Lecture 2: Assessing ML Model Accuracy

Beidi Qiang

SIUE

Training and Loss

In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss.

- Assume that there is some relationship between the label Y and features $\mathbf{x}$. Say
$$Y = f(\mathbf{x}) + \epsilon$$
- Training a model simply means learning (determining) good values for all model parameters in f.
- The result of training can be expressed as an estimate of the function $f(\mathbf{x})$, denoted as $\hat{f}$. which takes features $\mathbf{x}$ as input and that generates a prediction $\hat{Y}$.
- Loss is the penalty for a bad prediction. That is, loss is a number indicating how bad the model's prediction was on a single.

Reducible error and irreducible error

The accuracy of a prediction for Y depends on two quantities, which we will call the reducible error and the irreducible error.

- $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error. This error is reducible because we can potentially improve the accuracy of $\hat{f}$
- variability associated with ϵ also affects the accuracy of our predictions. This is known as the irreducible error, because no matter how well we estimate f, we cannot reduce the error.

Split the Data

A machine learning model aims to make good predictions on new, previously unseen data. To see how model performs on unseen data, we introduced the idea of dividing your data set into training and test. Our test set serves as a proxy for new data.

- Training set—a subset to train a model.
- Test set—a subset to test the model.

Make sure that your test set meets the following conditions:

- Is large enough to yield statistically meaningful results.
- Is representative of the data set as a whole. In other words, don't pick a test set with different characteristics than the training set.

Basic assumptions of ML models

The following are basic assumptions when collecting data and train the ML model:

- We draw examples independently and identically (i.i.d) at random from the distribution. This is usually achieved by a simple random sample from the population.
- The distribution is stationary; that is the distribution doesn't change within the data set.
- We draw examples from partitions from the same distribution.

Assessing Model Accuracy: Regression

- In the regression setting, the most commonly-used measure for loss is the mean squared error:

$$MSE = \frac{1}{n}\sum(y_i - \hat{f}(x_i))^2.$$

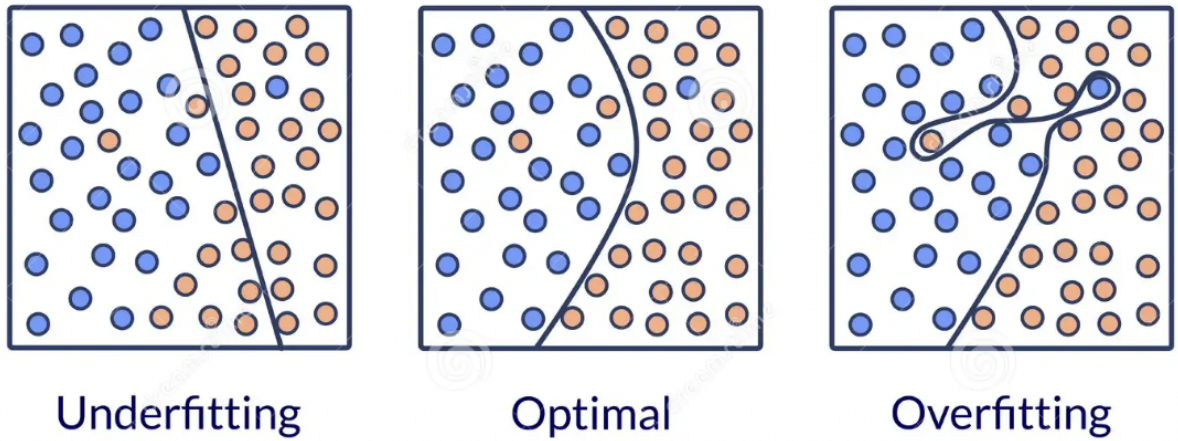
- The MSE computed using the training data is referred to as the training MSE.
- The MSE computed using the testing data (previously unseen) is referred to as the testing MSE.
- We want to choose the method that gives the test MSE as small as possible.

Assessing Model Accuracy: Classification

- In the classification setting, the most commonly-used measure is the error rate:
$$\frac{1}{n}\sum I(y_i \neq \hat{f}(x_i)).$$
- The error rate computed using the training data is referred to as the training error. The error computed using the testing data is referred to as the testing error.
- A good classifier is one for which the test error is smallest..

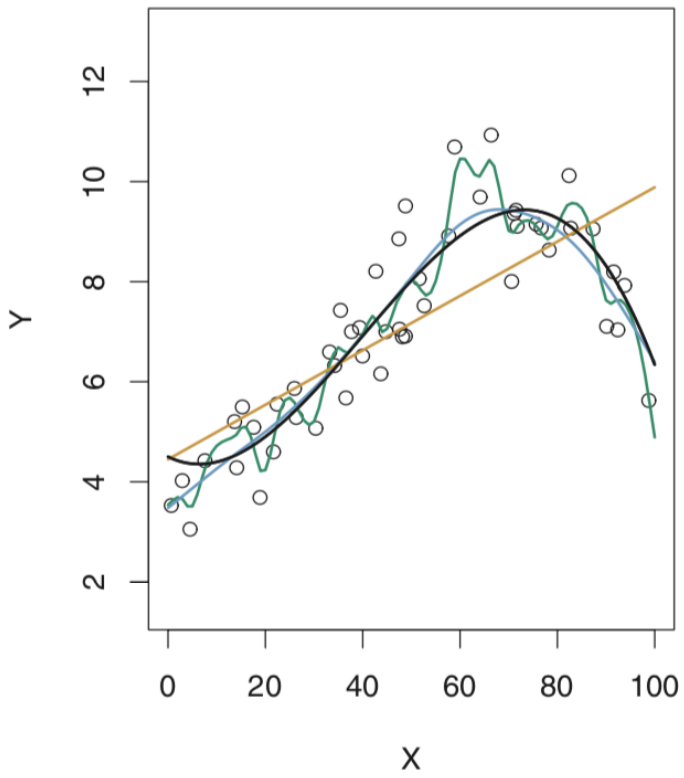
Peril of Overfitting

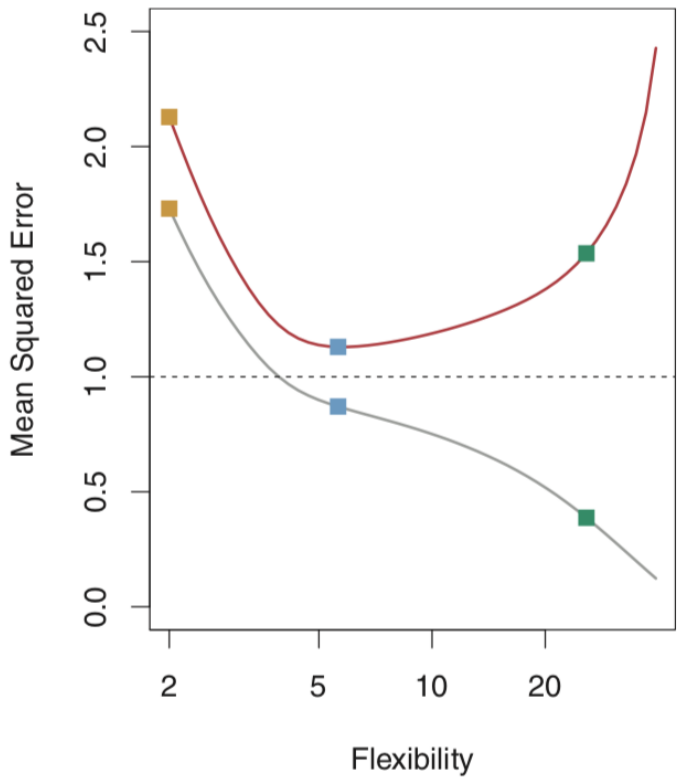
Overfitting occurs when a model tries to fit the training data so closely that it does not generalize well to new data.



Model Complexity, simulated example

In the follow plot, we have data simulated from a distribution f (black). Three estimates of f are based on the data are shown: the linear regression line (orange), and two higher order polynomial fits (blue and green curves). Training data MSE (grey), test MSE data (red). It is clear that as the level of flexibility (complexity) increases, the curves fit the observed data more closely, but not necessarily the true f.





The Bias-Variance Trade-Of

The U-shape observed in the test MSE curves turns out to be the result of two competing properties:

$$\mathrm{E}\left[y_0 - \hat{f}(x_0)\right]^2 = \mathrm{Var}\left(\hat{f}(x_0)\right) + \left[Bias(\hat{f}(x_0))\right]^2 + \mathrm{Var}\left(\epsilon\right)$$

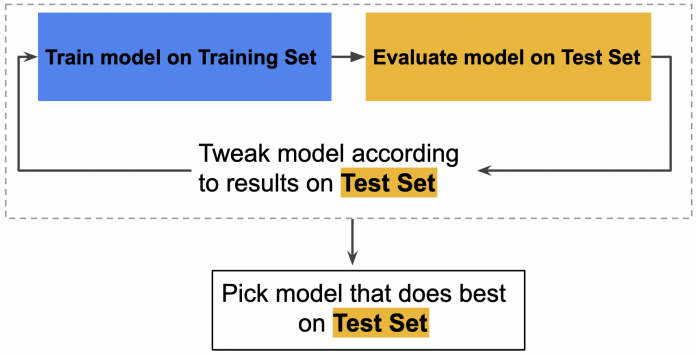
- $\mathrm{E}\left[y_0 - \hat{f}(x_0)\right]^2$ is the expected test MSE.
- $Bias(\hat{f}(x_0)) = \mathrm{E}\left[y_0 - \hat{f}(x_0)\right]$.
- To minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias.

The Bias-Variance Trade-Of

- Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. In general, more flexible statistical methods have higher variance.
- Bias refers to the error in estimating y that is introduced by using simplified model. It is unlikely that any real-life problem truly can be represented by any simple model. Generally, more flexible methods result in less bias.
- As we use more flexible methods, the variance will increase and the bias will decrease. The relative rate of change of these two quantities determines whether the test MSE increases or decreases.

Model Tweak and Another Partition

Partitioning a data set into a training set and a test set enabled you to train on one set of examples and then to test the model against a different set of examples. With two partitions, the workflow could look as follows:



You can greatly reduce your chances of overfitting by partitioning the data set into the three subsets by introducing a validation set to tweak your model. This is a better workflow because it creates fewer exposures to the test set.