Project T Final

Topic: 9. Training/Testing, Cross-Validation, Bias-Variance Quiz

Sean Betancourt, Erik Fisher, and Jenny Wang

December 2020

1 Terminology Tumble

1. The _____ set is data used to tune a model's weights.
2. The _____ set is data used to find the final prediction accuracy of the model.
3. The _____ set is data used to test a model's prediction error before changing some parameters for the model.
4. K-fold cross validation is helpful when your dataset is __small/large__.
5. _____ expresses the extent to which a model's predictions are swayed by its training data.
6. _____ expresses the extent to which a model makes assumptions about the data.

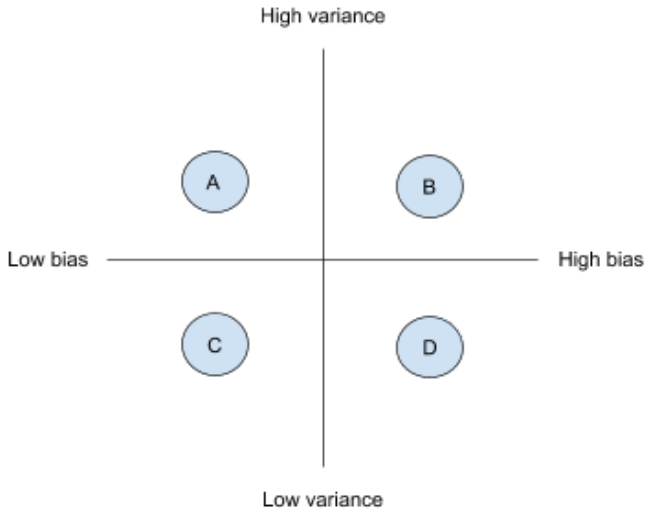
2 Gearing Up

Mike owns a shop that sells a lot of gears. The market is good, but he is having a hard time predicting customer demand for his products in the wintertime! He gives you a dataset containing the items he sold in the past few years, the times he sold them, and some features about them. Could you devise a business strategy for building a model that predicts the customer demand for an item that follows good practices for train/test workflows?

3 The Bias Variance Match

We're predicting startup success from their product price, resource cost, quarterly earnings, and more. Your friend Xavier shows you the predictions of a few

of the prototype models he trained and asks for help. Based on each model's performance, what are their levels of bias and variance? Match each model to a letter.



1. Model 1 confidently predicts success for a certain startup profile, but confidently predicts the opposite if the profile's resource cost is decreased by only 0.1%.
2. Model 2 tends to classify all startups with quarterly earnings greater than \$50k as successful regardless of country and its local average earnings.
3. Model 3's worst mistake was misclassifying a certain startup in the test set as unsuccessful, but a very similar startup in the training set was also labeled unsuccessful
4. Model 4's prediction is always the first training point's success rate.

4 Dataset Dilemma

Diego wants to estimate turtle age based on shell size, shell shape, and other relevant features. He drives to the aquarium and obtains a nice and large dataset, but there is a lot of noise in it! He notes a common tradeoff between bias and variance for a model. When picking a model, which is more important– the bias or the variance of the model? Why?

5 Uncovering the Unknown

To demonstrate the butterfly effect, Donna wants to train a model to predict this year's proportion of green-eyed babies based on Jeff Bezos's yearly earnings, the proportion of blue-painted cars in existence, and the number of plant species that went extinct that year. Assume she has an infinite number of data points and she successfully tunes her model to be the best it can possibly be at training time. Do you expect the model to achieve high training accuracy? How about test accuracy?

6 The Divided Trio

Meggie is using the below structure to train and test her model, where KFold is k-fold cross validation in the sklearn library. Draw how the data is split in each loop. An example is in the slides.

```
kf = KFold(n_splits=4)
for train_index, val_index in kf.split(X):
    X_train, X_val = X[train_index], X[val_index]
    y_train, y_val = y[train_index], y[val_index]

    # ... train then calculate validation accuracy
# ... calculate test accuracy with X_test
```

7 Time Saved or Time Wasted?

Craig is doing object recognition for an underwater dataset. He knows the dataset is moderately small, but he decides against using k-fold cross validation in order to save some computational power. Instead, he divides the original dataset into a training, validation, and test set in a 6:2:2 ratio and does not reuse data. What effects will this have on his training and test accuracies? How can this be described in terms of bias and variance?

8 MSE Bias Variance Derivation

Try it yourself! Can you massage the mean squared error expression $MSE = E[(h(x|D) - f(x))^2]$ into the form containing bias and variance? It is given by $(E[h(x|D)] - f(s))^2 + Var(h(x|D)) + Var(Z)$, where $h(x|D)$ is the model's prediction given a training dataset, $f(x)$ is the true label, and Z is the inherent noise in the labels.

9 MSE Bias Variance at Training and Test Times

As seen in **MSE Bias Variance Derivation**, the MSE error has bias, variance, and irreducible error terms. Is there anything you can say about the value of irreducible error at training time? How about test time?

10 Test Test Test

Your friend from Stanford wants use the test set to tune parameters in her model to get a better accuracy on the test set. Is this bad practice? Explain why. If it is bad practice, what is a better solution?

11 K-Fold Cross Validation Mechanical

How do you perform k-fold cross validation? How do you compute validation accuracy?