paraphrasing (replacing words by different tokens) generally alters the histogram counts, so $\mathbf{h}$ is *not* invariant.

(c) Invariance of h under different vocabulary order

False The histogram $\mathbf{h}$ is a mapping from "word" → "count." If vocabulary construction start from 'a' to 'z' then the word starts from 'a' comes earlier and words start from 'z' comes later. Now if you change the vocabulary construction from 'z' to 'a' then the whole ordering gets reversed and the histogram gets flipped so now it is not invariant.

(d) "The Euclidean distance over $\mathbf{h}_i$ and $\mathbf{h}_j$ is invariant under the order in which the vocabulary is constructed"

False. After performing operation P2. Before and after the histogram will change completely because it replaces the words with its synonym. so euclidean distance between $\mathbf{h}_i$ and $\mathbf{h}_j$ will change so it's not invariant.

(b) are True.

Problem 8 from other set:

Consider a document is represented by a histogram of the words in the document. $\mathbf{h}$ 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) $\mathbf{h}$ is invariant to the P1

(b) $\mathbf{h}$ is not invariant to the P1

(c) $\mathbf{h}$ 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".)

Which of the above statements are True? ————————

Solution

(a) and (b): Invariance of h under Paraphrasing P_1

The histogram h captures how many times each word appears in the entire document. When we apply the operation P_1, we are merely reordering the words and sentences; the total count of each word does not change. Therefore, h remains exactly the same under P_1.

(c) Invariance of h under different vocabulary order

False The histogram $\mathbf{h}$ is a mapping from "word" → "count." If vocabulary construction start from 'a' to 'z' then the word starts from 'a' comes earlier and words start from 'z' comes later. Now if you change the vocabulary construction from 'z' to 'a' then the whole ordering gets reversed and the histogram gets flipped so now it is not invariant.

(d) Invariance of Euclidean distance under vocabulary reordering

h is invariant to the P1 operation. so performing the P1 operation will not change the histogram but changing the vocabulary construction form a-z to z-a will flip both histograms. but the Euclidean distance will remain the same.

Original Histograms: Let

$$h = [a, b, c, d] \quad \text{and} \quad g = [e, f, g, h].$$

The Euclidean distance between them is given by:

$$\|h - g\|_2 = \sqrt{(a-e)^2 + (b-f)^2 + (c-g)^2 + (d-h)^2}.$$

Reordered Histograms: If we reorder both histograms by reversing the order of the entries, we have:

$$h' = [d, c, b, a] \quad \text{and} \quad g' = [h, g, f, e].$$

Then the Euclidean distance becomes:

$$\|h' - g'\|_2 = \sqrt{(d-h)^2 + (c-g)^2 + (b-f)^2 + (a-e)^2}.$$

Since addition is commutative, the two distances are equal:

$$\|h - g\|_2 = \|h' - g'\|_2.$$

Hence, (d) is true.

(a) and (d) are True.

---

9. Consider the covariance matrix Σ

   (a) Σ is symmetric
   (b) Σ is PSD
   (c) Σ is Diagonal if the distribution is Normal.
   (d) Σ can not be Diagonal if the distribution is Normal.
   (e) None of the above are true

   Which of the above statements are True? ————————

   Solution:

- **(a) Σ is symmetric.**
  By definition, the covariance matrix Σ of a random vector $\mathbf{X}$ has entries

$$\Sigma_{ij} \;=\; \mathrm{Cov}(X_i, X_j) \;=\; \mathbb{E}\big[(X_i - \mu_i)(X_j - \mu_j)\big],$$

  where $\mu_i = \mathbb{E}[X_i]$. Since $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$, the matrix Σ is always symmetric. *True.*

- **(b) Σ is PSD (positive semidefinite).**
  A covariance matrix Σ is always positive semidefinite. Formally, for any vector $\mathbf{v} \in \mathbb{R}^n$,

$$\mathbf{v}^\top \Sigma\, \mathbf{v} \;=\; \mathrm{Var}\big(\mathbf{v}^\top \mathbf{X}\big) \;\geq\; 0.$$

  Hence Σ must be PSD. *True.*

- **(c) Σ is diagonal if the distribution is Normal.**
  A Normal (Gaussian) distribution can have *any* valid covariance matrix (symmetric and PSD), not necessarily diagonal. The only time Σ is diagonal for a Normal distribution is if the components are mutually *independent* (and hence uncorrelated). A general multivariate Normal can have nonzero off-diagonal entries representing correlations. Thus statement (c) is *not necessarily true.*

- **(d) Σ cannot be diagonal if the distribution is Normal.**
  This is also false. It *can* be diagonal if the components happen to be independent. So there is no prohibition against a Normal distribution having a diagonal covariance; it simply corresponds to the case of independent components. Thus (d) is *not necessarily true* either.

(a) and (b) are True.

---

10. Consider the following statements:

    (a) Product of Eigen values is Determinant of a matrix

    (b) A matrix of $m \times n$ can have max (m,n) non-zero eigen values

    (c) If determinant of a matrix is zero, means at least one of the eigen value is zero.

    (d) Eigen vectors are orthogonal to each other (i.e., $\mathbf{v_1}^T \mathbf{v_2} = 0$)

    (e) All the above are true

    Which of the above statements are True? ————————

    Solution:

- **(a) "Product of the eigenvalues = determinant"**
  For any *square* matrix $A \in \mathbb{R}^{n\times n}$, if its eigenvalues (counted with algebraic multiplicities) are $\lambda_1, \ldots, \lambda_n$, then
  $$\det(A) \;=\; \lambda_1 \lambda_2 \cdots \lambda_n.$$
  Hence statement (a) is *True* (provided A is square, which is the usual setting for talking about eigenvalues).

- **(b) "A matrix of size $m\times n$ can have at most $\min(m,n)$ non-zero eigenvalues."**
  Eigenvalues are only defined for *square* matrices. If $m \neq n$, an $m \times n$ matrix does not have eigenvalues in the standard sense. It is true that an $m \times n$ matrix can have at most $\min(m,n)$ non-zero *singular* values, or that its rank is at most $\min(m,n)$. But the statement as written refers to "non-zero *eigenvalues*" for an $m \times n$ matrix. That is generally *not* a well-defined statement if $m \neq n$.

- **(c) "If the determinant is zero, at least one eigenvalue is zero."**
  For a square matrix A, $\det(A)$ is the product of its eigenvalues. Hence if $\det(A) = 0$, that product is zero, which implies at least one $\lambda_i = 0$. *True* for square matrices.

- **(d) "Eigenvectors are orthogonal to each other."**
  In general, for an arbitrary matrix $A \in \mathbb{R}^{n\times n}$, there is *no* requirement that distinct eigenvectors be orthogonal. Orthogonality of eigenvectors holds if A is *symmetric* (or Hermitian, or more generally *normal* in the complex case). But a general matrix can have non-orthogonal eigenvectors. Hence statement (d) is *False* in general.

(a) and (c) are True.

---

11. We saw the loss function for linear regression as $J(\theta) = (Y - X\theta)^T(Y - X\theta)$ We saw that we get a closed form solution for θ by solving $\frac{\partial J(\theta)}{\partial \theta} = 0$:

$$\frac{\partial}{\partial \theta}\left(Y^TY - 2Y^TX\theta + \theta^TX^TX\theta\right) = 0; \implies -2X^TY + 2X^TX\theta = 0; \implies \theta = (X^TX)^{-1}X^TY$$

Now find the closed form solution that minimizes this loss function (assume A is symmetric):

$$J(\theta) = (Y - X\theta)^TA(Y - X\theta)$$

(a) $\theta = (X^TAX)^{-1}X^TY$
(b) $\theta = (X^TX)^{-1}X^TAY$
(c) $\theta = (X^TAX)^{-1}X^TAY$
(d) $\theta = (X^TAX)^{-1}X^TA^{-1}Y$
(e) None of these

Write all correct options ——————-

$$J(\theta) = (Y - X\theta)^T A(Y - X\theta).$$

Symmetry of A: Since $A = A^T$, certain terms will be equal.

Let us expand the quadratic form:

$$\begin{aligned} J(\theta) &= (Y - X\theta)^T A(Y - X\theta) \\ &= Y^T AY - Y^T A(X\theta) - (X\theta)^T AY + (X\theta)^T A(X\theta). \end{aligned}$$

$$(X\theta)^T = \theta^T X^T.$$

Thus, we can rewrite the expression as:

$$J(\theta) = Y^T AY - Y^T AX\theta - \theta^T X^T AY + \theta^T X^T AX\theta.$$

Since A is symmetric, we have

$$Y^T AX\theta = \theta^T X^T AY,$$

because both are scalars (and the transpose of a scalar is the same scalar). Therefore, the two middle terms can be combined:

$$-Y^T AX\theta - \theta^T X^T AY = -2\,\theta^T X^T AY.$$

Thus, the loss function becomes:

$$J(\theta) = Y^T AY - 2\,\theta^T X^T AY + \theta^T X^T AX\theta.$$

We now compute the gradient $J(\theta)$.

(a) For a constant vector $\mathbf{c}$,

$$\frac{\partial}{\partial\theta}(\theta^T \mathbf{c}) = \mathbf{c}.$$

(b) For a symmetric matrix B,

$$\frac{\partial}{\partial\theta}(\theta^T B\theta) = 2B\theta.$$

Apply these to each term:

- The term $Y^T AY$ is independent of θ, so

$$\frac{\partial}{\partial\theta}(Y^T AY) = 0.$$

- The linear term $-2\,\theta^T X^T AY$ yields

$$\frac{\partial}{\partial\theta}\Big(-2\,\theta^T X^T AY\Big) = -2\,X^T AY.$$

- The quadratic term $\theta^T X^T A X \theta$ involves the symmetric matrix $B = X^T A X$ (note that B is symmetric because A is symmetric). Thus,

$$\frac{\partial}{\partial \theta}\Big(\theta^T X^T A X \theta\Big) = 2\, X^T A X\, \theta.$$

Therefore, the gradient is:

$$\frac{\partial}{\partial \theta}(J(\theta)) = 0 - 2\, X^T A Y + 2\, X^T A X\, \theta = 2\, X^T A X\, \theta - 2\, X^T A Y.$$

To find the minimum, we set the gradient equal to the zero vector:

$$2\, X^T A X\, \theta - 2\, X^T A Y = 0.$$

Dividing both sides by 2 (a scalar division, which does not affect the equality) gives:

$$X^T A X\, \theta - X^T A Y = 0.$$

Rearrange to obtain:

$$X^T A X\, \theta = X^T A Y.$$

Assuming that $X^T A X$ is invertible, we can solve for θ by multiplying both sides by $(X^T A X)^{-1}$:

$$\theta = (X^T A X)^{-1} X^T A Y.$$

Final Answer:

$$\boxed{\theta = (X^T A X)^{-1} X^T A Y, \quad \text{(Option (c) is correct.)}}$$

---

12. Consider the function

$$f(w) = w^2 + w + 1$$

We want to find the minima of the function using gradient descent. We start at $w^0 = 5.0$.
Write update equation for computing w^{k+1} from w^k. Ans: ————-
What should be the learning rate η so that we reach the minima in a single step?
Ans: ————————

Derivative of $f(w)$.

$$f(w) = w^2 + w + 1 \implies f'(w) = 2\, w + 1.$$

The gradient descent update equation.

In standard gradient descent, the update for the parameter w is

$$w^{(k+1)} = w^{(k)} - \eta \left.\frac{\partial f}{\partial w}\right|_{w=w^{(k)}}.$$

Since $\frac{\partial f}{\partial w} = 2w + 1$, we get

$$w^{(k+1)} = w^{(k)} - \eta\left(2w^{(k)} + 1\right).$$

This is the *update equation* we use at each iteration k.

Identify the true minimum of $f(w)$.

To find the exact minimizer, set the derivative to zero:

$$f'(w^*) = 2w^* + 1 = 0 \quad \Longrightarrow \quad w^* = -\tfrac{1}{2}.$$

Determine η so we reach $w^* = -\frac{1}{2}$ in one step from $w^{(0)} = 5$.

After one iteration (k=0 to k=1), we have

$$w^{(1)} = w^{(0)} - \eta\left(2w^{(0)} + 1\right).$$

Plug in $w^{(0)} = 5$:

$$w^{(1)} = 5 - \eta\left(2 \cdot 5 + 1\right) = 5 - \eta \times 11.$$

We want $w^{(1)}$ to be exactly the minimizer $-\frac{1}{2}$. So,

$$5 - 11\eta = -\tfrac{1}{2}.$$

Rearrange and solve for η:

$$-11\eta = -\tfrac{1}{2} - 5 = -\tfrac{1}{2} - \tfrac{10}{2} = -\tfrac{11}{2},$$

$$\eta = \frac{-\frac{11}{2}}{-11} = \tfrac{1}{2}.$$

Answer:

(a) *Update equation:*

$$w^{(k+1)} = w^{(k)} - \eta\left(2w^{(k)} + 1\right).$$

(b) *Learning rate for a single-step solution:*

$$\boxed{\eta = \tfrac{1}{2}.}$$

With $\eta = 0.5$, starting at $w^{(0)} = 5$, we immediately jump to $w^* = -\frac{1}{2}$ in one update step.

Problem 12 of other set:

Consider the function

$$f(w) = w^2 + w + 1$$

We want to find the minima of the function using gradient descent. We start at $w^0 = -5.0$.

Write update equation for computing w^{k+1} from w^k. Ans: ————-

What should be the learning rate η so that we reach the minima in a single step?

Ans: ————————

Solution:

Compute the derivative $\frac{df}{dw}$.

$$f'(w) \;=\; \frac{d}{dw}(w^2 + w + 1) \;=\; 2w + 1.$$

Write the gradient descent update equation.

In gradient descent, the parameter update is:

$$w^{(k+1)} \;=\; w^{(k)} \;-\; \eta \left.\frac{df}{dw}\right|_{w=w^{(k)}}.$$

Since $\frac{df}{dw} = 2w + 1$, we have:

$$\boxed{w^{(k+1)} \;=\; w^{(k)} \;-\; \eta\,(2\,w^{(k)} + 1).}$$

Compute $w^{(1)}$ from $w^{(0)} = 5$ with $\eta = 0.1$.

$$w^{(1)} \;=\; w^{(0)} \;-\; 0.1\,(2\,w^{(0)} + 1).$$

Plugging in $w^{(0)} = 5$:

$$w^{(1)} \;=\; 5 \;-\; 0.1\,(2 \cdot 5 + 1) \;=\; 5 \;-\; 0.1 \times 11 \;=\; 5 \;-\; 1.1 \;=\; 3.9.$$

Answer:

(a) *Update Equation:*

$$w^{(k+1)} = w^{(k)} - \eta\,(2\,w^{(k)} + 1).$$

(b) *Value of $w^{(1)}$ for $\eta = 0.1$ starting at $w^{(0)} = 5$:*

$$\boxed{w^{(1)} = 3.9.}$$