DS 100/200: Principles and Techniques of Data Science Date: March 8, 2019

Discussion #6 Exam Prep

Name:

1. Suppose in some universe, the true relationship between the measured luminosity of a single star Y can be written in terms of a single feature ϕ of that same star as

$$Y = \theta^* \phi + \epsilon$$

where $\phi \in \mathbb{R}$ is some non-random scalar feature, $\theta^* \in \mathbb{R}$ is a non-random scalar parameter, and ϵ is a random variable with $\mathbb{E}[\epsilon] = 0$ and $\text{var}(\epsilon) = \sigma^2$. For each star, you have a set of features $\mathbf{\Phi} = \begin{bmatrix}\phi_1 & \phi_2 & \dots & \phi_n\end{bmatrix}^T$ and luminosity measurements $\mathbf{y} = \begin{bmatrix}y_1 & y_2 & \dots & y_n\end{bmatrix}^T$ generated by this relationship. Your Φ may or may not include the feature ϕ described above. The ϵ_i for the various y_i have the same probability distribution and are independent of each other.

   (a) What is $\mathbb{E}[Y]$?

   ◯ A. 0 ◯ B. $\theta^* \phi$ ◯ C. $\phi(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{y}$
   ◯ D. θ^* ◯ E. None of the above

   (b) What is $\text{var}(Y)$?

   ◯ A. $\frac{\sigma^2}{n}$ ◯ B. $\frac{\sigma^2}{n^2}$ ◯ C. 0
   ◯ D. $\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2$ ◯ E. None of the above

2. What parameter estimate would minimize the following regularized loss function:

$$\ell(\theta) = \lambda(\theta - 4)^2 + \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2 \tag{1}$$

   ◯ A. $\hat{\theta} = \frac{1}{\lambda n}\sum_{i=1}^{n} x_i$

   ◯ B. $\hat{\theta} = 4 + \frac{1}{\lambda n}\sum_{i=1}^{n} x_i$

   ◯ C. $\hat{\theta} = \frac{1}{n(\lambda+1)}\sum_{i=1}^{n} x_i$

   ◯ D. $\hat{\theta} = \frac{\lambda}{\lambda+1} + \frac{1}{n(\lambda+1)}\sum_{i=1}^{n}(x_i - 4)$

   ◯ E. $\hat{\theta} = \frac{4\lambda}{\lambda+1} + \frac{1}{n(\lambda+1)}\sum_{i=1}^{n} x_i$

3. Suppose $X_1, \ldots, X_n$ are random variables with $\mathbb{E}[X_i] = \mu^*$ and $\mathbf{Var}[X_i] = \theta^*$. Consider the following loss function

$$\ell(\theta) = \log(\theta) + \frac{1}{n\theta} \sum_{i=1}^{n} X_i^2.$$

Let $\widehat{\theta}$ denote the minimizer for $\ell(\theta)$. What is $\mathbb{E}[\,\widehat{\theta}\,]$?

◯ A. θ^* ◯ B. $\theta^* + \mu^*$ ◯ C. $\theta^* + \mu^*/2$ ◯ D. $\mathbb{E}\left[\theta^* + \mu^*\right]$ ◯ E. $\theta^* + (\mu^*)^2$
◯ F. $(\theta^* + \mu^*)^2$

4. Let $x_1, \ldots, x_n$ denote any collection of numbers with average $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

   (a) $\sum_{i=1}^{n}(x_i - \overline{x})^2 \leq \sum_{i=1}^{n}(x_i - c)^2$ for all c.

   ◯ A. True ◯ B. False

   (b) $\sum_{i=1}^{n}|x_i - \overline{x}| \leq \sum_{i=1}^{n}|x_i - c|$ for all c.

   ◯ A. True ◯ B. False

5. Consider the following loss function based on data $x_1, \ldots, x_n$:

$$\ell(\mu, \sigma) = \log(\sigma^2) + \frac{1}{n\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

   (a) Which estimator $\hat{\mu}$ is a minimizer for μ, i.e. satisfies $\ell(\hat{\mu}, \sigma^2) \leq \ell(\mu, \sigma^2)$ for any μ, σ?

   ◯ A. $\hat{\mu} = 0$

   ◯ B. $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- ◯ C. $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i + \log\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$
- ◯ D. $\hat{\mu} = \frac{1}{n\sigma^2}\sum_{i=1}^{n} x_i + \log(\sigma^2)$
- ◯ E. $\hat{\mu} = \texttt{median}(x_1, \ldots, x_n)$.

(b) Which of the following is the result of solving $\ell\sigma = 0$ for σ (for fixed μ)?

- ◯ A. $\sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$.
- ◯ B. $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$.
- ◯ C. $\sigma = \frac{2}{n}\sum_{i=1}^{n}(\mu - x_i)$.
- ◯ D. $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - x_j)^2}$.

6. Suppose we create a new loss function called the OINK loss, defined as follows for a single observation:

$$L_{OINK}(\theta, x, y) = \begin{cases} a(f_\theta(x) - y) & f_\theta(x) \geq y \\ b(y - f_\theta(x)) & f_\theta(x) < y \end{cases}$$

You decide to use the constant model (given on the left) and average OINK loss (given on the right).

$$f_\theta(x) = \theta \qquad\qquad L(\theta, \mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} L_{OINK}(\theta, x_i, y_i)$$

The data are given below. Find the optimal $\hat{\theta}$ that minimizes the loss.

x	3	1	5	4	2	0	6
y	40	0	50	30	20	60	10

(a) when $a = b = 1$

(b) when $a = 1$, $b = 5$

(c) when $a = 3$, $b = 6$