AI & ML Course
MidSem 2(Mar 21, 2024)

Time: 70 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	Total
Points:	5	5	10	10	30
Score:					

Read Carefully: In the questions the following notations will be used.' Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y_i)|y_i \in \mathcal{Y}, \mathbf{x}^{(i)} \in \mathbb{R}^d, i \in [n]\}$ denote a training dataset of n observations. Soft margin SVM classifier, $f(\mathbf{x}) =$ sign $(\mathbf{w}^\top\mathbf{x}+b)$, on $\mathcal{D}$ with $\mathcal{Y} = \{1, -1\}$ is obtained by solving

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

$$\text{subject to } y_i(\mathbf{w}^\top\mathbf{x}^{(i)} \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i \in [n]$$

where $\mathbf{w} = \sum_{i=1}^{n} \lambda_i y_i \mathbf{x}^{(i)}$.
The SVM regression problem defined on dataset $\mathcal{D}$ with $\mathcal{Y} = \mathbb{R}$.

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

$$\text{subject to } y_i - (\mathbf{w}^\top\mathbf{x}^{(i)} + b) \leq \epsilon + \xi_i,$$
$$\mathbf{w}^\top\mathbf{x}^{(i)} + b - y_i \leq \epsilon + \xi_i^*,$$
$$\xi_i, \xi_i^* \geq 0, \quad i \in [n]$$

At optimality $\mathbf{w} = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\mathbf{x}^{(i)}$.

1. The soft-margin SVM optimization problem was solved for some $C > 0$ and the following was observed.

$$\frac{100}{n}|\{i|\xi_i = 0\}| = 80, \ \frac{100}{n}|\{i|0 < \xi_i \leq 1\}| = 5, \ \frac{100}{n}|\{i|1 < \xi_i \leq 2\}| = 10, \ \frac{100}{n}|\{i|2 < \xi_i \leq 3\}| = 5$$

   Based on the information provided answer the following questions related to $f(\mathbf{x})$, the associated classifier.

   (a) (1 point) $\frac{100}{n}|\{i|f(\mathbf{x}^{(i)}) = y_i\}| =$ _____85_____
   (b) (1 point) $\frac{100}{n}|\{i|\mathbf{x}^{(i)}$ is a support vector$\}| \geq$ _____20_____
   (points will be given for the best bound)
   (c) (1 point) $\frac{100}{n}|\{i|f(\mathbf{x}^{(i)}) = y_i, \lambda_i = C\}| =$ _____5_____
   (d) (1 point) $\frac{100}{n}|\{i|f(\mathbf{x}^{(i)}) \neq y_i, \lambda_i = C\}| =$ _____15_____
   (e) (1 point) Answer True or false. For a different choice of C the SVM problem was solved and it was found that $\sum_{i=1}^{n} \xi = 0.5n$. The chosen value of C was higher than the original value __**F**__.

2. (5 points) Consider SVM regression problem for $C = 1, \epsilon = 0.1$. Following was found for $i = 1$ and $i = 2$

$$y_1 - \mathbf{w}^\top\mathbf{x}^{(1)} - b = -0.05, y_2 - \mathbf{w}^\top\mathbf{x}^{(2)} - b = 1$$

   Find $\alpha_1 =$ _____0_____, $\alpha_1^* =$ _____**0**_____, $\alpha_2 =$ _____1_____, $\alpha_2^* =$ _____0_____.

3. The ridge regression problem on $\mathcal{D}$ is given by

$$\mathbf{w}_{RR} = \texttt{argmin}_{\mathbf{w}} \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top\mathbf{x}^{(i)})^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

   We will solve the Ridge regression problem by using the following parametrization of

$$\mathbf{w} = \sum_{i=1}^{n}\beta_i\mathbf{x}^{(i)} + \mathbf{v}, \mathbf{v}^\top\mathbf{x}^{(i)} = 0, \forall i \in [n]$$

   Let $X^\top = [\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}], \quad Y = [y_1, \ldots, y_n]^\top, \beta = [\beta_1, \ldots, \beta_n]^\top$. Assume n is more than d. Express the answers to the following question in terms of $X, Y, \beta, \mathbf{v}$. The answers need to be simplified as much as possible.

(a) (3 points) Restate the Ridge regression optimization problem using the parametrization. The problem statement should not involve $\mathbf{w}$.

Solution: By direct substitution.

$$\min_{\beta,\mathbf{v}} \frac{1}{2}\|Y - XX^\top\beta\|^2 + \frac{\lambda}{2}\left(\|X^\top\beta\|^2 + \|\mathbf{v}\|^2\right)$$

Since this is minimum at $\mathbf{v} = 0$ the new problem is

$$\min_{\beta} \frac{1}{2}\|Y - XX^\top\beta\|^2 + \frac{\lambda}{2}\|X^\top\beta\|^2$$

(b) (5 points) State the optimality conditions of the problem and from them find β^*, the optimal solution.

Solution: The stated problem is convex. Hence setting the gradient to zero is sufficient for optimality

$$XX^\top(XX^\top\beta - Y) + \lambda XX^\top\beta = 0$$
$$XX^\top\left(XX^\top + \lambda I)\beta - Y\right) = 0$$
$$\beta^* = (XX^\top + \lambda I)^{-1}\top Y$$

is the optimal solution.

(c) (2 points) Compare the computational complexity of the problem with the standard Ridge regression problem.

Solution: The standard Ridge regression problem has complexity $O(d^3)$ but this has $O(n^3)$.

4. Let $X_1, X_2, \ldots, X_n \overset{\texttt{IID}}{\sim} \mathcal{P}$ be a sample of continuous random variable. It is further given that $E(X) = \mu, Var(X) = \sigma^2, X \sim \mathcal{P}$. Consider two estimators T_1 and T_2 where

$$T_1(n) = X_{n-1}, \quad T_2(n) = \frac{1}{n+1}\sum_{i=1}^{n} X_i$$

(a) (2 points) Recall that the MSE of estimator $T(n)$ of θ is $MSE(T(n)) = E(T(n) - \theta)^2$ where n is the sample size. Find an upperbound on $P(|T(n) - \theta| \geq \epsilon)$ in terms of $MSE(T(n))$. The bias of $T(n)$ is not known.

Solution: Let $Y = T(n) - \theta$. Apply Markov's inequality on Y^2 to obtain

$$P(|Y| \geq \epsilon) = P(|Y|^2 \geq \epsilon^2) \leq \frac{E(Y^2)}{\epsilon^2} = \frac{MSE(T(n))}{\epsilon^2}$$

(b) (2 points) Determine if $T_1(n)$ and $T_2(n)$ are unbiased?

Solution: Note that $E(X_i) = \mu$. By computation $E(T_1(n)) = E(X_{n-1}) = \mu$ and hence it is unbiased. Similarly $E(T_2(n)) = E(\frac{1}{n+1}\sum_{i=1}^{n} X_i) = \frac{n}{n+1}\mu$ and hence it is biased.

(c) (2 points) Compute the variance of $T_1(n)$ and $T_2(n)$?

Solution: Note that $Var(X_i) = \sigma^2$
Hence $Var(T_1(n)) = \sigma^2$ and

$$Var(T_2(n)) = \frac{1}{(n+1)^2}\sum_{i=1}^{n} Var(X_i) = \frac{n}{(n+1)^2}\sigma^2.$$

(d) (4 points) Are $T_1(n)$ and $T_2(n)$ asymptotically consistent?

Solution:
Let F be the cdf of X_i. For $T_1(n)$, we observe that

$$P(|T_1(n) - \mu| < \epsilon) = P(-\epsilon < T_1(n) - \mu < \epsilon) = F(\mu + \epsilon) - F(\mu - \epsilon)$$

Since X_i are continuous there exists ϵ for which the RHS is not zerobut less than 1. For such a choice of ϵ

$$P(|T_1(n) - \mu| > \epsilon) = 1 - P(-\epsilon < T_1(n) - \mu < \epsilon) = 1 - F(\mu + \epsilon) + F(\mu - \epsilon)$$

the RHS is not zero. Hence it is not consistent.
For $T_2(n)$ we use the MSE to check for consistency. Observe that $Bias(T_2(n)) = -\frac{1}{n+1}\mu$ and hence

$$P(|T_2(n) - \mu| > \epsilon) \leq \frac{MSE(T_2(n))}{\epsilon^2} = \frac{1}{(n+1)}\frac{\frac{1}{n+1}\mu^2 + \frac{n}{n+1}\sigma^2}{\epsilon^2}$$

Clearly then as $n \lim \infty$ the RHS goes to zero.
Note: An unbiased estimator mayn't be consistent and a biased estimator maybe consistent

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