Type: MCQ

Q1. Consider the equation of hyperplane for SVM is x1-x2+2= 0, Determine the margin.(0.5)

1. ** 1/√2
2. √2
3. 1
4. 2

Q2. Determine the nature of the curve y=x^2-e^x at x=0 and x=1 respectively. (0.5)

1. **Convex, Concave.
2. Concave, Convex.
3. Concave, Concave.
4. Convex, Convex.

Q3. The feature selection technique that determines the relevance of features by only considering the statistical properties between the feature vectors is (0.5)

1. Wrapper method
2. **Filter method
3. Embedded methods
4. Regularization

Q4. If A is a 3x3 matrix, then the trace of A is equal to (0.5)

1. Product of eigen values
2. Product of diagonal elements
3. **Sum of eigen values
4. Sum of diagonal elements

Q5. The co-ordinates, where the the curve changes from nature from concave to convex or vice versa is called (0.5)

1. Co-ordinates of interation
2. Co-ordinates of regularization
3. **Co-ordinates of inflection
4. Co-ordinates of Norm

Q6. A researcher applies K-Means to a dataset with non-spherical clusters. The algorithm performs poorly. What is the most likely reason? (0.5)

1. ** K-Means assumes clusters are spherical and may not work well for non-spherical clusters.
2. The dataset has too many missing values.
3. The dataset is too large for K-Means to handle.
4. K-Means is not sensitive to cluster shape.

Q7. A data scientist is using the Expectation-Maximization (EM) algorithm to cluster a dataset where each point is believed to belong to one of several Gaussian distributions. During the iterative process, the algorithm alternates between updating the probabilities of data points belonging to clusters and updating the cluster parameters. What is the primary reason for using EM in this scenario? (0.5)

1. EM can directly assign each data point to a single cluster with certainty.
2. ** EM is useful when data has missing or hidden variables, such as latent cluster memberships.
3. EM finds the exact maximum likelihood solution in one iteration.
4. EM does not require an initial guess for the model parameters.

Q8. A researcher is using a Gaussian Mixture Model (GMM) for clustering and observes that the Expectation-Maximization (EM) algorithm iteratively updates the parameters. The probability density function of a GMM is given by:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where $\pi_k$ are the mixing coefficients, $\mathcal{N}(x|\mu_k, \Sigma_k)$ represents a multivariate Gaussian distribution with mean $\mu_k$ and covariance $\Sigma_k$.

During the Expectation step (E-step), what does the algorithm compute? (0.5)

1. It updates the means $\mu_k$, covariances $\Sigma_k$ , and mixing coefficients $\pi_k$.
2. ** It computes the posterior probabilities (responsibilities) of each data point belonging to each Gaussian component.
3. It maximizes the likelihood function to obtain updated parameter values.
4. It assigns each data point to the most likely cluster with hard clustering.

Q9. A psychologist is studying anxiety and depression symptoms in a group of patients. They collect responses to 10 different questionnaire items and suspect that these symptoms are influenced by two **latent factors**: **General Anxiety** and **Depressive Mood**. To identify these factors, they apply **Factor Analysis (FA)** using the model:

$$X = LF + \epsilon$$

where $X$ represents observed symptoms, $L$ is the factor loading matrix, $F$ contains the latent factors, and $\epsilon$ is the unique noise component.

Which of the following correctly describes how Factor Analysis (FA) differs from Principal Component Analysis (PCA)?(0.5)

1. **FA captures underlying latent factors that cause correlations between variables, while PCA simply finds orthogonal projections that maximize variance.
2. FA and PCA are identical in their mathematical approach and always yield the same results.
3. FA only works if the variables are independent, whereas PCA allows correlated variables.

4. FA maximizes the variance in the data, whereas PCA finds latent variables that explain shared variance.

Q10. A company applies K-Means clustering to segment customers based on their purchasing behavior. Given the following four customer data points, they initialize two cluster centroids randomly. (Assume Initial Centroids: Centroid 1: (500, 10) and Centroid 2: (900, 20)

**Customer Data:**

| Customer ID | Total Spend ($) | Number of Purchases |
|---|---|---|
| 101 | 500 | 10 |
| 102 | 1200 | 25 |
| 103 | 300 | 7 |
| 104 | 900 | 20 |

Using the Euclidean distance formula,

$$d(A, B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

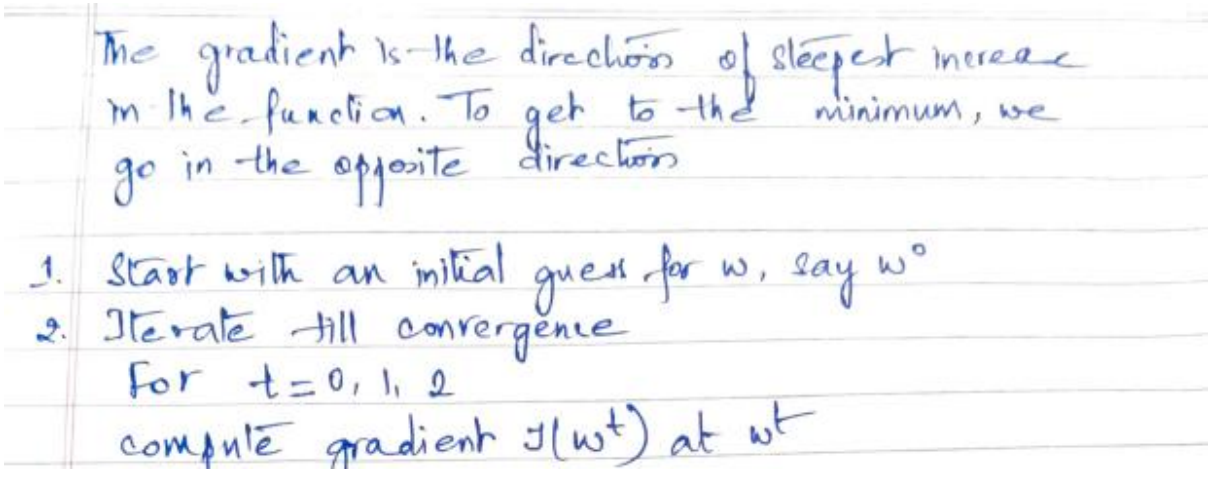
compute the distance of Customer 102 (1200, 25) from both centroids and determine which cluster it will be assigned to. (0.5)

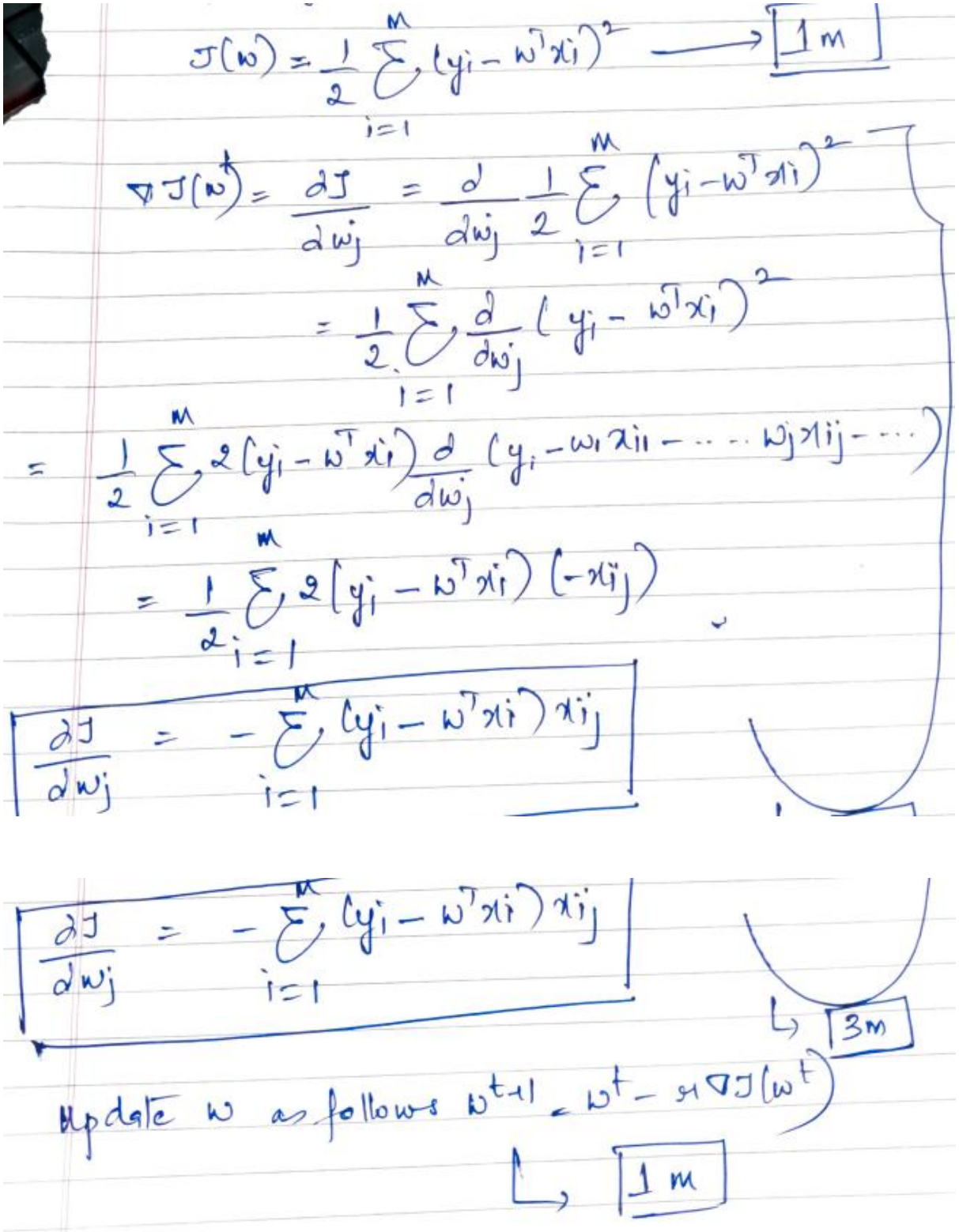
1. ** Cluster 2
2. Cluster 1
3. Cannot be determined
4. Both clusters

Type: DES

Q11. Mathematically, illustrate the gradient descent algorithm used in LMS for updating the weights. (5)

Answer:






Q12. A company wants to group its customers based on their spending behavior. The dataset contains the following information for each customer:

| Customer ID | Total Spend ($) | Number of Purchases |
|---|---|---|
| 1 | 200 | 3 |
| 2 | 450 | 7 |
| 3 | 700 | 10 |
| 4 | 850 | 12 |
| 5 | 900 | 15 |

The company applies K-Means clustering with K=2 and initializes the centroids randomly as:

Centroid 1: (200, 3) and Centroid 2: (900, 15).

a. Compute the Euclidean distance of each customer from both centroids using the formula:

$$d(A,B) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

b. Assign each customer to the nearest centroid based on the computed distance.

c. Compute the new centroids after reassigning the points to clusters.

d. Repeat the process for one more iteration and update the centroids again. (5)

**Answer: Iteration 1 – 2.5M, Iteration 2 – 2.5M**

a.

| | | |
|---|---|---|
| (200, 3) | $\sqrt{(200-200)^2 + (3-3)^2} = 0.000$ | $\sqrt{(200-900)^2 + (3-15)^2} = 700.103$ |
| (450, 7) | $\sqrt{(450-200)^2 + (7-3)^2} = 250.032$ | $\sqrt{(450-900)^2 + (7-15)^2} = 450.071$ |
| (700, 10) | $\sqrt{(700-200)^2 + (10-3)^2} = 500.049$ | $\sqrt{(700-900)^2 + (10-15)^2} = 200.062$ |
| (850, 12) | $\sqrt{(850-200)^2 + (12-3)^2} = 650.062$ | $\sqrt{(850-900)^2 + (12-15)^2} = 50.090$ |
| (900, 15) | $\sqrt{(900-200)^2 + (15-3)^2} = 700.103$ | $\sqrt{(900-900)^2 + (15-15)^2} = 0.000$ |

b.

| Distance to C1 (200,3) | Distance to C2 (900,15) | Assigned Cluster |
|---|---|---|
| 0.000 | 700.103 | **Cluster 1** |
| 250.032 | 450.071 | **Cluster 1** |
| 500.049 | 200.062 | **Cluster 2** |
| 650.062 | 50.090 | **Cluster 2** |
| 700.103 | 0.000 | **Cluster 2** |

c. Compute new centroids

For points (200,3) and (450,7)

$$x_{\text{new}} = \frac{200 + 450}{2} = 325$$

$$y_{\text{new}} = \frac{3 + 7}{2} = 5$$

For points (700,10), (850,12), and (900,15)

$$x_{\text{new}} = \frac{700 + 850 + 900}{3} = 816.67$$

$$y_{\text{new}} = \frac{10 + 12 + 15}{3} = 12.33$$

Repeat the process for one more iteration

| | | |
|---|---|---|
| $\sqrt{(200-325)^2+(3-5)^2} = 125.016$ | $\sqrt{(200-816.67)^2+(3-12.33)^2} = 617.022$ | Cluster 1 |
| $\sqrt{(450-325)^2+(7-5)^2} = 125.016$ | $\sqrt{(450-816.67)^2+(7-12.33)^2} = 366.235$ | Cluster 1 |
| $\sqrt{(700-325)^2+(10-5)^2} = 375.033$ | $\sqrt{(700-816.67)^2+(10-12.33)^2} = 117.540$ | Cluster 2 |
| $\sqrt{(850-325)^2+(12-5)^2} = 525.047$ | $\sqrt{(850-816.67)^2+(12-12.33)^2} = 33.33$ | Cluster 2 |
| $\sqrt{(900-325)^2+(15-5)^2} = 575.087$ | $\sqrt{(900-816.67)^2+(15-12.33)^2} = 84.07$ | Cluster 2 |

**Final Centroids**

**Centroid 1**: (325, 5)

**Centroid 2**: (816.67, 12.33)

Q13. Determine the gradient of the quadratic function f(x) by considering $\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j$ (3)

$$f(x) = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j.$$

To take the partial derivative, we'll consider the terms including $x_k$ and $x_k^2$ factors separately:

$$\begin{aligned}\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j \\ &= \frac{\partial}{\partial x_k}\left[\sum_{i\neq k}\sum_{j\neq k} A_{ij}x_i x_j + \sum_{i\neq k} A_{ik}x_i x_k + \sum_{j\neq k} A_{kj}x_k x_j + A_{kk}x_k^2\right] \\ &= \sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j + 2A_{kk}x_k\end{aligned}$$

Step 2:- 1.5m

Step 3:-1.5m

Q14. Consider, the initial weights w1 = 1.2, w2 = 0.6, threshold =1, and learning rate = 0.5, graphically illustrate and implement an AND gate with binary inputs using single layer perceptron.(3)

For Training Instance 1: A=0, B=0 and Target = 0

wi.xi = 0*1.2 + 0*0.6 = 0

This is not greater than the threshold of 1, so the output = 0, Here the target is same as calculated output.

For Training Instance 2: A=0, B=1 and Target = 0

wi.xi = 0*1.2 + 1*0.6 = 0.6

This is not greater than the threshold of 1, so the output = 0. Here the target is same as calculated output.

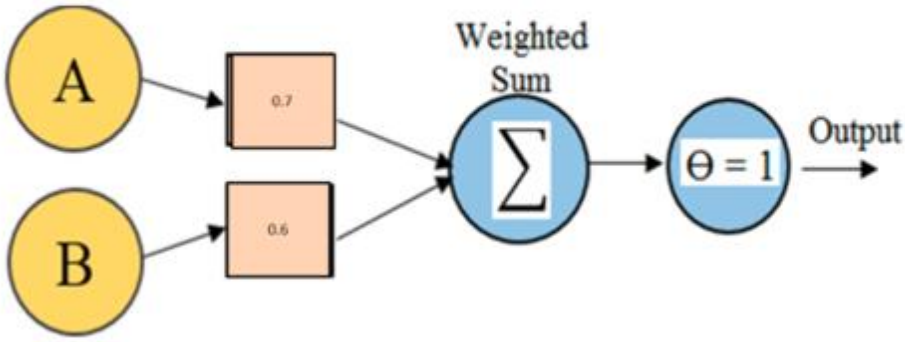
wi.xi = 1*1.2 + 0*0.6 = 1.2

This is greater than the threshold of 1, so the output = 1. Here the target does not match with the calculated output.

Hence we need to update the weights.

$$wi = wi + n(t-o)xi$$
$$w1 = 1.2 + 0.5(0-1)1 = 0.7$$
$$w2 = 0.6 + 0.5(0-1)0 = 0.6$$

Hence the final weights are w1= 0.7 and w2 = 0.6, Threshold = 1 and Learning Rate n = 0.5.




**Equation- 1M**

**Computation-1m**

**Diagram-1m**

Q15. A researcher is using the Expectation-Maximization (EM) algorithm to fit a **Gaussian Mixture Model (GMM)** with two components to a dataset consisting of five values: X={2,3,5,8,9}. The initial parameters for the two Gaussian components are as follows: Component 1 has a mean $\mu_1$=3, variance $\sigma_1^2$=1, and weight $w_1$=0.5, while Component 2 has a mean $\mu_2$=8, variance $\sigma_2^2$=1 and weight $w_2$=0.5. In the Expectation Step (E-step), compute the responsibility $r_{i,1}$ that the first Gaussian component takes for the data point x=5. Use the formula:

$$r_{i,1} = \frac{w_1 \mathcal{N}(x_i|\mu_1, \sigma_1^2)}{w_1 \mathcal{N}(x_i|\mu_1, \sigma_1^2) + w_2 \mathcal{N}(x_i|\mu_2, \sigma_2^2)}$$

where $\mathcal{N}(x|\mu, \sigma^2)$ represents the Gaussian probability density function.

In the Maximization Step (M-step), explain how the means $\mu_1$ and $\mu_2$ should be updated using the computed responsibilities. (3)

Iteration 1:

**Expectation Step →1M**

For **Component 1** ($\mu_1 = 3, \sigma_1^2 = 1$):

$$P(x|3,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-3)^2}{2}}$$

For **Component 2** ($\mu_2 = 8, \sigma_2^2 = 1$):

$$P(x|8,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-8)^2}{2}}$$

Responsibilities:

| $x$ | $r_{i,1}$ (responsibility for C1) |
|---|---|
| 2 | 0.999 |
| 3 | 0.998 |
| 5 | 0.971 |
| 8 | 0.002 |
| 9 | 0.001 |

Similarly, $r_{i,2} = 1 - r_{i,1}$.

**Maximization step→1M**

For **Component 1 (C1, $\mu_1$)**:

$$\mu_1 = \frac{(0.999 \times 2) + (0.998 \times 3) + (0.971 \times 5) + (0.002 \times 8) + (0.001 \times 9)}{0.999 + 0.998 + 0.971 + 0.002 + 0.001}$$

$$\mu_1 = \frac{1.998 + 2.994 + 4.855 + 0.016 + 0.009}{3.971}$$

$$\mu_1 \approx 3.50$$

For **Component 2 (C2, $\mu_2$)**:

$$\mu_2 = \frac{(0.001 \times 2) + (0.002 \times 3) + (0.029 \times 5) + (0.998 \times 8) + (0.999 \times 9)}{0.001 + 0.002 + 0.029 + 0.998 + 0.999}$$

$$\mu_2 = \frac{0.002 + 0.006 + 0.145 + 7.984 + 8.991}{2.029}$$

$$\mu_2 \approx 8.5$$

After **one EM iteration**:

- New $\mu_1$ = 3.50
- New $\mu_2$ = 8.50

Iteration 2: (1M)

| $x$ | $r_{i,1}$ (responsibility for C1) |
|---|---|
| 2 | 0.999 |
| 3 | 0.999 |
| 5 | 0.968 |
| 8 | 0.022 |
| 9 | 0.001 |

For **Component 1** (C1, $\mu_1$):

$$\mu_1 = \frac{(0.999 \times 2) + (0.999 \times 3) + (0.968 \times 5) + (0.022 \times 8) + (0.001 \times 9)}{0.999 + 0.999 + 0.968 + 0.022 + 0.001}$$

$$\mu_1 = \frac{1.998 + 2.997 + 4.840 + 0.176 + 0.009}{3.989}$$

$$\mu_1 \approx 3.50$$

For **Component 2** (C2, $\mu_2$):

$$\mu_2 = \frac{(0.001 \times 2) + (0.001 \times 3) + (0.032 \times 5) + (0.978 \times 8) + (0.999 \times 9)}{0.001 + 0.001 + 0.032 + 0.978 + 0.999}$$

$$\mu_2 = \frac{0.002 + 0.003 + 0.160 + 7.824 + 8.991}{2.011}$$

$$\mu_2 \approx 8.50$$

- New $\mu_1$ = 3.50
- New $\mu_2$ = 8.50

The means have not changed significantly, indicating that the EM algorithm has converged.

Q16.

Consider a binary classification problem where a model is trained on a dataset with a small number of samples and high model complexity. The model achieves nearly 100% accuracy on the training data but performs poorly on new test data.

**a.** Explain this scenario in terms of the **bias-variance tradeoff**.

**b.** Suppose you want to provide a probabilistic guarantee on the generalization error of your model. Which of **Hoeffding's inequality** or **Chernoff bound** would be more suitable, and why?

**c.** What role does **VC dimension** play in determining the sample complexity needed to ensure good generalization? (3)

Answer:

**(a) Bias-Variance Tradeoff Explanation (1 Mark)**

The scenario describes a model with high complexity that achieves nearly 100% accuracy on training data but performs poorly on test data. This indicates overfitting, where the model has low bias (it learns the training data very well) but high variance (it does not generalize to unseen data).

- A model with low bias captures complex patterns but also noise, leading to poor performance on new data.
- A model with high variance is overly sensitive to training data variations, causing poor generalization.

- A balanced model should have an optimal tradeoff between bias and variance to ensure good generalization.

**(b) Choosing Hoeffding's Inequality vs. Chernoff Bound (1 Mark)**

To provide a probabilistic guarantee on the generalization error, Hoeffding's inequality is more suitable.

- Hoeffding's inequality provides an upper bound on the probability that the empirical mean deviates from the expected mean, making it useful for worst-case guarantees on generalization error.
- Chernoff bound is tighter and useful for exponential decay analysis in large-sample settings, but Hoeffding's inequality is preferred when dealing with small sample sizes, as in this scenario.

**(c) Role of VC Dimension in Sample Complexity (1 Mark)**

The Vapnik-Chervonenkis (VC) dimension measures a model's capacity to fit various patterns in data.

- A higher VC dimension means the model can fit more complex functions but may require more training samples to generalize well.
- The sample complexity needed for good generalization depends on the VC dimension

$$N \propto \frac{VC}{\epsilon^2}$$

where N is the number of samples, VC is the VC dimension, and ϵ is the desired generalization error.

- If the training dataset is too small relative to the VC dimension, the model will overfit, leading to poor generalization.

Q17. A hospital is developing a machine learning model to predict whether a patient has diabetes based on medical features such as blood glucose level, BMI, age, and blood pressure.

a. What type of learning approach should be used for this problem? Justify your answer.

b. If the dataset contains 5000 patients, with 4000 non-diabetic and 1000 diabetic cases, what challenge might arise during training, and how can it be addressed?

c. Given a new patient with feature values x=[120,25,45,80], how would a trained logistic regression model predict the probability of diabetes? (Assume model parameters are $w_0 = -5$, $w_1 = 0.04$, $w_2 = 0.1$, $w_3 = 0.02$, and $w_4 = 0.05$, The general equation for Logistic Regression is:)

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(w_0 + \sum_{i=1}^{n} w_i x_i)}}$$ (3)

Answer:

a. The appropriate learning approach for this problem is supervised learning, specifically classification.

Justification: The dataset consists of labeled data (patients are either diabetic or non-diabetic), making supervised learning the best approach. **(0.5M)**

b. Challenges: **(0.5M)**

i. The model may become biased towards the majority class (non-diabetic) and predict "non-diabetic" most of the time.
ii. Poor recall for the minority class (diabetic), leading to missed diagnoses.

Solutions: **(0.5M)**

i. Resampling Methods:

Oversampling: Increase the diabetic class instances (e.g., using SMOTE).

Undersampling: Reduce the non-diabetic class instances.

ii. Class Weighting: Assign higher weights to diabetic cases during training.

c. LR **(1.5M)**

Compute Linear combination

$$z = -5 + 4.8 + 2.5 + 0.9 + 4.0$$

$$z = -5 + 12.2$$

$$z = 7.2$$

Apply sigmoid function

The probability of having diabetes is given by:

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}$$

$$P(y = 1|x) = \frac{1}{1 + e^{-7.2}}$$

Approximating $e^{-7.2} \approx 0.00075$:

$$P(y = 1|x) = \frac{1}{1 + 0.00075} \approx \frac{1}{1.00075} \approx 0.99925$$