AI & ML Course
Final(Apr 29, 2024)

Time: 120 minutes

Instructions

- Answer all questions
- All answers must be written in the provided spaces. Answers written outside the boxes will not be not graded.
- Last five pages are for rough work. Will not be graded.

Name: ____________________ SRNO:____________________

Room no: __________ Serial Number:__________

Question:	1	2	3	4	5	Total
Points:	9	10	10	10	10	49
Score:						

Read Carefully: In the questions the following notations will be used. (X, Y) be a random instance drawn from a distribution $\mathcal{P}$ where $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$.

$$h : \mathcal{X} \subseteq \mathbb{R}^d \to \mathcal{Y}$$

will denote a classifier. For Binary classification we will use the following notation,

$$\mathcal{Y} = \{-1, 1\}, \eta(x) = P(Y = 1|X = \mathbf{x})$$

The loss function will be denoted by $\ell(h(\mathbf{x}), y)$ where $(\mathbf{x}, y)$ is an instance of observation and label. $\mathcal{D}_n = \{(\mathbf{x}^{(i)}, y_i)|\mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, 1\}, i \in [n]\}$ will denote a dataset of n iid draws from $\mathcal{P}$ $\mathbf{I}$ will denote the identity matrix,dimension will be clear from the context. $N(\mathbf{x}|\mu, C)$ is as defined in the class.

1. (a) (2 points) On a linearly separable dataset of 100 instances it was observed that an SVM classifier gave a LOO error of 4. The number of support vectors
A. is less than 4 **B. is more than** 4 **C. is equal to** 4 D. has no relationship with LOO error.

   (b) (7 points) Consider dataset D_n. Pose the following problem as an Quadratic Optimization problem (Quadratic objective with linear constraints). At optimality state the value of $\mathbf{w}$ in terms of the Data.

$$min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} max(0, 1 - y_i(\mathbf{w}^\top \mathbf{x}^{(i)}))$$

Solution: We use the following observation, for any $a, b \in \mathbb{R}$

$$max(a, b) = min_t t, \quad \text{s.t. } t \geq a,\ t \geq b$$

Using the observation, introduce new variables ξ_i to obtain

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

$$\xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}^{(i)}))$$

$$\xi_i \geq 0$$

It is a Quadratic program.

2. University **UNI** has devised a score to determine whether students are addicted to Internet. If they are addicted there will be a rehab program. The class $Y = 1$ corresponds to persons who have Internet addiction. It is also noted that adminstering the program to non-addicted students maybe counterproductive. You have modelled this problem as a Classification problem with the following loss function

$$\ell(1, 1) = \ell(-1, -1) = 0, \ell(1, -1) = c, \ell(-1, 1) = 1.$$

Let the score be x and it is given that $\eta(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$.

We are interested in classifiers of the form

$$f(x) = sign(x - t), \quad t \text{ is the threshold}$$

(a) (2 points) Define the risk of the classifier for a given value of t?

Solution:

$$R(f) = E_{(x,y)\sim P}\ell(f(x), y)$$

(b) (3 points) Compute t so that the classifier has minimum misclassification risk?

Solution: It is attained by the Bayes classifier

$$f_B(x) = sign(2\eta(x) - 1)$$

Thus the decision boundary is $\eta(x) = \frac{1}{2}$. Now $\eta(t) = \frac{1}{2}$ which yields $t = \frac{1}{2}$.

(c) (4 points) Compute t so that the classifier achieves the minimum risk?

Solution: The minimum is attained by the following classifier

$$f(x) = 1, \quad E_{Y|X}\ell(1, Y) < E_{Y|X}\ell(-1, Y)$$

$$f(x) = -1, \quad E_{Y|X}\ell(1, Y) > E_{Y|X}\ell(-1, Y)$$

$$E_{Y|X}\ell(1, Y) < E_{Y|X}\ell(-1, Y) \implies c(1 - \eta(x)) < \eta(x)$$

hence t satisfies

$$\eta(t) = c(1 - \eta(t)), \implies ; \eta(t) = \frac{c}{1 + c}$$

This yields $t = \frac{c}{1+c}$

(d) (1 point) Comment on the effect of c on your results.

Solution: If $c = 1$ we recover the Bayes classifier. If c is very large it implies that there is a large cost for mis-classifying an healthy person. Hence the classifier would prefer labelling *most examples* as -1, and hence the value of t will be high. Similarly if c is small the classifier would prefer labelling most examples as 1 and the value of t should be low.

3. Kernel functions

(a) (2 points) Define a Kernel function?

Solution: State the symmetry and Positive definite property.

(b) Using the definition answer the following

i. (4 points) For any normalized kernel function, $k(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x}$, is the following true

$$k^2(\mathbf{x}, \mathbf{z}) \leq 1$$

Justify.

Solution: True. Due to positive semidefinite property, for any pair of examples $(\mathbf{x}, \mathbf{z})$ the matrix

$$K_{11} = k(\mathbf{x}, \mathbf{x}) = 1, K_{22} = k(\mathbf{z}, \mathbf{z}) = 1, K_{12} = k(\mathbf{x}, \mathbf{z}) = K_{21}$$

is positive semidefinite. This implies the determinant of the matrix must be greater than 0. This implies that

$$K_{11}K_{22} - K_{12}^2 \geq 0, \implies 1 \geq k^2(\mathbf{x}, \mathbf{z})$$

ii. (4 points) Prove or Disprove using the definition stated in Q3a. If k is a kernel function $g(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z})k(\mathbf{x}, \mathbf{z})$ is a Kernel function using the definition stated above.

Solution: Prove that $\mathbf{x}^\top \mathbf{z}$ is a Kernel function. Check the proof that the product of two Kernels is a kernel.

4. Let $X_1, \ldots, X_n \overset{IID}{\sim} U(0, a)$. Consider the estimator $\hat{a} = max_i X_i$.

(a) (2 points) The bias of the estimator is $E(\hat{a}) - a$. Is it positive or negative? Give reasons

Solution: It is negative because all $X_i < a$ for all i.

(b) (3 points) What is the cdf of $\hat{a}$?

Solution:

$$P(\hat{a} \leq t) = P(X_1 \leq t, \ldots, X_n \leq t) = \Pi_{i=1}^{n} P(X_i \leq t) = \begin{cases} 0 & t < 0 \\ \left(\frac{t}{a}\right)^n & 0 \leq t \leq a \\ 1 & t \geq a \end{cases}$$

(c) (5 points) Construct $\tilde{a}$, an unbiased estimator of a, from $\hat{a}$

Solution: The density of $\hat{a}$ is

$$f(t) = \begin{cases} n\left(\frac{t}{a}\right)^{n-1} & 0 \leq t \leq a \\ 0 & \text{otherwise} \end{cases}$$

Hence $E(\hat{a}) = \int_{-\infty}^{\infty} f(t)tdt = \frac{n}{n+1}a$. Thus $\tilde{a} = \frac{n+1}{n}\hat{a}$ is an unbiased estimator.

5. Consider learning a mixture of Gaussian distribution of the following form.

$$P(X = x|\Theta) = \sum_{i=1}^{k} \alpha_i N(\mathbf{x}|\mu_i, C_i)$$

$$\Theta = \{(\alpha_i, \mu_i, C_i)|i \in [k]\}$$

$$\alpha_i > 0, \sum_{i=1}^{k} \alpha_i = 1, 0 < \alpha_i < 1, \mu_i \in \mathbb{R}^d, C_i \in \mathbb{R}^{d\times d}$$

From Dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}_{(N)}\}$ we wish to learn the parameters Θ through the EM algorithm. Let $\bar{\Theta}$ be the current estimate of the parameters and the probabilities

$$q_i^{(j)} = \frac{\bar{\alpha}_i N(\mathbf{x}|\bar{\mu}_i, \bar{C}_i)}{\sum_{l=1}^{k} \bar{\alpha}_l N(\mathbf{x}|\bar{\mu}_l, \bar{C}_l)}$$

(a) (2 points) The EM algorithm computes the new estimate of the parameters by solving the following problem

$$\Theta^{(new)} = argmax_{\Theta} Q(\Theta; \bar{\Theta})$$

State $Q(\Theta|\bar{\Theta})$ in terms of $q_i^{(j)}$, Data and the parameters.

Solution:

$$Q(\Theta, \bar{\Theta}) = \sum_{i=1}^{k}\sum_{j=1}^{N} q_i^j \log \alpha_i + \sum_{j=1}^{N}\left(\sum_{i=1}^{k}\left(q_i^{(j)}\left(-\frac{1}{2}(\mathbf{x}^{(j)} - \mu_i)^\top C_i^{(-1)}(\mathbf{x}^{(j)} - \mu_i) - \frac{1}{2}log\,(2\pi)^d\,|C_i|\right)\right)\right)$$

(b) (3 points) Solving the above problem find $\mu_i^{(new)}$.

Solution: Finding μ_i reduces to solving

$$min_{\mu_i} \frac{1}{2} \sum_{j=1}^{N} \left(q_i^{(j)} (\mathbf{x}^{(j)} - \mu_i)^\top C_i^{(-1)} (\mathbf{x}^{(j)} - \mu_i) \right.$$

The minimizer is

$$\sum_{j=1}^{N} \left(q_i^{(j)} (\mathbf{x}^{(j)} - \mu_i) \right) = 0$$

$$\mu_i = \frac{1}{N_i} \sum_{j=1}^{N} \left(q_i^{(j)} \mathbf{x}^{(j)} \right), \quad N_i = \sum_{j=1}^{N} q_i^{(j)}$$

(c) (5 points) Suppose it was given that $C_i = \sigma_i^2 \mathbf{I}$. How will your answer change in the above two questions. Find σ_i^{new}?

Solution:

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded

Space for Rough Work: Not to be graded