MATH560 HW 2

Jared Andreatta

2025-03-19

library(ggplot2)  
library(tidyr)

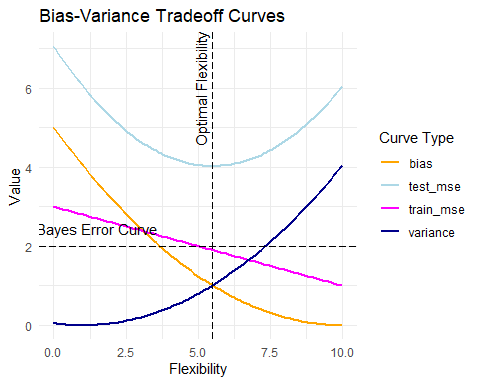
# Section 2.4

## Problem 3

### a.

x <- seq(0,10, length.out=100) # x-axis  
variance <- (x-1)^2 / 20 # variance  
bias <- (10-x)^2 / 20 # Bias  
test\_mse <- variance + bias + 2 # test MSE  
train\_mse <- 3 - 0.2 \* x   
  
df <- data.frame(x, test\_mse, train\_mse, variance, bias)  
  
df\_long <- pivot\_longer(df, cols = c(test\_mse, train\_mse, variance, bias),  
 names\_to = "curve", values\_to = "value")  
  
ggplot(df\_long, aes(x = x, y = value, color = curve)) +  
 geom\_line(size = 1) +   
 geom\_vline(xintercept = 5.5, linetype = "longdash") +  
 annotate("text", x = 5.5, y = 6,   
 label = "Optimal Flexibility",   
 angle = 90, vjust = -0.5) +  
 geom\_hline(yintercept = 2, linetype = "longdash") +  
 annotate("text", x = 1.5, y = 2,   
 label = "Bayes Error Curve",   
 vjust = -1) +  
 labs(title = "Bias-Variance Tradeoff Curves",  
 x = "Flexibility",  
 y = "Value",  
 color = "Curve Type") +  
 scale\_color\_manual(values = c("orange", "lightblue", "magenta", "darkblue")) +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



### b.

* Bias is downward sloping because rigid models tend to provide biased estimates that consistently under/overestimate the true relationship of the data.
* Variance is upward sloping because more flexible models tend to overfit on noise and cannot generalize as well to unseen data.
* The test MSE follows a quadratic shape because it is minimized at the optimal level of bias and variance in a model. Too much bias or too much variance can cause the test MSE to increase.
* The train MSE curve is monotonically decreasing; as model complexity increases, it will fit the to the training set better as the model complexity increases.
* The Bayes error curve is a constant line that increases the test MSE because it is an intrinsic error that cannot be reduced, no matter the complexity of the model.
* The Optimal Flexibility line is the desired model flexibility that minimizes the test MSE, which is a single point on the test MSE curve.

# Section 5.4

## Problem 3

### a.

Here is a short rundown of how k-Fold CV works:

1. Initialize first fold as a validation set.

2. Train on remaining folds and validate on the held-out fold.

3. Compute where denotes the fold being held-out as a validation set.

4. Repeat this times with each iteration using a different group of observations as a validation set.

Essentially, it looks like this mathematically

### b.

#### k-Fold vs Validation Set

The advantages of k-Fold CV is that the estimates are more accurate, as the estimates are averaged out over folds instead of 1 random sample. Also, k-Fold doesn’t contain all of the data to one hold-out set.

The disadvantage of k-Fold over Validation Set CV is the computational cost, as it requires more iterations than the simple validation set approach.

#### k-Fold vs LOOCV

One big advantage over LOOCV is the computational cost; LOOCV is a special case where , but when , it is easier to compute, and often it is unnecessary to estimate the MSE times. Also, the MSE estimate from k-Fold has less variance since it leaves out more than one observation, which gives a more stable error estimate.

The disadvantage is that LOOCV is more accurate (in theory) in estimating MSE. k-Fold also gives a more biased estimate of MSE since k-Fold uses smaller training sets.

## Problem 8

### a.

set.seed(1)  
 x <-rnorm(100)  
 y <- x - 2 \* x^2 + rnorm(100)

In this data, and . The equation is