AI & ML Course
MidSem 1(Feb 17, 2024)

Time: 90 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 three pages are for rough work. Will not be graded.

Name: _____ SRNO:

Question:	1	2	3	Total
Points:	5	10	15	30
Score:				

Read Carefully: In the following we will use the following notations. (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. $\mathcal{D}_n = \{(\mathbf{x}^{(i)}, y_i) | \mathbf{x}_i \in \mathcal{X}, y_i \in \{-1, 1\}, i \in [n]\}$ will denote a dataset of n iid draws from $\mathcal{P}$ For multicategory classification,

$$\mathcal{Y} = \{1, \ldots, C\}, P_i(x) = P(X = x | Y = i), p_i = P(Y = i), i \in \{1, \ldots, K\}.$$

1. Answer all questions
   (a) Answer True or False
      i. (1 point) Consider multi-category classification with C classes. Let $p_i = \frac{1}{C}$. Consider the classifier
      $$h(\mathbf{x}) = \text{argmax}_{i \in [C]} log P_i(\mathbf{x})$$
      This classifier does not achieve the Bayes error-rate under the $0 - 1$ loss. **F**
      ii. (1 point) Naive Bayes Classifiers are always linear. **F**

   (b) (1 point) On a dataset $\mathcal{D}_n$ the following classifier is obtained after solving an SVM problem. It is given that the $n = 10$ and number of observations from each class are roughly equal.
   $$h(\mathbf{x}) = sign\left((\mathbf{x}^{(1)} - \mathbf{x}^{(3)})^\top \mathbf{x}\right)$$
   The value of $h(\mathbf{x}^{(3)}) =$ _____−1_____.

   (c) (2 points) Let $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ be defined over the dataset $\mathcal{D}_n$. It is found that (a.) $K_{12} = -1$ and (b.) $\sum_{i,j=1}^{n} K_{ij} = -5$. From this info. it was deduced that k is not a valid kernel function. This is true because for a valid kernel function
      A. (a) and (b) are both false
      B. (a) can be true but (b) is false
      C. (a) is false but (b) can be true

2. Let $\mathcal{D}_n$ be such that there exist $\mathbf{w}^* \in \mathbb{R}^d, \|\mathbf{w}^*\| = 1$ and $\gamma > 0$ which satisfies

$$\gamma = min_{i\in[n]} \quad y_i((\mathbf{x}^{(i)})^\top \mathbf{w}^*)$$

You have implemented the Perceptron Algorithm with initial weight vector $\mathbf{w}^{(0))} = \mathbf{v}, \quad \mathbf{v}^\top \mathbf{w}^* = 0$. Let $\mathbf{w}^{(k)}$ denote the weight vector after k updates. Assume that $\|\mathbf{x}\| = 1, \mathbf{x} \in \mathcal{D}_n$.

(a) (3 points) Find smallest value of $t \in \mathbb{R}$ such that

$$\|\mathbf{w}^{(k+1)} - t\mathbf{w}^*\|^2 \leq \|\mathbf{w}^{(k)} - t\mathbf{w}^*\|^2 - 1$$

for all k.

Solution: For $k+1$th update there must exist $i \in [n]$ such that

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + y_i\mathbf{x}^{(i)}, \qquad y_i\mathbf{w}^{(k)\top}\mathbf{x}^{(i)} < 0$$

Thus

$$\|\mathbf{w}^{(k+1)} - t\mathbf{w}^*\|^2 = \|\mathbf{w}^{(k)} - t\mathbf{w}^*\|^2 + 2y_i(\mathbf{w}^{(k)} - t\mathbf{w}^*)^\top\mathbf{x}^{(i)} + \|\mathbf{x}^{(i)}\|^2$$

$$\|\mathbf{w}^{(k+1)} - t\mathbf{w}^*\|^2 \leq \|\mathbf{w}^{(k)} - t\mathbf{w}^*\|^2 - 2ty_i\mathbf{w}^{*\top}\mathbf{x}^{(i)} + 1$$

$$\|\mathbf{w}^{(k+1)} - t\mathbf{w}^*\|^2 \leq \|\mathbf{w}^{(k)} - t\mathbf{w}^*\|^2 - 2t\gamma + 1$$

The desired inequality holds as soon as $-2t\gamma + 1 \leq -1$ and hence $t = \frac{1}{\gamma}$ is the smallest value.

(b) (4 points) Is there any upper bound on T, the maximum number of updates, such that the Perceptron Algorithm will converge for your implementation. If so find it. If not give reasons.

Solution: Using the above inequality we conclude that after T updates

$$\|\mathbf{w}^{(T)} - t\mathbf{w}^*\|^2 \leq \|\mathbf{w}^{(0)} - t\mathbf{w}^*\|^2 - T$$

Thus

$$T \leq \|\mathbf{w}^{(0)} - t\mathbf{w}^*\|^2 = \|\mathbf{v}\|^2 + \frac{1}{\gamma^2}$$

(c) (3 points) What would be the best choice of $\mathbf{v}$? Give reasons

Solution: From the above expression $\mathbf{v} = 0$ yields the smallest bound on T.

3. It is known that for a given $a^{(i)} \in \mathbb{R}^d, i \in [n]$ there does not exist any $\mathbf{x} \in \mathbb{R}^d$ such that

$$a^{(i)\top}\mathbf{x} \geq 1$$

is true for all $i \in [n]$. Though there exists $\mathbf{x}$ which can satisfy some of the constraints.

(a) (1 point) Consider the following optimization problem

$$min_{\mathbf{x}\in\mathbb{R}^d,\mathbf{y}\in\mathbb{R}^n} \quad \frac{1}{2}\|\mathbf{x}\|^2$$
$$\text{subject to } a^{(i)\top}\mathbf{x} \geq 1 - y_i, \quad y_i \geq 0, i \in [n]$$

Find the optimal point and optimal value.

Solution: Note that $\mathbf{x} = 0$ yields the minimum possible value of the objective and it is also feasible as soon as we choose $y_i \geq 1$. Thus the optimal value is 0 and it is attained at $\mathbf{x} = 0$ and $y_i \geq 1, i \in [n]$.

(b) Consider the following optimization problem

$$min_{\mathbf{x}\in\mathbb{R}^d,\mathbf{y}\in\mathbb{R}^n} \quad \frac{1}{2}\|\mathbf{x}\|^2 + C\sum_{i=1}^{n} y_i$$
$$\text{subject to } a^{(i)\top}\mathbf{x} \geq 1 - y_i, \quad y_i \geq 0, i \in [n]$$

Let the Lagrangian of the problem be

$$\mathcal{L}(\mathbf{x}, \mathbf{y}, \lambda^{(1)}, \lambda^{(2)}) = \frac{1}{2}\|\mathbf{x}\|^2 + C\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \lambda_i^{(1)} \left(a^{(i)\top}\mathbf{x} - 1 + y_i\right) - \sum_{i=1}^{n} \lambda_i^{(2)} y_i, i \in [n]$$

i. (1 point) Find $\mathbf{x}, \mathbf{y}$ which satisfy the constraints? If not possible, give reasons.

Solution:

$$\mathbf{x} = 0, y_i = 1, i \in [n]$$

ii. (2 points) Let $\mathbf{x}^*, \mathbf{y}^*, \lambda^{(1*)}, \lambda^{(2*)}$ be the KKT point for the problem. Suppose $\mathbf{a}^{(1)\top}\mathbf{x}^* = 4$. Then

$$\lambda_1^{(1*)} = \underline{\quad \mathbf{0} \quad}, \quad \lambda_1^{(2*)} = \underline{\quad \mathbf{C} \quad}, \quad y_1^* = \underline{\quad \mathbf{0} \quad}$$

Justify.

Solution: One of the KKT condition is

$$\lambda_i^{(1*)} + \lambda_i^{(2*)} = C \forall i \in [n]$$

The inequality

$$a^{(1)\top}\mathbf{x} - 1 + y_1 \geq 0 \implies 3 + y_1 \geq 0$$

is strictly positive for any $y_1 \geq 0$. Thus $\lambda_1^{(1*)} = 0$ and hence by the KKT condition $\lambda_1^{(2*)} = C$. This implies $y_1 = 0$ by other KKT conditions.

iii. (2 points) Let $\mathbf{x}^*, \mathbf{y}^*, \lambda^{(1*)}, \lambda^{(2*)}$ be the KKT point for the problem. Suppose $\mathbf{a}^{(2)\top}\mathbf{x}^* = -1$. Then

$$\lambda_2^{(1*)} = \underline{\quad \mathbf{C} \quad}, \quad \lambda_2^{(2*)} = \underline{\quad \mathbf{0} \quad}, \quad y_2^* = \underline{\quad \mathbf{2} \quad}$$

Justify.

Solution: Similar to question above.

$$a^{(2)\top}\mathbf{x} - 1 + y_2 \geq 0$$

Thus $-2 + y_2 \geq 0$ implies y_2 is positive. By KKT condition $y_2 = 2$. This implies $\lambda_2^{(2*)} = 0$ and hence $\lambda_2^{(1*)} = C$

iv. (9 points) The Wolfe Dual of the above problem can be written as

$$max_{\alpha \in \mathbb{R}^n} \alpha^\top \mathbf{e} - \frac{1}{2}\alpha^\top Q \alpha$$

subject to $0 \leq \alpha \leq C$.
$\mathbf{e} = [1, \ldots, 1]^\top$ and Q is a $n \times n$ matrix.

$$Q_{ij} = \underline{\quad \mathbf{a}^{(i)\top}\mathbf{a}^{(j)} \quad}$$

Express α_i in terms of $\lambda^{(1)}, \lambda^{(2)}$

Solution:

$$\alpha_i = \lambda_i^{(1)}$$

Justify:

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