Data 100, Spring 2024

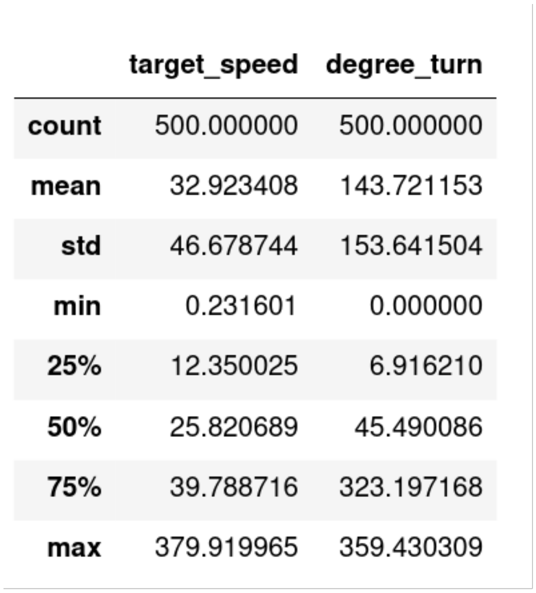
Discussion #6

Note: Your TA will probably not cover all the problems. This is totally fine, the discussion worksheets are not designed to be finished within an hour. They are deliberately made slightly longer so they can serve as resources you can use to practice, reinforce, and build upon concepts discussed in lectures, labs, and homework.

Driving with a Constant Model

1. Lillian is trying to use modeling to drive her car autonomously. To do this, she collects a lot of data from driving around her neighborhood and stores it in `drive`. She wants your help to design a model that can drive on her behalf in the future using the outputs of the models you design. First, she wants to tackle two aspects of this autonomous car modeling framework: going forward and turning.

   Some statistics from the collected dataset are shown below using `drive.describe()`, which returns the mean, standard deviation, quartiles, minimum, and maximum for the two columns in the dataset: `target_speed` and `degree_turn`.



	target_speed	degree_turn
count	500.000000	500.000000
mean	32.923408	143.721153
std	46.678744	153.641504
min	0.231601	0.000000
25%	12.350025	6.916210
50%	25.820689	45.490086
75%	39.788716	323.197168
max	379.919965	359.430309

(a) Suppose the first part of the model predicts the target speed of the car. Using constant models trained on the speeds of the collected data shown above with L_1 and L_2 loss functions, which of the following is true?

○ A. The model trained with the L_1 loss will always drive slower than the model trained with L_2 loss.

○ B. The model trained with the L_2 loss will always drive slower than the model trained with L_1 loss.

○ C. The model trained with the L_1 loss will sometimes drive slower than the model trained with L_2 loss.

○ D. The model trained with the L_2 loss will sometimes drive slower than the model trained with L_1 loss.

(b) Finding that the model trained with the L_2 loss drives too slowly, Lillian changes the loss function for the constant model where the loss is penalized **more** if the true speed is higher. That way, in order to minimize loss, the model would have to output predictions closer to the true value, particularly as speeds get faster, the end result being a higher constant speed. Lillian writes this as $L(y, \hat{y}) = y(y - \hat{y})^2$.

Find the optimal $\hat{\theta_0}$ for the constant model using the new empirical risk function $R(\theta_0)$ below:

$$R(\theta_0) = \frac{1}{n}\sum_i y_i(y_i - \theta_0)^2$$

(c) Lillian's friend, Yash, also begins working on a model that predicts the degree of turning at a particular time between 0 and 359 degrees using the data in the `degree_turn` column. Explain why a constant model is likely inappropriate in this use case.

Extra: If you've studied some physics, you may recognize the behavior of our constant model!

(d) Suppose we finally expand our modeling framework to use simple linear regression (i.e. $f_\theta(x) = \theta_{w,0} + \theta_{w,1}x$). For our first simple linear regression model, we predict the turn angle (y) using target speed (x). Our optimal parameters are: $\hat{\theta}_{w,1} = 0.019$ and $\hat{\theta}_{w,0} = 143.1$.

However, we realize that we actually want a model that predicts target speed (our new y) using turn angle, our new x (instead of the other way around)! What are our new optimal parameters for this new model?

Geometry of Least Squares

2. Using the geometry of least squares, let's answer a few questions about Ordinary Least Squares (OLS)!

   (a) Which of the following are true about the optimal solution $\hat{\theta}$ to OLS? Recall that the least squares estimate $\hat{\theta}$ solves the normal equation $(\mathbb{X}^T\mathbb{X})\theta = \mathbb{X}^T\mathbb{Y}$.

   $$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

   □ A. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_2 loss.

   □ B. Using the normal equation, we can derive an optimal solution for simple linear regression with an L_1 loss.

   □ C. Using the normal equation, we can derive an optimal solution for a constant model with an L_2 loss.

   □ D. Using the normal equation, we can derive an optimal solution for a constant model with an L_1 loss.

   □ E. Using the normal equation, we can derive an optimal solution for the model $\hat{y} = \theta_1 x + \theta_2 \sin(x^2)$.

   □ F. Using the normal equation, we can derive an optimal solution for the model $\hat{y} = \theta_1\theta_2 + \theta_2 x^2$.

   (b) Which of the following conditions are required for the least squares estimate in the previous subpart?

   □ A. $\mathbb{X}$ must be full column rank.

   □ B. $\mathbb{Y}$ must be full column rank.

   □ C. $\mathbb{X}$ must be invertible.

   □ D. $\mathbb{X}^{\mathbb{T}}$ must be invertible.

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

   □ A. They are orthogonal to the column space of the design matrix.

   □ 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.

(d) Which of the following 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 design matrix.

□ B. They are linear combinations of the features.

□ C. They are orthogonal to the residuals.

□ D. They are orthogonal to the column space of the features.

□ E. None of the above.

(e) Which of the following is true of the mystery quantity $\vec{v} = (I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T)\mathbb{Y}$?

□ A. The vector $\vec{v}$ represents the residuals for any linear model.

□ B. If the $\mathbb{X}$ matrix contains the $\vec{1}$ vector, then the sum of the elements in vector $\vec{v}$ is 0 (i.e. $\sum_i v_i = 0$).

□ C. All the column vectors x_i of $\mathbb{X}$ are orthogonal to $\vec{v}$.

□ D. If $\mathbb{X}$ is of shape n by p, there are p elements in vector $\vec{v}$.

□ E. For any $\vec{\alpha}$, $\mathbb{X}\vec{\alpha}$ is orthogonal to $\vec{v}$.

Modeling using Multiple Regression

3. Ishani wants to model exam grades for DS100 students. She collects various information about student habits, such as how many hours they studied, how many hours they slept before the exam, and how many lectures they attended and observes how well they did on the exam. Suppose she collected such information on n students, and wishes to use a multiple-regression model to predict exam grades.

   (a) Using the data of the n individuals, she constructs a design matrix $\mathbb{X}$, and uses the OLS formula to obtain the following $\hat{\theta}$:

   $$\hat{\theta} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

   The design matrix $\mathbb{X}$ was constructed such that the first column represents how many hours each of the n students studied, the second represents how many hours each student slept before the exam, and the third represents how many lectures each student attended. With this knowledge, give an interpretation of what each entry of $\hat{\theta}$ means in context.

   (b) After fitting this model, we would like to predict the exam grades for two individuals using these variables. Suppose for Individual 1, they slept 10 hours, studied 15 hours, and attended 4 lectures. Suppose also for Individual 2, they slept 5 hours, studied 20 hours, and attended 10 lectures. Construct a matrix $\mathbb{X}'$ such that, if you computed $\mathbb{X}'\hat{\theta}$, you would obtain a vector of each individual's predicted exam scores.

   (c) Denote y' as a 2×1 vector that represents the actual exam scores of the individuals Ishani is predicting on. Write out an expression that evaluates to give the Mean Squared Error (MSE) of our predictions using matrix notation.