CS5691: Pattern recognition and machine learning
Mid-term exam - Solutions
Course Instructor : Prashanth L. A.

I. Short Answer Questions

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 $\mathbf{x}_i^\intercal$. Let

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

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

Answer: Let b be a n-vector with each entry 3, A be a $(n \times d)$ matrix with rows $\mathbf{x}_i^\intercal$. Then, $A\left(\widehat{\mathbf{u}}_{\text{ML}} + 2\widehat{\mathbf{v}}_{\text{ML}}\right)^\intercal = b$.

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: Let C be a diagonal matrix with entries a_i, A be a $n \times d$ matrix with rows $\mathbf{x}_i^\intercal$, and Y be a n-dimensional vector with entries y_i. Then, the minimiser $\mathbf{w}^*$ of $R(\mathbf{w})$ is given by

$$\mathbf{w}^* = \left(A^\intercal C A\right)^{-1} A^\intercal C Y.$$

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, $\min_w \sum_{i=1}^{3} (x_i^\intercal w - y_i)^2$ is

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

   Answer: (c)

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: No.

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: (b)

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: $\hat{\mu}_{\mathrm{ML}} = \left(\sum_{i=1}^n \frac{1}{\sigma_i^2}\right)^{-1} \sum_{i=1}^n \frac{x_i}{\sigma_i^2}$.

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 $C = (A^\top A + \lambda I)^{-1} A^\top$, and A is a $(n \times d)$ matrix with rows $\mathbf{x}_i^\top$. Is C a projection matrix?

   Answer: No.

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

   Answer: Gamma(α, β).

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 $W^\top \mathbf{x}_i > 0$, for $i = 1,\ldots,6$.

   Answer: $W^* = (-1,4)$.

   (b) Suppose the perceptron algorithm is run on this dataset. Using $\|W^*\|$, $M = \max_{i=1,\ldots,6} \|\mathbf{x}_i\|^2$, and $\beta = \min_{i=1,\ldots,6} \mathbf{x}_i^\top W^*$, 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: The required bound is $\frac{\|W^*\|^2 M}{\beta^2} = \frac{17 \times 18}{1} = 306$.

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 $h_1(\tilde{x})$.

   Answer:

   (a) Yes, because it obeys Bayesian classification rule, and it says, if $q_0 > q_1$ predict 0 else predict 1, i.e.,

   $$\frac{1}{(2\pi)^{n/2}\sigma} exp\left(\frac{-(x-\mu_0)^\intercal (x-\mu_0)}{2\sigma^2}\right) > \frac{1}{(2\pi)^{n/2}\sigma} exp\left(\frac{-(x-\mu_1)^\intercal (x-\mu_1)}{2\sigma^2}\right)$$
   $$\implies -\|x-\mu_0\|^2 > -\|x-\mu_1\|^2$$
   $$\implies \|x-\mu_0\| < -\|x-\mu_1\|$$

   (b) No, h_1 is not optimal.

   $$q_i(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}} exp\left[\frac{-1}{2}(x-\mu_i)^\intercal \left(\frac{1}{(ac-b^2)}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}\right)(x-\mu_i)\right]$$
   $$\implies h_2(x) = 0 \quad \text{if } (x-\mu_0)^\intercal \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}(x-\mu_0) < (x-\mu_1)^\intercal \begin{bmatrix} c & -b \\ -b & a \end{bmatrix}(x-\mu_1)$$
   $$h_2(x) = 1 \quad \text{otherwise.}$$

   $h_2(x)$ is the optimal classifier. Now, $h_2(x) = h_1(x)$ if $a = c$ and $b = 0$, otherwise $h_1 \neq h_2$ and thus h_1 is not optimal.

(c) For $h_2(\tilde{x})$,

$$(\tilde{x}-\mu_0)^{\top}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}(\tilde{x}-\mu_0) = \begin{bmatrix} 1 & 2.2 \end{bmatrix}\begin{bmatrix} 1.9 & -0.3 \\ -0.3 & 1.1 \end{bmatrix}\begin{bmatrix} 1 \\ 2.2 \end{bmatrix}$$
$$= \begin{bmatrix} 1.24 & 2.12 \end{bmatrix}\begin{bmatrix} 1 \\ 2.2 \end{bmatrix} = 5.904$$
$$(\tilde{x}-\mu_1)^{\top}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}(\tilde{x}-\mu_1) = \begin{bmatrix} -2 & -0.8 \end{bmatrix}\begin{bmatrix} 1.9 & -0.3 \\ -0.3 & 1.1 \end{bmatrix}\begin{bmatrix} -2 \\ -0.8 \end{bmatrix}$$
$$= \begin{bmatrix} -3.56 & -0.28 \end{bmatrix}\begin{bmatrix} -2 \\ -0.8 \end{bmatrix} = 7.344$$

Thus, $h_2(\tilde{x}) = 0$ as

$$(\tilde{x}-\mu_0)^{\top}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}(\tilde{x}-\mu_0) < (\tilde{x}-\mu_1)^{\top}\begin{bmatrix} c & -b \\ -b & a \end{bmatrix}(\tilde{x}-\mu_1)$$

Now, for $h_1(\tilde{x})$,

$$\|\tilde{x}-\mu_0\| = \sqrt{1^2+2.2^2} = 2.416$$
$$\|\tilde{x}-\mu_1\| = \sqrt{(-2)^2+(-0.8)^2} = 2.154$$

Thus, $h_1(\tilde{x}) = 1$ as $\|\tilde{x}-\mu_0\| > \|\tilde{x}-\mu_0\|$.

2. Suppose that the target variable y is given by $y = W^{\top}X + \epsilon$, 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.,

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

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

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

Hint: Use the fact that $\mathrm{Var}(CY) = C\mathrm{Var}(Y)C^{\top}$, 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.

Answer:

(a)

$$y = w^{\top}x + \epsilon \implies Y = AW + E \implies \mathbb{E}[Y] = AW$$
$$\widehat{W}(\lambda) = (A^{\top}A + \lambda\mathcal{I})^{-1}A^{\top}Y$$
$$\mathbb{E}[\widehat{W}(\lambda)] = (A^{\top}A + \lambda\mathcal{I})^{-1}A^{\top}\mathbb{E}[Y]$$
$$= (A^{\top}A + \lambda\mathcal{I})^{-1}A^{\top}AW \neq W \quad \text{for} \quad \lambda > 0.$$

(b) We have $Var(Y) = \mathbb{E}[YY^\intercal] - \mathbb{E}[Y]\mathbb{E}[Y^\intercal]$

$$\begin{aligned}
\text{Also, } \mathbb{E}[Y] &= AW \\
\mathbb{E}[Y^\intercal] &= W^\intercal A^\intercal \\
\mathbb{E}[Y]\mathbb{E}[Y^\intercal] &= AWW^\intercal A^\intercal \\
\mathbb{E}[YY^\intercal] &= \mathbb{E}[(AW + E)(W^\intercal A^\intercal + E^\intercal)] \\
\implies \mathbb{E}[YY^\intercal] &= \mathbb{E}[AWW^\intercal A^\intercal + EW^\intercal A^\intercal + AWE^\intercal + EE^\intercal] \\
\implies \mathbb{E}[YY^\intercal] &= AWW^\intercal A^\intercal + \frac{\mathcal{I}}{\beta}
\end{aligned}$$

Thus,

$$Var(Y) = AWW^\intercal A^\intercal + \frac{\mathcal{I}}{\beta} - AWW^\intercal A^\intercal = \frac{\mathcal{I}}{\beta}$$

Now,

$$\begin{aligned}
\mathrm{Var}(\widehat{W}(\lambda)) &= Var(A^\intercal A + \lambda\mathcal{I})^{-1} A^\intercal Y) \\
&= (A^\intercal A + \lambda\mathcal{I})^{-1} A^\intercal Var(Y) A (A^\intercal A + \lambda\mathcal{I})^{-1} \\
&= \frac{1}{\beta}(A^\intercal A + \lambda\mathcal{I})^{-1} A^\intercal A (A^\intercal A + \lambda\mathcal{I})^{-1}
\end{aligned}$$