CS5691: Pattern recognition and machine learning
Quiz - 2 Solutions
Course Instructor : Prashanth L. A.
Date : Feb-26, 2019 **Duration** : 40 minutes

I. Multiple Choice Questions

1. Let $\{X_1, \ldots, X_n\}$ be i.i.d. samples from $\mathbb{N}(\mu, \sigma^2)$, with $\sigma > 0$. Letting $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, which of the following statements is true?
   (a) $\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 = \sum_{i=1}^n (X_i - \mu)^2$.
   (b) $\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 \leq \sum_{i=1}^n (X_i - \mu)^2$.
   (c) $\sum_{i=1}^n (X_i - \hat{\mu}_n)^2 > \sum_{i=1}^n (X_i - \mu)^2$.
   (d) An inequality/equality relating $\sum_{i=1}^n (X_i - \hat{\mu}_n)^2$ and $\sum_{i=1}^n (X_i - \mu)^2$ does not always hold.

Solution: (b)
Observe that

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n ([X_i - \hat{\mu}_n] + [\hat{\mu}_n - \mu])^2 = \sum_{i=1}^n (X_i - \hat{\mu}_n)^2 + \sum_{i=1}^n (\hat{\mu}_n - \mu)^2 \text{(the cross term vanishes)},$$

leading to the claim in part (b).

2. Consider a Bayesian estimation problem, with data $\{X_1, \ldots, X_n\}$ i.i.d. from $\mathbb{N}(\theta, 1)$, and a $\mathbb{N}(0, 1)$ prior. Letting $S_n = \sum_{i=1}^n X_i$, the posterior mean is
   (a) $\frac{S_n}{n}$
   (b) $\frac{S_n}{n+1}$
   (c) $\frac{nS_n}{n+1}$
   (d) $\frac{S_n + 1}{n+2}$

Solution: (b). Use the expression for posterior mean (derived in the class), subsitute the prior mean/variance values to arrive at the answer.

3. Let $X \sim$ Unif$[0, \theta]$. Then, the maximum likelihood estimate of θ, given i.i.d. samples $\{X_1, \ldots, X_n\}$ is
   (a) $\sum_{i=1}^n \frac{S_n}{n}$.
   (b) $\min_{i=1,\ldots,n} X_i$.
   (c) $\max_{i=1,\ldots,n} X_i$.
   (d) $\frac{1}{2}\left(\max_{i=1,\ldots,n} X_i - \min_{i=1,\ldots,n} X_i\right)$.

Solution: (c). The likelihood function is given by

$$L(\theta) = \frac{1}{\theta^n} \text{ for } 0 \leq X_i \leq \theta, \text{ and 0 elsewhere.}$$

The maximizer of $\frac{1}{\theta^n}$ subject to $X_i \leq \theta$ is $\max_i X_i$. The simpler case of one sample, say X_1, is easy to think about. The uniform density $f_\theta(X_1)$, as a function of θ, is zero if $\theta < X_1$, is $\frac{1}{X_1}$ at $\theta = X_1$, and decreases thereafter, i.e., for $\theta > X_1$. Hence, the ML estimate in the one sample case is X_1.

4. Suppose that we are trying to fit a linear and 10th degree polynomial to data coming from a cubic function, corrupted by standard Gaussian noise. Let M_1 and M_2 denote the models corresponding to the linear and 10 degree polynomial. Then,
   (a) Bias$(M_1) \leq$ Bias(M_2), Variance$(M_1) \leq$ Variance(M_2).
   (b) Bias$(M_1) \leq$ Bias(M_2), Variance$(M_1) \geq$ Variance(M_2).
   (c) Bias$(M_1) \geq$ Bias(M_2), Variance$(M_1) \leq$ Variance(M_2).
   (d) Bias$(M_1) \geq$ Bias(M_2), Variance$(M_1) \geq$ Variance(M_2).

Solution: (c). From the bias-variance tradeoff discussion in class, it is apparent that a linear fit will have a higher bias than a fit using a higher-degree polynomial, while the reverse is true w.r.t. variance, since the training is on a finite dataset.

5. Consider a regression problem, with scalar input $X \in \mathbb{R}$, and target $Y \in \mathbb{R}$. Suppose (X, Y) is bivariate normal with non-zero means, positive variances, and non-zero correlation. Then, the optimal predictor, for the square loss, as a function of X is

(a) Quadratic.
(b) Constant.
(c) Linear.
(d) None of the above.

Solution: (c). Recall that $\mathbb{E}(Y \mid X)$ is the optimal predictor for the square loss. Now, when (X, Y) is bivariate normal, with non-zero correlation, then $\mathbb{E}(Y \mid X)$ is a linear function of X (Why?).

II. A problem that requires a detailed solution

1. Consider a distribution over (X, Y) given by the following assumptions:
$Y \in \{-1, +1\}, X \in \{0, 1\}^3$.
$\mathbb{P}(Y = +1) = a, \mathbb{P}(Y = -1) = 1 - a$,
$X|Y = -1 \sim \text{Bern}(\theta_1) \times \text{Bern}(\theta_2) \times \text{Bern}(\theta_3)$,
$X|Y = +1 \sim \text{Bern}(\tau_1) \times \text{Bern}(\tau_2) \times \text{Bern}(\tau_3)$.
We have 10 training points from the above distribution, given by the table below.

X_1	X_2	X_3	Y
1	0	0	$+1$
0	1	1	-1
0	1	0	$+1$
1	1	0	$+1$
1	1	1	-1
1	0	0	$+1$
1	0	1	$+1$
0	0	1	-1
0	1	1	$+1$
0	0	0	-1

(a) Give the ML estimates for $a, \theta_1, \theta_2, \theta_3, \tau_1, \tau_2, \tau_3$. (3 marks)

(b) For all the 8 points in the instance space $\{0, 1\}^3$, give the estimate of the posterior probability $\mathbb{P}(Y = +1 \mid X)$, and give the prediction that minimises the mis-classification rate (or the Bayes classifier for the zero-one loss), in the form of a table with 8 rows. (2 marks)

Solution: The ML estimate of a Bernoulli parameter p from n samples is simply $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Therefore the ML parameters are given by:

$$\widehat{a} = \frac{6}{10},$$
$$\widehat{\theta}_1 = \frac{1}{4}, \quad \widehat{\theta}_2 = \frac{2}{4}, \quad \widehat{\theta}_3 = \frac{3}{4},$$
$$\widehat{\tau}_1 = \frac{4}{6}, \quad \widehat{\tau}_2 = \frac{3}{6}, \quad \widehat{\tau}_3 = \frac{2}{6}.$$

The table of posterior probabilities, and Bayes classifier's prediction is given by

X_1	X_2	X_3	$P(X\|Y = -1)$	$P(X\|Y = +1)$	$P(Y = +1\|X)$	$h^*(X)$
0	0	0	$\frac{3}{32}$	$\frac{1}{9}$	0.64	$+1$
0	0	1	$\frac{9}{32}$	$\frac{1}{18}$	0.22	-1
0	1	0	$\frac{3}{32}$	$\frac{1}{9}$	0.64	$+1$
0	1	1	$\frac{9}{32}$	$\frac{1}{18}$	0.22	-1
1	0	0	$\frac{1}{32}$	$\frac{2}{9}$	0.914	$+1$
1	0	1	$\frac{3}{32}$	$\frac{1}{9}$	0.64	$+1$
1	1	0	$\frac{1}{32}$	$\frac{2}{9}$	0.914	$+1$
1	1	1	$\frac{3}{32}$	$\frac{1}{9}$	0.64	$+1$