Data 100, Spring 2024

Exam Prep Section #5

Flight Delay

1. Professor Norouzi is building a flight delay linear regression model using a dataset of historical flight information. Suppose the flight information dataset contains 400 flight records and 10 columns.
   Of the 10 columns, 1 contains the observed flight delays in minutes, and 9 columns have non-constant numerical features related to the flight (quantitative variables).
   Assume that we find the least squares estimate for parameter vector θ (with intercept) as $\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$ for a design matrix $\mathbb{X}$ and true output vector $\mathbb{Y}$.

   (a) What is the shape of the design matrix $\mathbb{X}$?

   (b) What is the shape of the parameter vector θ?

   (c) What is the shape of the true output vector $\mathbb{Y}$?

   (d) What is the shape of the predicted output vector $\hat{\mathbb{Y}}$?

(e) What is the residual vector? Write down the equation for calculating the residual vector using $\mathbb{Y}$, $\mathbb{X}$, and $\hat{\theta}$.

(f) What is always true about the residuals in any Ordinary Least Squares (OLS) regression? Select all that apply.

- ☐ A. They are orthogonal to the column space of the design matrix ($\mathbb{X}$).
- ☐ B. Their sum is equal to the mean squared error.
- ☐ C. Their sum is equal to zero.
- ☐ D. They are orthogonal to the vector of predictions ($\hat{\mathbb{Y}}$)

Linear Regression Fundamentals

2. In this problem, we will review some of the core concepts of linear regression. Suppose we create a linear model with parameters $\hat{\theta} = [\hat{\theta}_0, \ldots, \hat{\theta}_p]$. Given an observation $\vec{x}$, this model predicts $\hat{y} = \hat{\theta} \cdot \vec{x} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \ldots + \hat{\theta}_p x_p$. The design matrix and observations used to construct $\hat{\theta}$ are $\mathbb{X}$ and $\mathbb{Y}$.

   (a) Suppose $\hat{\theta} = [2, 0, 1]$ and we receive an observation $\vec{x_1} = [1, 2, 5]$. What $\hat{y_1}$ value will this model predict for the given observation?

   (b) Suppose the true y_1 was 3. What will be the L_2 loss for our prediction $\hat{y_1}$ from the previous part?

   (c) Which of the following statements are always true? **Select all that apply**

   - ☐ A. $\vec{x}_1^T \hat{\theta} = 0$
   - ☐ B. $\vec{x}_1^T \hat{\theta} = \hat{y}_1$
   - ☐ C. $\vec{x}_1^T \hat{\theta} - y_1 = 0$
   - ☐ D. $\mathbb{X}\hat{\theta} - \mathbb{Y} = 0$
   - ☐ E. $\mathbb{X}^T(\mathbb{X}\hat{\theta} - \mathbb{Y}) = 0$

   (d) (T/F) Define the residuals of this model as $e_i = y_i - \hat{y}_i$. For all data points x_i and y_i in $\mathbb{X}$, $\mathbb{Y}$, the sum of residuals $\sum_i e_i = 0$. Justify why.

   (e) Suppose we arbitrarily removed a feature from the design matrix. Which of the following could happen to the new optimal loss compared to the old optimal loss?

   - ○ A. The optimal loss must decrease.
   - ○ B. The optimal loss may decrease, but it must not increase.
   - ○ C. The optimal loss may increase, but it must not decrease.
   - ○ D. The optimal loss must increase.