Project T Final

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

Quiz Solutions

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.

Solution:

Training, Test, Validation, Small, Variance, Bias

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?

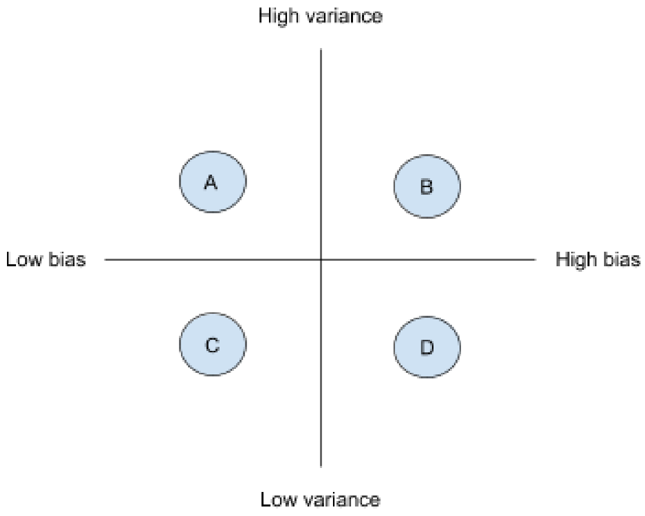
Solution:

Mike's training and testing workflow can look as follows:

1. Visualize the data to sanity check that trends can exist in it
2. Split the data into train and test sets
3. Choose a model to use, based on the shapes of data seen in visualization
4. Separate out a validation set from the train set then tune your model formulation based on validation accuracy
5. Find test accuracy and wrap up the model with a GUI for use

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.

Solution:

- Model 1: A. It seems the model is very sensitive to small fluctuations.
- Model 2: D. The model displays erroneous assumptions while ignoring other features.
- Model 3: C. This is probably from irreducible error. It does well otherwise.
- Model 4: B. The model has erroneous assumptions and is very dependent on its training set.

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?

Solution:

Low variance is more important because he will be dealing with very noisy data; the model should be robust to noise in its training set. Although low bias would also be nice, a low bias + high variance model would most likely do worse than a high bias + low variance model in this case.

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?

Solution:

The training accuracy can be arbitrarily high because we can always increase the dimensionality of the model to overfit to the training set. The test accuracy, however, is probably going to be low because none of the features logically predict the proportion of green-eyed babies, so the model cannot do well on new data. A model cannot magically find patterns where there are none.

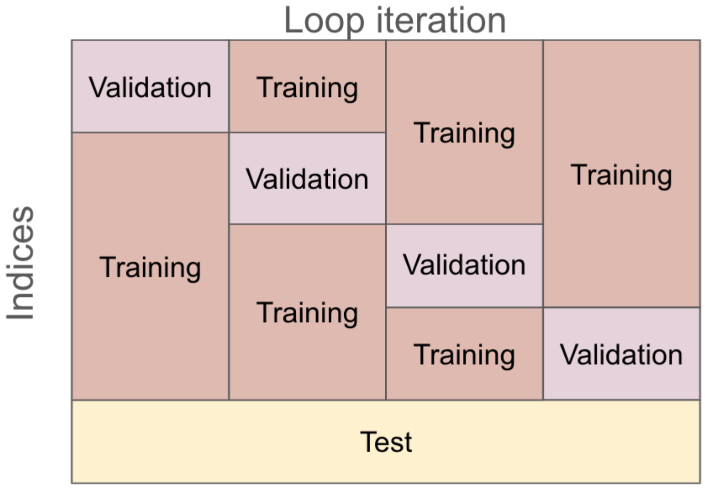
6 The Divided Trio

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
```

Solution



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?

Solution:

Since his training data is very limited (only 60% of an already small dataset!), his model is likely to overfit on the training data. This will lead to bad generalization even if he uses a validation set. So training accuracy can be high, but test accuracy will be much lower than that. This means the training procedure and the way the dataset is created will make the model have high bias and high 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.

Solution:

The derivation is in the notes pdf. This problem acts as a sanity check for whether students have learned how to manipulate probabilistic equations.

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?

Solution:

Remember that the observation of the label Y, so the inherent noise in the labels Z, are only used at test time. Thus, the irreducible error at training time is 0. There is nothing you can say about Z or irreducible error at test time unless you use a Gaussian prior or the like.

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?

Solution:

It is bad practice! The test set is used to see whether your model can generalize to data it has not seen before before shipping it out to the real world. If you tune your model on the test set, you lose that sanity check in your workflow. A better solution would be to create a validation set from the training set, or perform k-fold cross validation if data is limited.

11 K-Fold Cross Validation Mechanical

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

Solution:

Set K to be some number, for example 4. Split the training set into K folds. Pick one of them to be the validation set. Train the model on the rest of the folds and calculate accuracy from the validation set. Then do the same for each of the other K folds. The validation accuracy is the average of the K accuracy metrics.