An Introduction to Statistical Learning with Applications in R

1/29/23 Day 1

Today. I'm starting by reading through §2.1

1/31/23 Day 2

2.2 Assessing Model Accuracy

There is no free lunch in statistics: No one method dominates all others over all possible data sets.

2.1.5 Measuring the Quality of Fit

The mean squared error of a model $\hat{f}$ predicting a dataset $\{(x_i, y_i)\}_{i=1}^n$, given by

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

is the most common measure to quantify how well the model predicts the data.

- MSE is calculated on the training data, hence it is better called training MSE.
- We are actually interested in test MSE.

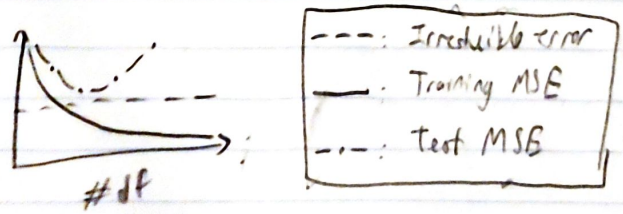
How to minimize test MSE?

- Use testing data if you have it.

If a model has small training MSE but large test MSE, you're overfitting. Overfitting specifically refers to the case when a less flexible model would have had smaller test MSE.

Will discuss cross-validation later on

MSE as a function of degrees of freedom.



2.2.2: The Bias Variance Trade-Off

The U-shape of test MSE curves turns out to be due to two competing properties of statistical learning methods.

It can be shown that

$$\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \text{Var}(\hat{f}(x_0)) + [\text{Bias}(\hat{f}(x_0))]^2 + \text{Var}(\varepsilon).$$

(The expected test MSE for x_0.

- Need to achieve low variance and low bias.

Variance is the amount $\hat{f}$ would change if estimated on a different training set.

- More flexible SL methods have higher variance.

Bias refers to the error introduced by approximating a real-life problem with a simpler model.

As flexibility increases bias decreases.

The bias-variance trade-off is one of the most important themes in this book.

2.2.3. The Classification Setting

Estimate f from training data $\{(x_i, y_i)\}_{i=1}^n$, where y_i is qualitative.

We seek to minimize the training error rate

$$1/n \sum_{i=1}^n I(y_i \neq \hat{y}_i)$$

We're actually interested in the test error rate

$$\text{Avg}(I(y_0 \neq \hat{y}_0)),$$

and we want to minimize that.

The Bayes Classifier

- Can prove that the test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class given its predictor values, i.e. to class j where

$$P(Y=j|X=x_0)$$

is largest.

- This very simple classifier is called the Bayes classifier. The lowest possible error rate is the Bayes error rate.

The error rate at $X=x_0$ is related to the fact that

$$P(Y\neq j|X=x_0) = 1 - P(Y=j|X=x_0).$$

The error rate is $1 - \max_j P(Y=j|X=x_0)$.

K-Nearest Neighbors

Given $K \in \mathbb{N}$, test observation x_0, KNN first identifies the K points in the training data closest to x_0, represented by N_0. It then estimates the conditional probability for class j as a fraction of points in N_0 whose response values equal j.

$$P(Y=j|X=x_0) = \frac{1}{K}\sum_{i\in N_0} I(y_i=j).$$

Finally, KNN classifies x_0 to the class with the largest probability.

- KNN is often pretty close!
- The choice of K has a big effect.
  - For small K, the decision boundary may be too flexible, causing overfitting.
  - For large K, have low-variance, high bias.

2.3 Lab: Introduction to R