Model Selection: Bias Variance Tradeoff and Cross Validation

Astrini Sie

asie@seattleu.edu

Week 6a

ECEGR4750 - Introduction to Machine Learning
Seattle University

October 24, 2023

Recap and Updates

- Paper Presentation this Tuesday and Thursday
- Office Hours
  - T, Th 12-1p at Bannan 224
  - W 7-9p via Zoom
  - F 9-9.45a via Zoom
- Zoom Link: `https://seattleu.zoom.us/j/7519782079?pwd=cnhCM2tPcHJKVWwxZVArS2VHSUNJZz09`
  - Meeting ID: 751 978 2079
  - Passcode: 22498122
- Dimensionality Reduction - Feature Selection (continued.)

Overview

A Machine Learning Process

Bias Variance Tradeoff

Bias Variance Tradeoff

Which model is better?

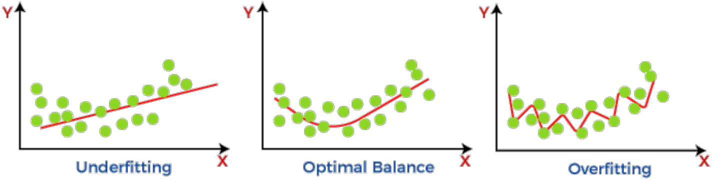
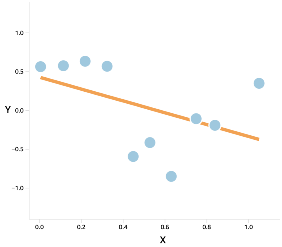


Image source

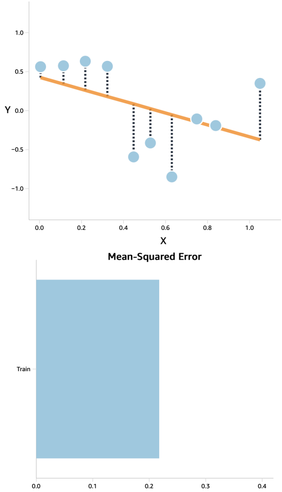
A very simple model

Image source



A very simple model

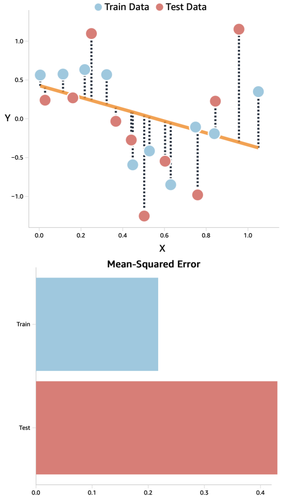
Image source



- Isn't the best at modeling the relationship!
- Mean-squared-error (MSE) of the training data is huge
- Room for improvement!
- **Rule of thumb: Tune model until training error drops to a desired (low value)**

A very simple model

Image source



- MSE of the testing data is even larger!
- **Model is underfitting**
- It is too simple it fails to adequately capture the relationship in between the features x and the labels y

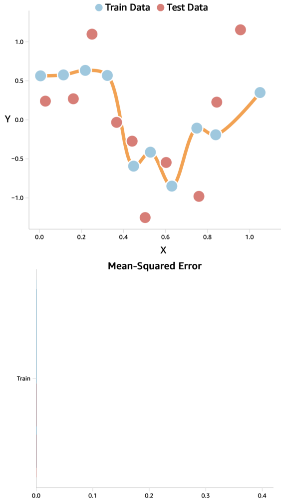
A very complex model

Image source

- The most complex model: predicts every point in the training data perfectly

A very complex model

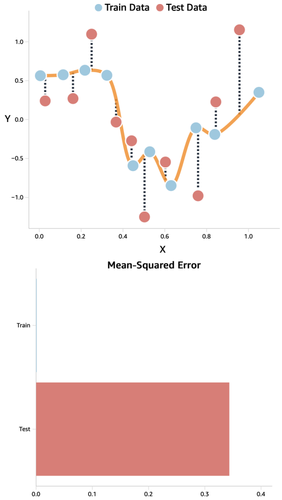
Image source



- MSE of the training data is zero!
- Is this as good as it seems?

A very complex model

Image source



- MSE of the testing data is high!
- **Model is overfitting**
- Instead of learning the true relationship between the features x and labels y, it memorized the noise
- Not able to generalize to unseen data
- **Rule of thumb: Train model such that both training and testing errors are low**

Error Decomposition

Testing error can be decomposed into three components:

Error

$$Error = Bias^2 + Variance + Noise$$

$$Error(x) = \left(\mathbf{E}[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - \mathbf{E}[\hat{f}(x)]\right)^2\right] + Noise$$

Remember from Lab 1 exercise, there's very little we can do about the Noise contribution of the error (*Irreducible Error*), but we can certainly reduce the error contributions from the bias and the variance terms.

Bias

The difference between the average prediction and the true value:

Bias

$$Bias^2(x) = \left(E[\hat{f}(x)] - f(x)\right)^2$$

where $E[\hat{f}(x)]$ is the average prediction of the model over several iterations of independent datasets.

Bias

The difference between the average prediction and the true value:

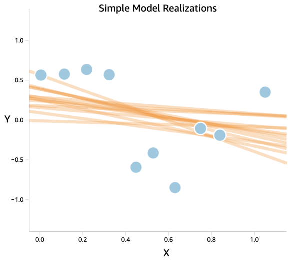
Bias

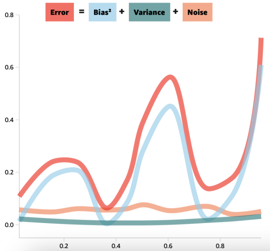
$$Bias^2(x) = \left(E[\hat{f}(x)] - f(x)\right)^2$$

where $E[\hat{f}(x)]$ is the average prediction of the model over several iterations of independent datasets.

In simple, underfit models, majority of the error comes from the bias term!

Image source





Variance

How much, on average, predictions vary for a given data point:

Variance

$$Var(x) = E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

where $E[\hat{f}(x)]$ is the average prediction of the model over several iterations of independent datasets.

Variance

How much, on average, predictions vary for a given data point:

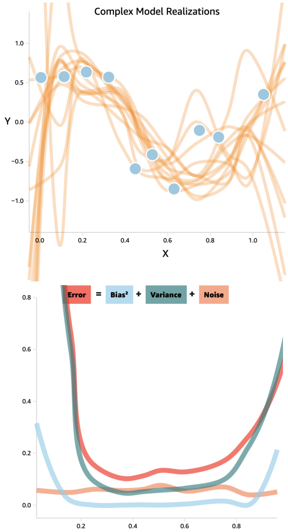
Variance

$$Var(x) = E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right]$$

where $E[\hat{f}(x)]$ is the average prediction of the model over several iterations of independent datasets.

In complex, overfit models, majority of the error comes from the variance term! Unseen data points will be predicted with high error.

Image source



Finding a balance – Bias Variance Tradeoff

No free lunch theorem!
We can't have both low bias **and** low variance.

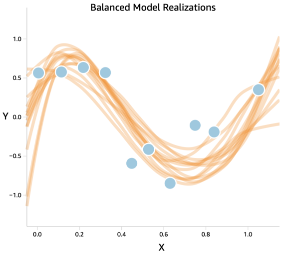
By trading some bias for variance, and without going overboard we can find a balanced model for our dataset.

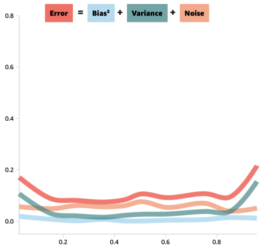
Finding a balance – Bias Variance Tradeoff

No free lunch theorem!
We can't have both low bias **and** low variance.

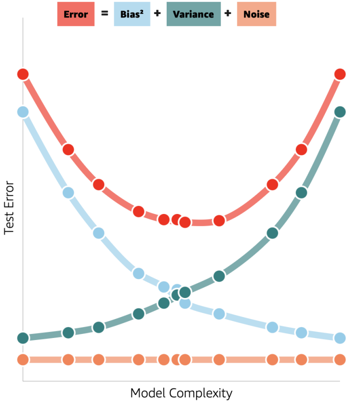
By trading some bias for variance, and without going overboard we can find a balanced model for our dataset.

Image source

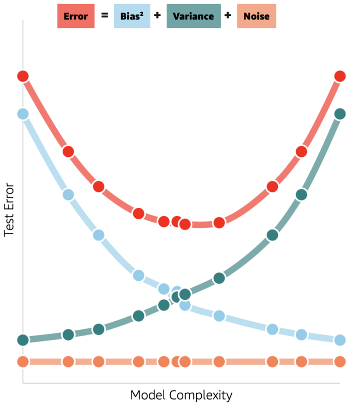




Finding a balance – Bias Variance Tradeoff



Finding a balance – Bias Variance Tradeoff



At the sweet spot where the Bias and the Variance curve intersects, the error is the lowest. This is called the **number of iteration or epoch**, and is one of the hyperparameters of the model.

Model Selection

The importance of data splitting

How do we know what hyperparameters to pick?

- Features selected
- Model Hyperparameters
  - Learning Rate *alpha*
  - Epoch *e*

Data splitting!

The importance of data splitting

How do we know what hyperparameters to pick?

- Features selected
- Model Hyperparameters
  - Learning Rate *alpha*
  - Epoch *e*

Data splitting!

- **Training Set**. Learn patterns and relationships between features and labels.
- **Validation Set**. Understand model performance across various hyperparameters.
- **Test Set**. Once we got the best model, evaluate the model performance for real data in the wild.

Taking it one step further...

How do we select the best model based on all the hyperparameters?

Taking it one step further...

How do we select the best model based on all the hyperparameters?

Cross Validation

Typical Train, Validation, Test Split

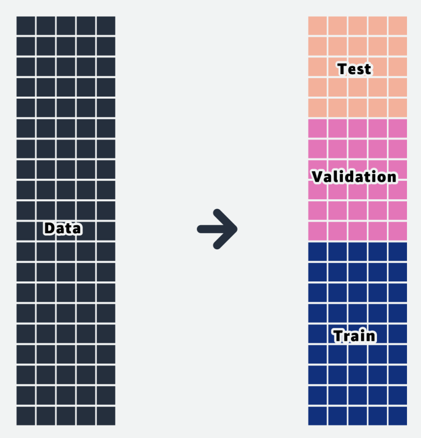


Image source

k-Fold Cross Validation

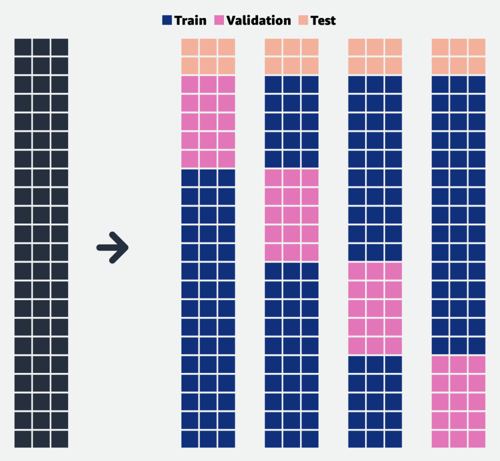


Image source

Leave One Out Cross Validation (LOOV)

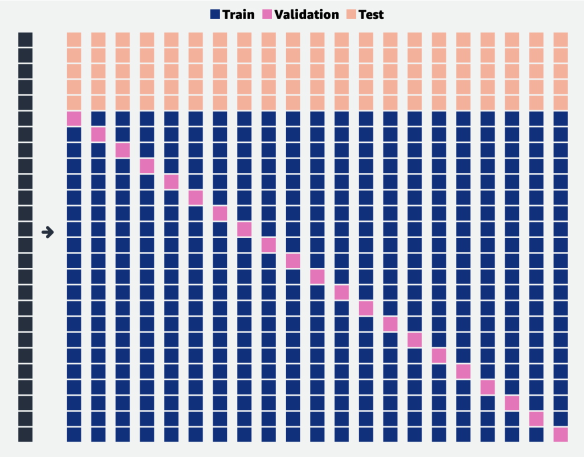


Image source

Illustration

Image source

Illustration

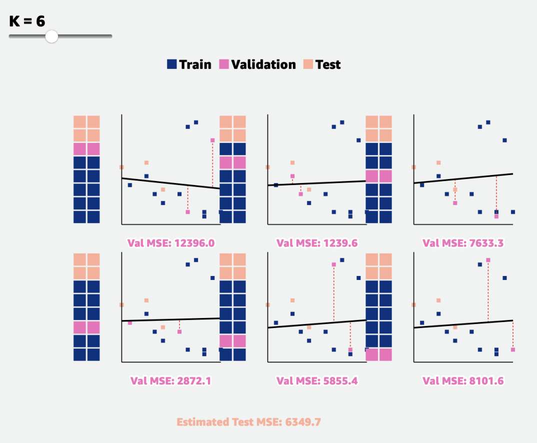


Image source

Illustration

Image source

Illustration

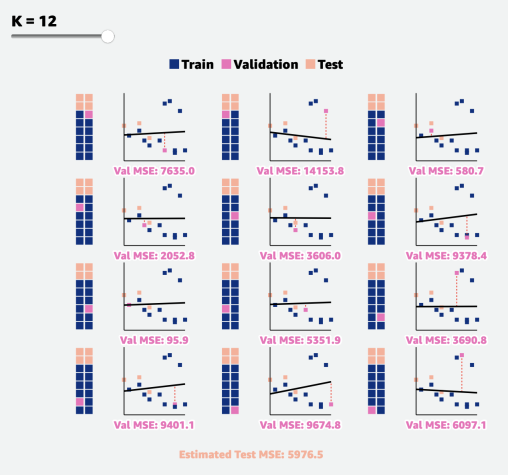


Image source

Split Proportion Selection

How do we decide what's the train, validation, test split proportion?
How do we decide what k should be?

Depends on the size of your data! If your data is small, making train size too small can cause too high of a bias!

References

MLU-Explain
https://mlu-explain.github.io/bias-variance/