CS5691: Pattern recognition and machine learning
Mid-term exam
Course Instructor : Prashanth L. A.
Date : Mar-8, 2019 **Duration** : 100 minutes **Marks**: 20

Name of the student :
Roll No :

INSTRUCTIONS: For short answer questions, you do not have to justify the answer. For the rest, provide proper justification for the answers. Please use rough sheets for any calculations *if necessary.* Please **DO NOT** submit the rough sheets. DO NOT use pencil for writing the answers.

I. Short Answer Questions

Note: 1 mark for the correct answer. For MCQs only one answer is correct. Please write the choice code a, b, c or d in the answer box provided. For other questions fill in the blank with the appropriate number or expression or a yes/no.

1. Let $(\mathbf{x}_1, y_1, z_1), \ldots, (\mathbf{x}_n, y_n, z_n)$ be a set of data points such that $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i, z_i \in \mathbb{R}$. Let $y_i + 2z_i = 3$ for $i = 1, \ldots, n$. Let A be a $(n \times d)$ matrix with rows $\mathbf{x}_i^\mathsf{T}$. Let

$$\widehat{\mathbf{u}}_{\mathrm{ML}} = \mathrm{argmin}_{\mathbf{u} \in \mathbb{R}^d} \sum_i (\mathbf{u}^\mathsf{T}\mathbf{x}_i - y_i)^2, \quad \widehat{\mathbf{v}}_{\mathrm{ML}} = \mathrm{argmin}_{\mathbf{v} \in \mathbb{R}^d} \sum_i (\mathbf{v}^\mathsf{T}\mathbf{x}_i - z_i)^2$$

   Give an expression relating $\widehat{\mathbf{u}}_{\mathrm{ML}}$ and $\widehat{\mathbf{v}}_{\mathrm{ML}}$.

   Answer:

2. Let $a_1, a_2, \ldots, a_n$ be the importances of the data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. Consider the weighted least squares regression problem, with the following objective:

$$R(\mathbf{w}) = \sum_{i=1}^{n} a_i (\mathbf{w}^\top \mathbf{x}_i - y_i)^2.$$

   Give an expression for the minimiser of $R(\mathbf{w})$.

   Answer:

3. Suppose we have the following four points $x_1 = (1, 1), x_2 = (-1, 3), x_3 = (2, 4)$, and $(y_1, y_2, y_3) = (5, 11, 18)$. Then, $\min_w \sum_{i=1}^{3} (x_i^\mathsf{T} w - y_i)^2$ is

   (a) $\in (0, 10)$.
   (b) > 10.
   (c) $= 0$.
   (d) < 0.

   Answer: ☐

4. Consider a dataset for classification $\{(X_i, y_i), i = 1, \ldots, n\}$, with $y_i \in \{-1, +1\}$, formed using n i.i.d. samples, with equi-probable classes, and with univariate Gaussian class conditional densities. The means for the latter are 10 and -1, corresponding to class labels -1 and $+1$, respectively, while the variances are equal. Suppose that the perceptron algorithm is run on this dataset. Then, on any such dataset of n samples, is the perceptron algorithm guaranteed to converge? Provide a yes or no for the answer.

   Answer:

5. Consider a dataset with the following four data points: $(0, 0), (0, 1), (1, 0), (1, 1)$, with corresponding class labels $1, -1, -1, 1$, respectively. The dataset is clearly(?) not linearly separable. Consider adding another co-ordinate to each data point. Which of the following schemes will ensure that the resulting dataset in three dimensions is linearly separable?

   (a) Third co-ordinate value is equal to first one for each data point.

   (b) Third co-ordinate value is 1 for one of the data points, and 0 for the rest three of them.

   (c) Third co-ordinate value is the negative of the second value for each data point.

   (d) None of the above.

   Answer: ☐

6. Consider a dataset of n points $x_1, \ldots, x_n$, where x_i is drawn from a Gaussian distribution with mean μ, and variance $\sigma_i^2 > 0$, for $i = 1, \ldots, n$. What is the ML estimate for μ, when the variances $\sigma_1^2, \ldots, \sigma_n^2$ are known?

   Answer:

7. Given a dataset $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, $\forall i$. Consider the ridge regression solution $\widehat{W}(\lambda) = CY$, where $C = (A^\intercal A + \lambda I)^{-1} A^\intercal$, and A is a $(n \times d)$ matrix with rows $\mathbf{x}_i^\intercal$. Is C a projection matrix?

   Answer:

8. Specify a conjugate prior when the likelihood is an exponential distribution with parameter $\theta > 0$.

   Answer:

9. Consider a classification dataset, with two-dimensional inputs $(-1, 1), (1, 3), (-3, 3)$ having class label "-1", and input data points $(0, 1), (2, 2), (3, 1)$ having class label "1". Let $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ denote the inputs with class label -1, and $\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6$ denote the inputs with class label 1.

   Answer the following: (1 mark each)

   (a) Find a vector W^* such that $W^\intercal \mathbf{x}_i > 0$, for $i = 1, \ldots, 6$.

   Answer:

(b) Suppose the perceptron algorithm is run on this dataset. Using $\|W^*\|$, $M = \max_{i=1,\ldots,6} \|\mathbf{x}_i\|^2$, and $\beta = \min_{i=1,\ldots,6} \mathbf{x}_i^\intercal W^*$, provide an upper bound on the number of times the iterate, say w_k, of the perceptron algorithm is updated, before the stopping condition is reached (i.e., an iterate w_k that correctly classifies all the input data points).

Answer:

II. Problems that require a detailed solution

1. Consider a two class two-dimensional problem, where the class conditional densities are Gaussian with means μ_0 and μ_1. Assume equi-probable classes.

   Answer the following: (2+2+1 marks)

   (a) Suppose that the covariance matrix for each class is $\sigma^2 I$, for some $\sigma^2 > 0$. Consider the following classifier:

   $$h_1(x) = \begin{cases} 0 & \text{if } \|x - \mu_0\| > \|x - \mu_1\|, \\ 1 & \text{otherwise.} \end{cases}$$

   Is h_1 optimal for the zero-one loss function? Justify your answer.

   (b) Suppose that the covariance matrix is $\begin{bmatrix} a & b \\ b & c \end{bmatrix}$, for some positive constants a, b, c. Then, is h_1 optimal for the classification problem, with rest of the parameters as in the part above?

   (c) Let $\mu_1 = [0, 0]^\intercal$, $\mu_2 = [3, 3]^\intercal$, and the covariance matrix entries are given by $a = 1.1, b = 0.3, c = 1.9$. Classify the input vector $\tilde{x} = [1.0, 2.2]^\intercal$, and compare with the prediction $h_1(\tilde{x})$.

2. Suppose that the target variable y is given by $y = W^\intercal X + \epsilon$, where $X \in \mathbb{R}^d$ is the input vector, W is the unknown parameter, and ϵ is a zero-mean Gaussian random variable with precision (inverse variance) β. Given a dataset $\{(X_i, y_i), i = 1, \ldots, n\}$, let $\widehat{W}(\lambda)$ denote the estimate of W obtained using regularized least squares, i.e.,

$$\widehat{W}(\lambda) = \min_{\overline{W}} \frac{1}{2} \sum_{i=1}^{n} (y_i - X_i^\intercal \overline{W})^2 + \frac{\lambda}{2} \overline{W}^\intercal \overline{W}.$$

Answer the following: (2+3 marks)

(a) Is $\mathbb{E}\left(\widehat{W}(\lambda)\right) = W$ for $\lambda > 0$?

(b) Calculate the variance of $\widehat{W}(\lambda)$ defined by

$$\mathrm{Var}(\widehat{W}(\lambda)) = \mathbb{E}\left[\left(\widehat{W}(\lambda) - \mathbb{E}(\widehat{W}(\lambda))\right)\left(\widehat{W}(\lambda) - \mathbb{E}(\widehat{W}(\lambda))\right)^\intercal\right].$$

Hint: Use the fact that $\mathrm{Var}(CY) = C\mathrm{Var}(Y)C^\intercal$, when C is not random.

(c) BONUS (2 marks): Show that the variance of $\widehat{W}(\lambda)$ is smaller than $\widehat{W}(0)$, i.e., $\mathrm{Var}(\widehat{W}(0)) - \widehat{W}(\lambda)$ positive semi-definite.