DS 100/200: Principles and Techniques of Data Science	Date: April 3, 2020

Discussion #9

Name:

Cross Validation

1. Describe the k-fold cross validation procedure and why we might use it in developing models.

2. Give some limitations of cross-validation.

Feature Engineering

3. Consider the following model training script to estimate the training error:

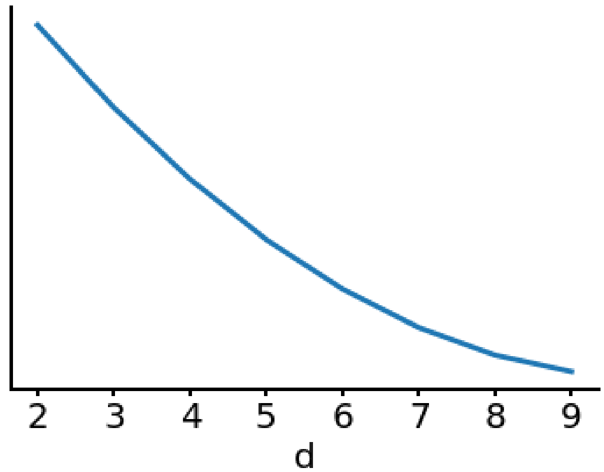
```
1  X_train, X_test, y_train, y_test =
2      train_test_split(X, y, test_size=0.1)
3
4  model = lm.LinearRegression(fit_intercept=True)
5  model.fit(X_test, y_test)
6
7  y_fitted = model.predict(X_train)
8  y_predicted = model.predict(X_test)
9
10 training_error = rmse(y_fitted, y_predicted)
```

4. There are two major mistakes in the code above. Identify the line where each mistake occurs and explain how you would fix them.

5. Which of the following techniques could be used to reduce over-fitting?
   - ◯ A. Adding noise to the training data
   - ◯ B. Cross-validation to remove features
   - ◯ C. Fitting the model on the test split
   - ◯ D. Adding features to the training data

6. Your team would like to train a machine learning model in order to predict the next YouTube video that a user will click on based on the videos the user has watched in the past. We extract m attributes (such as length of video, view count etc) from each video and our model will be based on the previous d videos watched by that user. Hence the number of features for each data point for the model is $m \cdot d$. You're not sure how many videos to consider.

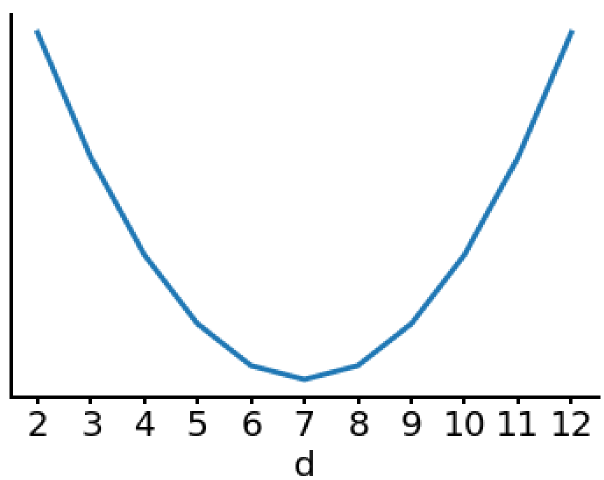
(a) Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

☐ A. Training Error
☐ B. Validation Error
☐ C. Bias
☐ D. Variance

(b) Your colleague generates the following plot, where the value d is on the x axis. However, they forgot to label the y-axis.



Which of the following could the y axis represent? Select all that apply.

☐ A. Training Error
☐ B. Validation Error
☐ C. Bias
☐ D. Variance

Dummy Variables/One-hot Encoding

In order to include a qualitative variable in a model, we convert it into a collection of dummy variables. These dummy variables take on only the values 0 and 1. For example, suppose we have a qualitative variable with 3 possible values, call them A, B, and C, respectively. For concreteness, we use a specific example with 10 observations:

$$[A, A, A, A, B, B, B, C, C, C]$$

We can represent this qualitative variable with 3 dummy variables that take on values 1 or 0 depending on the value of this qualitative variable. Specifically, the values of these 3 dummy variables for this dataset are $\vec{x}_A$, $\vec{x}_B$, and $\vec{x}_C$, arranged from left to right in the following design matrix, where we use the following indicator variable:

$$\vec{x}_{k,i} = \begin{cases} 1 & \text{if } i\text{-th observation has value } k \\ 0 & \text{otherwise.} \end{cases}$$

This representation is also called one-hot encoding.

$$\begin{bmatrix} | & | & | \\ \vec{x}_A & \vec{x}_B & \vec{x}_C \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

We will show that the fitted coefficients for $\vec{x}_A$, $\vec{x}_B$, and $\vec{x}_C$ are $\bar{y}_A$, $\bar{y}_B$, and $\bar{y}_C$, the average of the y_i values for each of the groups, respectively.

7. Show that the columns of $\mathbb{X}$ are orthogonal, (i.e., the dot product between any pair of column vectors is 0).

8. Show that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} n_A & 0 & 0 \\ 0 & n_B & 0 \\ 0 & 0 & n_C \end{bmatrix}$$

Here, n_A, n_B, n_C are the number of observations in each of the three groups defined by the levels of the qualitative variable.

9. Show that

$$\mathbb{X}^T\vec{y} = \begin{bmatrix} \sum_{i\in A} y_i \\ \sum_{i\in B} y_i \\ \sum_{i\in C} y_i \end{bmatrix}$$

10. Use the results from the previous questions to solve the normal equations for $\hat{\beta}$, i.e.,

$$\begin{aligned} \hat{\beta} &= [\mathbb{X}^T\mathbb{X}]^{-1}\mathbb{X}^T\vec{y} \\ &= \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix} \end{aligned}$$