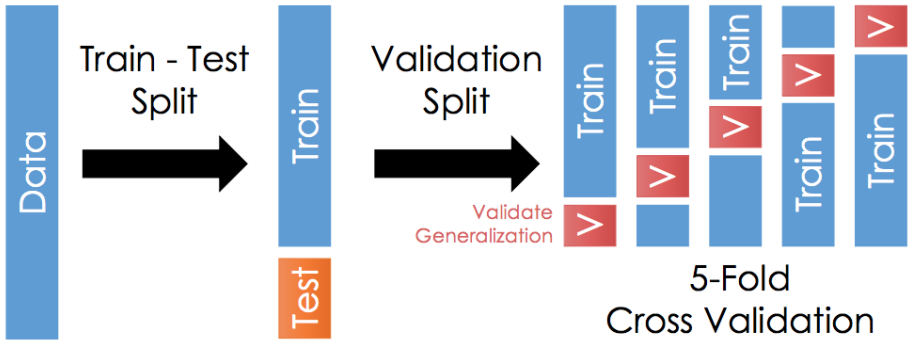
DS 100/200: Principles and Techniques of Data Science **Date: April 3, 2020**

Discussion #9 Solutions

Name:

Cross Validation

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



Solution:

We want to test that we are not overfitting to the training data, but we cannot evaluate our model on the test set until the model is finalized. Instead, to simulate this process, we can split the training set once more into a smaller sub-training set and a validation set. The validation error provides us with an idea of how well our model generalizes to new and unknown data points without ever having to touch the test set.

In k-fold cross validation, we repeat the procedure above k times but hold out a different subset of the training data for each fold. Our cross-validation error is the mean validation error across all k-folds.

2. Give some limitations of cross-validation.

Solution: If the system evolves in time or has spatial properties, then one should be very careful in selecting the folds for cross-validation to account for the structure of the data.

CV can be computationally intensive. If the training procedure takes a long time, then CV will be very slow as each fold will result in a separate training stage.

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.

Solution:

Line 5: We should be training our model on the training set, not the test set.

Line 10: Training error is the RMSE of our models predictions on the training data and the actual values of the training data. $y_{predicted}$ should be y_{train}

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

Solution:

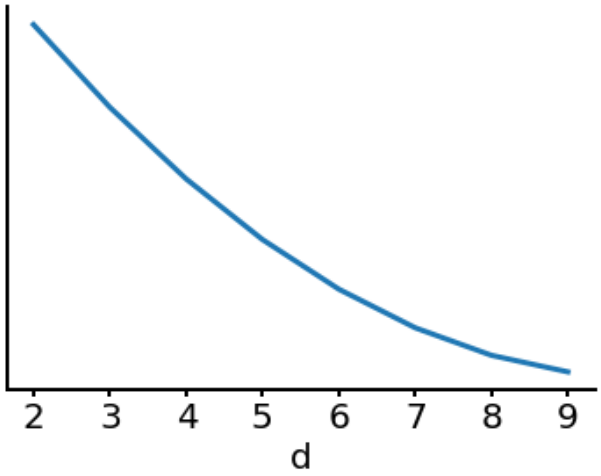
Fitting our model to the test split is not valid because our test data will become training data. Adding features makes our model more complex, which increases our chance of overfitting.

By cross-validating on the features and removing features that make our validation error worse, we are decreasing the complexity of our model as well as our chance of overfitting.

Adding noise to the training data can potentially work by making the training data harder to fit to(and consequently yielding a more general model).

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.

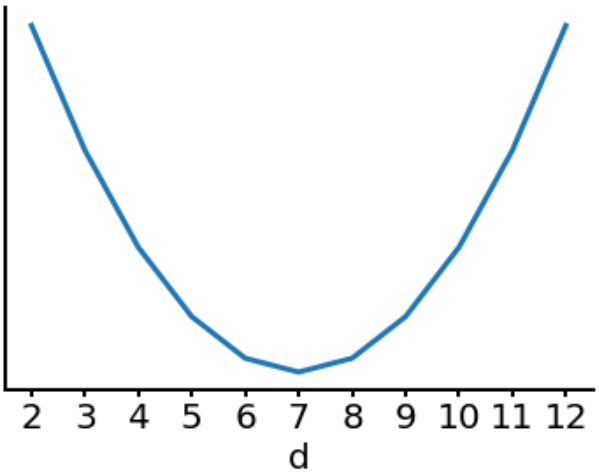


   Which of the following could the y axis represent? Select all that apply.

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

   Solution:
   Training Error: Training error decreases as we add more features
   Validation Error: Can be true depending on the underlying complexity of the data. In part (b), we see that increasing d does not necessarily lower our validation error
   Bias: Decreases with model complexity
   Variance: Increases with model complexity

   (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

Solution:
See solution in part (a)

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).

Solution: The argument is the same for any pair of $\vec{x}$s so we show the orthogonality for one pair, $\vec{x}_A \dot{\vec{x}}_B$.

$$\begin{aligned}\vec{x}_A \dot{\vec{x}}_B &= \sum_{i=1}^{10} x_{A,i} x_{B,i} \\ &= \sum_{i=1}^{4}(1 \times 0) + \sum_{i=5}^{7}(0 \times 1) + \sum_{i=8}^{10}(0 \times 0) \\ &= 0\end{aligned}$$

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.

Solution: Here, we note that

$$\mathbb{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

We also note that

$$\mathbb{X}^T\mathbb{X} = \begin{bmatrix} \vec{x}_A^T\vec{x}_A & \vec{x}_A^T\vec{x}_B & \vec{x}_A^T\vec{x}_C \\ \vec{x}_B^T\vec{x}_A & \vec{x}_B^T\vec{x}_B & \vec{x}_B^T\vec{x}_C \\ \vec{x}_C^T\vec{x}_A & \vec{x}_C^T\vec{x}_B & \vec{x}_C^T\vec{x}_C \end{bmatrix}$$

Since we earlier established the orthogonality of the vectors in $\mathbb{X}$, we find $\mathbb{X}^T\mathbb{X}$ to be the diagonal matrix:

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

9. Show that

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

Solution: Note in the previous solution we found $\mathbb{X}^t$. The solution follows from recognizing that for a row in $\mathbb{X}^t$, e.g., the first row, we have

$$\sum_{i=1}^{10} x_{A,i} \times y_i = \sum_{i=1}^{4} y_i = \sum_{i \in \text{ group A}} y_i$$

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

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

Solution: By inspection, we can find

$$[\mathbb{X}^T\mathbb{X}]^{-1} = \begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix}$$

When we pre-multiply $\mathbb{X}^T\vec{y}$ by this matrix, we get

$$\begin{bmatrix} \frac{1}{n_A} & 0 & 0 \\ 0 & \frac{1}{n_B} & 0 \\ 0 & 0 & \frac{1}{n_C} \end{bmatrix} \begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in B} y_i \\ \sum_{i \in C} y_i \end{bmatrix} = \begin{bmatrix} \bar{y}_A \\ \bar{y}_B \\ \bar{y}_C \end{bmatrix}$$