CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2024)
Quiz 3 (Set A)
Date of Examination: 08.11.2024 Duration: 1 hour Total Marks: 10 marks

Instructions
• Attempt all questions. State any assumptions you have made clearly.
• MCQs may have multiple correct options. No evaluation without suitable justification.
• Standard institute plagiarism policy holds.

---

1. **(1 mark)** Raju read about Support Vector Machines (SVM) some time ago and wrote down some statements. Help Raju by identifying which of the following statements is False with appropriate reasons/explanations:

a. Data points on support vectors are the easiest to classify.
b. Data must be linearly separable.
c. Normalised data does not have any effect on SVM performance.
d. SVM is a non-probabilistic model.

Note: check the assumptions and evaluate accordingly

Solution: A, B, and C are correct answers.

A. **False statement.** Hardest to classify.
B. **False statement.** Kernel methods can be used.
C. **False statement.** Normalized data is better for SVM
D. **True statement.** It gives a perfect prediction.

Rubrics: 1 mark for all correct answers + explanation

2. **(1 mark)** Explain how the regularisation parameter *C* affects the trade-off between margin size and classification errors.

The regularisation parameter *C* in SVM controls the trade-off between margin size and classification errors.

High *C* Value (Strong Regularization):

- **Emphasis on Correct Classification**: A large C assigns a higher penalty to classification errors. The model will prioritise minimising these errors over maximising the margin.
- **Narrower Margin**: To reduce the classification errors, the SVM may choose a hyperplane with a narrower margin that better fits the training data points.
- **Overfitting Risk**: Because the model focuses heavily on correctly classifying every training point, especially outliers, it can lead to overfitting. The decision boundary becomes more complex and sensitive to noise.

Low *C* Value (Weak Regularization):

- **Emphasis on Margin Maximization**: A smaller ***C*** places less importance on misclassification errors, allowing some data points to be within the margin or misclassified without a significant penalty.
- **Wider Margin**: The model prioritises finding a hyperplane that maximises the margin, even if it means allowing some misclassification.
- **Underfitting Risk**: With lower sensitivity to misclassifications, the model may oversimplify the decision boundary, potentially underfitting the data if the classes are not well-separated.

Rubrics- 0.5 for high C value and 0.5 for low C value. There can be more effects, if correct give marks.

3. **(1 mark)** Why would you use the Kernel Trick?

Solution:

When it comes to classification problems, the goal is to establish a decision boundary that maximises the margin between the classes. However, in the real world, this task can become difficult when dealing with **non-linearly separable data**. One approach to solve this problem is to perform a data transformation process, in which we map all the data points to a **higher dimension**, find the boundary and make the classification.

That sounds alright, however, when there are more and more dimensions, computations within that space become more and more expensive. In such cases, **the kernel trick allows us to operate in the original feature space without computing the coordinates of the data** in a higher-dimensional space and therefore offers a more efficient and less expensive way to transform data into higher dimensions.

There exist different kernel functions, such as:

1. **Linear** (no transformation): $K(x, x') = x^T x'$
2. **Polynomial**: $K(x, x') = (x^T x' + c)^d$
3. **RBF**: $K(x, x') = \exp\left(-\frac{|x - x'|^2}{2\sigma^2}\right)$
4. **Sigmoid**: $K(x, x') = \tanh(\kappa x^T x' + \Theta)$

Rubrics- 1 mark for the correct answer

4. **(1 mark)** What is the maximum possible value of the Radial Basis Function (RBF) Kernel? Give justification as well.

a. 0
b. 1
c. infinity
d. -1

Solution) (b) correct option

Explanation: The maximum value that the RBF kernel can be is 1 and occurs when d is 0, when the points are the same, i.e. X = X.

Rubrics- 1 mark for the correct answer and correct justification.

5. **(1 mark)** You are given a labelled binary classification data set with N data points and D features. Suppose that N < D. In training an SVM on this data set, which of the following kernels is likely to be most appropriate?

a. Linear kernel
b. Quadratic kernel
c. Higher-order polynomial kernel
d. RBF kernel

Answer: (a) Linear kernel

Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.

When the number of examples is less compared to the number of features, you would not have enough data to fit a non-linear SVM i.e. SVM with a non-linear kernel. SVM with a linear kernel (or without a kernel) is one way to go.

Rubrics- 1 mark for the correct answer and correct justification.

6. **(2 marks)** Consider the following XOR dataset:

a. (0,0) with class label +1

b. (0,1) with class label −1

c. (1,0) with class label −1

d. (1,1) with class label +1

Using a polynomial kernel transformation defined as: $(x_1,x_2) \rightarrow (z_1,z_2) = (x_1+x_2, x_1 \cdot x_2)$

a. Transform each point (x1,x2) to the new feature space (z_1,z_2).

b. Plot or describe the position of each transformed point and discuss if the points are now linearly separable in this transformed 2D space.

Note: they have to find the decision boundary for part (b)

Solution:

1. Apply the Transformation:

We will calculate (z_1, z_2) for each data point:

- For $(0, 0)$:

$$(z_1, z_2) = (0 + 0, 0 \cdot 0) = (0, 0) \text{ with label } +1$$

- For $(0, 1)$:

$$(z_1, z_2) = (0 + 1, 0 \cdot 1) = (1, 0) \text{ with label } -1$$

- For $(1, 0)$:

$$(z_1, z_2) = (1 + 0, 1 \cdot 0) = (1, 0) \text{ with label } -1$$

- For $(1, 1)$:

$$(z_1, z_2) = (1 + 1, 1 \cdot 1) = (2, 1) \text{ with label } +1$$

2. Describe the Position of Transformed Points:

In the new feature space (z1,z2) are:

- Point (0,0) has the label +1
- Point (1,0) has label −1(this point appears twice because both (0,1) and (1,0) map to (1,0))
- Point (2,1) has label +1

Analyse Linearly Separability

In the transformed 2D space:

- The points (0,0) and (2,1), both labelled +1, are on opposite ends of the feature space.
- The point (1,0), labelled −1, lies in between them.

To determine if they are linearly separable, consider a simple decision boundary.

$$z2 = z1 - 1$$

- For (0,0): z2 = 0 and z1 − 1 = −1. So 0 > −1, correctly classified as **+1**.
- For (2,1): z2 = 1 and z1 − 1= 1. So 1 = 1, correctly classified as **+1**.

- For (1,0): z2 = 0 and z1 − 1 = 0. So 0 = 0, making this point lie on the boundary, still correctly classified as **−1**.

Rubric- 1 mark for correct part a and 1 mark for correct part b. There can be more than one decision boundary. Consider all decision boundaries if correct. They need not find the equation for the decision boundary.

7. **(3 marks)** Derive dual form equations for soft margin SVM. (**added in class-** derive the complete formulation of hard margin SVM.)

Step 1: Primal Formulation (0.5 Marks)

Given n data points (x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$, we aim to find a hyperplane $w^T x + b = 0$ with maximum margin and allowable misclassification. Slack variables $\xi_i \geq 0$ are introduced for misclassified points.

Primal form:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n$$

where C controls the margin-misclassification trade-off.

Step 2: Lagrangian and Conditions (1.5 Marks)

Using Lagrange multipliers $\alpha_i \geq 0$ (margin) and $\mu_i \geq 0$ (slack), we form the Lagrangian:

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \mu_i \xi_i$$

Minimizing with respect to w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Minimizing with respect to b:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

Minimizing with respect to ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow 0 \leq \alpha_i \leq C$$

Step 3: Dual Formulation (1.0 Marks)

Substitute $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ to get the dual problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n$$

Rubrics:

- Primal Formulation (0.5 Marks).
- Lagrangian Setup and Conditions (1.5 Marks): Formulate Lagrangian, minimise w, b, ξi .

- Dual Formulation (1.0 Marks).

OR

Solution

Step 1: Define the Problem and Margins (0.5 Marks)

- Clearly state the optimization objective for maximizing the margin:

$$\max_{\gamma,w,b} \gamma \quad \text{subject to } y^{(i)}(w^T x^{(i)} + b) \geq \gamma \quad \forall i$$

- Define the relationship between the **geometric margin** and the constraints $||w|| = 1$.

Step 2: Reformulate to Eliminate γ (0.5 Marks)

- Reformulate the problem by scaling w and b such that $\gamma = 1$, leading to:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{subject to } y^{(i)}(w^T x^{(i)} + b) \geq 1 \quad \forall i$$

Step 3: Optimization Problem Setup (0.5 Marks)

- Clearly write down the **primal optimization problem**:

$$\min_{w,b} \frac{1}{2}||w||^2, \quad \text{subject to linear constraints.}$$

Step 4: Lagrangian and Dual Formulation (1 Mark)

- Construct the **Lagrangian**:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1$$

- Derive conditions for w and b using partial derivatives:

$$\nabla_w L = 0 \implies w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial L}{\partial b} = 0 \implies \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- Substitute back into the Lagrangian to get the **dual formulation**.

Step 5: Identify Support Vectors and Decision Boundary (0.5 Marks)

- Show how the **decision boundary** depends on support vectors:

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)} \cdot x) + b\right)$$

- Highlight that only **support vectors** (points with $\alpha_i > 0$) contribute to the classifier.

Rubrics:
Problem Definition and Margins: 0.5 Marks
Reformulation to Eliminate γ: 0.5 Marks
Primal Formulation: 0.5 Marks
Lagrangian and Dual Derivation: 1 Mark
Support Vectors and Decision Boundary: 0.5 Marks

CSE343/CSE543/ECE363/ECE563: Machine Learning Sec A (Monsoon 2024)
Quiz 3 (Set B)
Date of Examination: 08.11.2024 Duration: 1 hour Total Marks: 10 marks

Instructions
- Attempt all questions. State any assumptions you have made clearly.
- MCQs may have multiple correct options. No evaluation without suitable justification.
- Standard institute plagiarism policy holds.

---

1. **(1 mark)** Raju read about Support Vector Machines (SVM) some time ago and wrote down some statements. Help Raju by identifying which of the following statements is False with appropriate reasons/explanations:
   a. Data points on support vectors are the easiest to classify.
   b. Data must be linearly separable.
   c. Normalized data does not have any effect on SVM performance.
   d. SVM is a non-probabilistic model.

   Solution: A, B, and C are correct answers.
   E. False statement. Hardest to classify.
   F. False statement. Kernel methods can be used.
   G. False statement. Normalized data is better for SVM
   H. True statement. It gives a perfect prediction.
   Rubrics: 1 mark for all correct answers + explanation

2. **(1 mark)** Explain how the regularisation parameter *C* affects the trade-off between margin size and classification errors.

   The regularisation parameter *C* in SVM controls the trade-off between margin size and classification errors.

   High *C* Value (Strong Regularization):
   - **Emphasis on Correct Classification**: A large C assigns a higher penalty to classification errors. The model will prioritise minimising these errors over maximising the margin.
   - **Narrower Margin**: To reduce the classification errors, the SVM may choose a hyperplane with a narrower margin that better fits the training data points.
   - **Overfitting Risk**: Because the model focuses heavily on correctly classifying every training point, especially outliers, it can lead to overfitting. The decision boundary becomes more complex and sensitive to noise.

   Low *C* Value (Weak Regularization):
   - **Emphasis on Margin Maximization**: A smaller ***C*** places less importance on misclassification errors, allowing some data points to be within the margin or misclassified without a significant penalty.
   - **Wider Margin**: The model prioritises finding a hyperplane that maximises the margin, even if it means allowing some misclassification.
   - **Underfitting Risk**: With lower sensitivity to misclassifications, the model may oversimplify the decision boundary, potentially underfitting the data if the classes are not well-separated.

   Rubrics- 0.5 for high C value and 0.5 for low C value. There can be more effects, if correct give marks.

3. **(1 mark)**Why would you use the Kernel Trick?

Solution:

When it comes to classification problems, the goal is to establish a decision boundary that maximises the margin between the classes. However, in the real world, this task can become difficult when dealing with **non-linearly separable data**. One approach to solve this problem is to perform a data transformation process, in which we map all the data points to a **higher dimension**, find the boundary and make the classification.

That sounds alright, however, when there are more and more dimensions, computations within that space become more and more expensive. In such cases, **the kernel trick allows us to operate in the original feature space without computing the coordinates of the data** in a higher-dimensional space and therefore offers a more efficient and less expensive way to transform data into higher dimensions.

There exist different kernel functions, such as:

1. **Linear** (no transformation): $K(x, x') = x^T x'$
2. **Polynomial**: $K(x, x') = (x^T x' + c)^d$
3. **RBF**: $K(x, x') = \exp\left(-\frac{|x - x'|^2}{2\sigma^2}\right)$
4. **Sigmoid**: $K(x, x') = \tanh(\kappa x^T x' + \Theta)$

Rubrics- 1 mark for the correct answer

4. **(1 mark)** What is the maximum possible value of the Radial Basis Function (RBF) Kernel? Give justification as well.

a. 0
b. 1
c. infinity
d. -1

Solution) (b) correct option

Explanation: The maximum value that the RBF kernel can be is 1 and occurs when d is 0, when the points are the same, i.e. X = X.

Rubrics- 1 mark for the correct answer and correct justification.

5. **(1 mark)** You are given a labelled binary classification data set with N data points and D features. Suppose that N < D. In training an SVM on this data set, which of the following kernels is likely to be most appropriate?

a. Linear kernel
b. Quadratic kernel
c. Higher-order polynomial kernel
d. RBF kernel

Answer: (a) Linear kernel

Linear Kernel is used when the data is Linearly separable, that is, it can be separated using a single Line. It is one of the most common kernels to be used. It is mostly used when there are a Large number of Features in a particular Data Set.

When the number of examples is less compared to the number of features, you would not have enough data to fit a non-linear SVM i.e. SVM with a non-linear kernel. SVM with a linear kernel (or without a kernel) is one way to go.

Rubrics- 1 mark for the correct answer and correct justification.

6. **(2 marks)** Consider the training data samples and the corresponding Lagrange multipliers learned from them, as given in the following table.

i	x_i	y_i	α_i	i	x_i	y_i	α_i
1	(4,2.9)	1	0.414	6	(1.9,1.9)	-1	0
2	(4,4)	1	0	7	(3.5, 4)	1	0.018
3	(1,2.5)	-1	0	8	(0.5,1.5)	-1	0
4	(2.5,1)	-1	1.18	9	(2,2.1)	-1	0.414
5	(4.9,2.5)	1	0	10	(4.5,2.5)	1	0

From the given table above, answer the following questions:

a. What is the b for the SVM?

b. Identify the support vectors.

Solution: **1. Identifying the Support Vectors**

Support vectors are the training samples that are closest to the decision boundary. These samples have non-zero Lagrange multipliers (a_i>0). From the given data:

Looking at the table, the support vectors are the points where αi>0. From this, we can see that the support vectors are:

- **(4, 2.9)** with a_1=0.414
- **(2.5, 1)** with a_4=1.18
- **(3.5, 4)** with a_7=0.018
- **(2, 2.1)** with a_9=0.414

2. Finding the Bias

To find *b*, we can use the fact that for each support vector *Xi*, the following condition must hold:

$$y_i(w.X_i + b) = 1$$

Where *w* is the weight vector. For a support vector, αi>0, we can compute the weight vector *w* as:

$$w = \sum_i a_i y_i X_i$$

Let's compute *w* using the support vectors:

- Support vectors are the ones with non-zero α, which are at i=1,4,7,9.
- The corresponding y_i values for these support vectors are 1,−1,1,−1, respectively.

Now calculate the weight vector *w*: $w = a_1y_1X_1 + a_4y_4X_4 + a_7y_7X_7 + a_9y_9X_9$

Substituting the values: $w = 0.414 \cdot 1 \cdot (4,2.9)+1.18 \cdot (-1) \cdot (2.5,1)+0.018 \cdot 1 \cdot (3.5,4)+0.414 \cdot (-1) \cdot (2,2.1)$

Now compute this step-by-step:

- $0.414 \cdot (4,2.9)=(1.656,1.201)$
- $1.18 \cdot (-1) \cdot (2.5,1)=(-2.95,-1.18)$
- $0.018 \cdot (3.5,4)=(0.063,0.072)$
- $0.414 \cdot (-1) \cdot (2,2.1)=(-0.828,-0.8694)$

Now add these vectors: $w=(1.656,1.201)+(-2.95,-1.18)+(0.063,0.072)+(-0.828,-0.8694)$

$$w = (-2.059,-0.7764)$$

Now, we can use any support vector to find *b*.

Let's use the first support vector, X_1=(4,2.9), with y_1=1, to find *b*.

The equation for the support vector is: $y_i(w.X_i + b) = 1$

Substitute the values: $1 \cdot ((-2.059) \cdot 4 + (-0.7764) \cdot 2.9 + b) = 1$

Calculate the dot product: $(-2.059) \cdot 4 + (-0.7764) \cdot 2.9 + b = 1$

Now the equation becomes: $-8.236 - 2.25156 + b = 1$

$-10.48756 + b = 1$

Solving for *b*: $b = 1 + 10.4866 = 11.48756$

Final Answers

- **Support vectors**: (4, 2.9), (2.5, 1), (3.5, 4), (2, 2.1)
- **Bias *b***: 11.48756

Rubric- 1 mark for correct part a (The answer should correctly identify the support vectors based on the non-zero Lagrange multipliers $\alpha i > 0$ from the given data) and 1 mark for correct part b (The answer should correctly follow the steps for finding the bias *b*, using the weight vector and support vector equations. The computation should be correct)

7. **(3 marks)** Derive dual form equations for soft margin SVM.

Step 1: Primal Formulation (0.5 Marks)

Given n data points (x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$, we aim to find a hyperplane $w^T x + b = 0$ with maximum margin and allowable misclassification. Slack variables $\xi_i \geq 0$ are introduced for misclassified points.

Primal form:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n$$

where C controls the margin-misclassification trade-off.

Step 2: Lagrangian and Conditions (1.5 Marks)

Using Lagrange multipliers $\alpha_i \geq 0$ (margin) and $\mu_i \geq 0$ (slack), we form the Lagrangian:

$$L(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i (y_i(w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{n} \mu_i \xi_i$$

Minimizing with respect to w:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Minimizing with respect to b:

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0$$

Minimizing with respect to ξ_i:

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow 0 \leq \alpha_i \leq C$$

Step 3: Dual Formulation (1.0 Marks)

Substitute $w = \sum_{i=1}^{n} \alpha_i y_i x_i$ to get the dual problem:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

subject to:

$$\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n$$

Rubrics:

- Primal Formulation (0.5 Marks).
- Lagrangian Setup and Conditions (1.5 Marks): Formulate Lagrangian, minimise w, b, ξi .
- Dual Formulation (1.0 Marks).

OR

Solution

Step 1: Define the Problem and Margins (0.5 Marks)

- Clearly state the optimization objective for maximizing the margin:

$$\max_{\gamma, w, b} \gamma \quad \text{subject to } y^{(i)}(w^T x^{(i)} + b) \geq \gamma \quad \forall i$$

- Define the relationship between the **geometric margin** and the constraints $||w|| = 1$.

Step 2: Reformulate to Eliminate γ (0.5 Marks)

- Reformulate the problem by scaling w and b such that $\gamma = 1$, leading to:

$$\min_{w, b} \frac{1}{2} ||w||^2 \quad \text{subject to } y^{(i)}(w^T x^{(i)} + b) \geq 1 \quad \forall i$$

Step 3: Optimization Problem Setup (0.5 Marks)

- Clearly write down the **primal optimization problem**:

$$\min_{w,b} \frac{1}{2}||w||^2, \quad \text{subject to linear constraints.}$$

Step 4: Lagrangian and Dual Formulation (1 Mark)

- Construct the **Lagrangian**:

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1$$

- Derive conditions for w and b using partial derivatives:

$$\nabla_w L = 0 \implies w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial L}{\partial b} = 0 \implies \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

- Substitute back into the Lagrangian to get the **dual formulation**.

Step 5: Identify Support Vectors and Decision Boundary (0.5 Marks)

- Show how the **decision boundary** depends on support vectors:

$$f(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)} \cdot x) + b\right)$$

- Highlight that only **support vectors** (points with $\alpha_i > 0$) contribute to the classifier.

Rubrics:
Problem Definition and Margins: 0.5 Marks
Reformulation to Eliminate γ: 0.5 Marks
Primal Formulation: 0.5 Marks
Lagrangian and Dual Derivation: 1 Mark
Support Vectors and Decision Boundary: 0.5 Marks