DS 100/200: Principles and Techniques of Data Science Date: March 15, 2019

Discussion #7 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) Suppose you have information about the exact ϕ value for each star, but try to fit a linear model for Y that includes an intercept term θ_0.

$$Y = \theta_0 + \theta_1 \phi$$

Note the true relationship has no intercept term, so our model is not quite correct. Let $\hat{\theta}_0$ and $\hat{\theta}_1$ be the values that minimize the average L_2 loss. Let $\mathbf{y}$ be the actual observed data and $\hat{\mathbf{y}} = \hat{\theta}_0 + \hat{\theta}_1 \boldsymbol{\phi}$ be the fitted values.

i. Which of the following could possibly be the value of $\hat{\theta}_0$ after fitting our model? **Select all that apply; at least one is correct.**

☐ A. -1 ☐ B. 0 ☐ C. 1 ☐ D. 10

ii. Which of the following could possibly be the residual vector for our model? **Select all that apply; at least one is correct.**

☐ A. $\begin{bmatrix}-2 & -4 & 6\end{bmatrix}^T$ ☐ B. $\begin{bmatrix}0.0001 & 0.0003 & -0.0005\end{bmatrix}^T$
☐ C. $\begin{bmatrix}3 & 12 & -9\end{bmatrix}^T$ ☐ D. $\begin{bmatrix}1 & 1 & 1\end{bmatrix}^T$

2. Throughout this section we refer to "least squares regression", which is the process of minimizing the average L2 loss using a linear regression model. Ordinary least squares is the version of least squares regression where **we do not use regularization**. Assume throughout that **our model includes a bias term**.

(a) What is always true about the residuals in least squares regression? Select all that apply.

☐ A. They are orthogonal to the column space of the features.
☐ B. They represent the errors of the predictions.
☐ C. Their sum is equal to the mean squared error.
☐ D. Their sum is equal to zero.
☐ E. None of the above.

(b) Which are true about the predictions made by OLS? Select all that apply.

☐ A. They are projections of the observations onto the column space of the features.
☐ B. They are linear in the chosen features.
☐ C. They are orthogonal to the residuals.
☐ D. They are orthogonal to the column space of the features.
☐ E. None of the above.

(c) Which of the following would be true if you chose mean absolute error (L1) instead of mean squared error (L2) as your loss function? Select all that apply.

☐ A. The results of the regression would be more sensitive to outliers.
☐ B. You would not be able to use gradient descent to find the regression line.
☐ C. You would not be able to use the normal equation to calculate your parameters.
☐ D. The sum of the residuals would now be zero.
☐ E. None of the above.

3. Let $\hat{\boldsymbol{y}} \in \mathbb{R}^n$ be the vector of fitted values in the ordinary least squares regression of $\boldsymbol{y} \in \mathbb{R}^n$ on the full column-rank feature matrix $\boldsymbol{\Phi} \in \mathbb{R}^{n \times d}$ with n much larger than d. Denote the fitted coefficients as $\hat{\boldsymbol{\beta}} \in \mathbb{R}^d$ and the vector of residuals as $\boldsymbol{e} \in \mathbb{R}^n$.

(a) What is $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$?

◯ A. $\mathbf{0}$ ◯ B. $\hat{\boldsymbol{y}}$ ◯ C. $\boldsymbol{e}$ ◯ D. $\hat{\boldsymbol{\beta}}$ ◯ E. 1 ◯ F. None of the above

(b) What is $\boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\hat{\boldsymbol{y}}$? Notice: This problem has a hat in $\hat{\boldsymbol{y}}$.

◯ A. $\mathbf{0}$ ◯ B. $\hat{\boldsymbol{y}}$ ◯ C. $\boldsymbol{e}$ ◯ D. $\hat{\boldsymbol{\beta}}$ ◯ E. 1 ◯ F. None of the above

Suppose $\boldsymbol{e} \neq \mathbf{0}$. Define a new feature matrix $\boldsymbol{\Psi}$ by appending the residual vector $\boldsymbol{e}$ to the feature matrix $\boldsymbol{\Phi}$. In other words,

$$\boldsymbol{\Psi} = \begin{bmatrix} | & | & \vdots & | & | \\ \boldsymbol{\Phi_{:,1}} & \boldsymbol{\Phi_{:,2}} & \cdots & \boldsymbol{\Phi_{:,d}} & \boldsymbol{e} \\ | & | & \vdots & | & | \end{bmatrix}$$

(c) We now want to fit the model $\boldsymbol{y} = \boldsymbol{\Psi\gamma} = \gamma_1\boldsymbol{\Phi_{:,1}} + \gamma_2\boldsymbol{\Phi_{:,2}} + \cdots + \gamma_d\boldsymbol{\Phi_{:,d}} + \gamma_{d+1}\boldsymbol{e}$ by choosing $\hat{\boldsymbol{\gamma}} = [\hat{\gamma}_1 \ldots \hat{\gamma}_{d+1}]^T$ to minimize the L_2 loss. What is $\hat{\gamma}_{d+1}$?

◯ A. 0 ◯ B. 1 ◯ C. $\boldsymbol{e}^T\boldsymbol{y}$ ◯ D. $1 - \hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}$
◯ E. $(\Phi^T\Phi)^{-1}\Phi^T$ ◯ F. None of the above

4. We collect some data $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$ and decide to model the relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$ as

$$\boldsymbol{y} = \beta_1\boldsymbol{\Phi}_{:,1} + \beta_2\boldsymbol{\Phi}_{:,2}$$

where $\boldsymbol{\Phi}_{i,:} = \begin{bmatrix} 1 & x_i \end{bmatrix}$ We found the estimates $\hat{\beta}_1 = 2$ and $\hat{\beta}_2 = 5$ for the coefficients by minimizing the L_2 loss. Given that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \begin{bmatrix} 4 & 2 \\ 2 & 5 \end{bmatrix}$, answer the following problems. If not enough information is given, write "Cannot be determined."

(a) What was the sample size n? Hint: Consider the form of the feature matrix.

(b) What must $\boldsymbol{\Phi}^T\boldsymbol{y}$ be for this data set?

5. Consider the following loss function based on data $x_1, \ldots, x_n$ with mean $\overline{x}$:

$$\ell(\beta) = \log\beta + \frac{\overline{x}}{\beta} + \frac{1}{n}\sum_{i=1}^{n} e^{-x_i/\beta}$$

Given an estimate $\beta^{(t)}$, write out the update $\beta^{(t+1)}$ after one iteration of gradient descent with step size α.