Weekly Exercise - Week 1

Benedikt Sojka

2024-02-11

Write a function that implements the following pseudo code:

```
set.seed(42)
pseudo = function(sample_size) {
  x = runif(n = sample_size, min = -3, max = 3)
  y = 8*sin(x) + rnorm(sample_size)
  return(data.frame(x, y))
}
```

Use this function to generate a training set of size 50 and a test set of size 10000.

```
train = pseudo(50)
test = pseudo(10000)
```

Train two polynomial regression models, one with a degree of 3 and one with a degree of 15, and estimate their mean squared errors (MSE) on the test set. Repeat the process with a training set of 10000.

```
fit3 = lm(y ~ poly(x, degree = 3), data = train)
train_pred_3 = predict(fit3, newdata = train)
test_pred_3 = predict(fit3, newdata = test)
train_err_3 = mean((train$y - train_pred_3)^2)
test_err_3 = mean((test$y - test_pred_3)^2)

fit15 = lm(y ~ poly(x, degree = 15), data = train)
train_pred_15 = predict(fit15, newdata = train)
test_pred_15 = predict(fit15, newdata = test)
train_err_15 = mean((train$y - train_pred_15)^2)
test_err_15 = mean((test$y - test_pred_15)^2)

train2 = pseudo(10000)
test2 = pseudo(10000)

fit3.1 = lm(y ~ poly(x, degree = 3), data = train2)
train_pred_3.1 = predict(fit3.1, newdata = train2)
```

```
test_pred_3.1 = predict(fit3.1, newdata = test2)
train_err_3.1 = mean((train2$y - train_pred_3.1)^2)
test_err_3.1 = mean((test2$y - test_pred_3.1)^2)

fit15.1 = lm(y ~ poly(x, degree = 15), data = train2)
train_pred_15.1 = predict(fit15.1, newdata = train2)
test_pred_15.1 = predict(fit15.1, newdata = test2)
train_err_15.1 = mean((train2$y - train_pred_15.1)^2)
test_err_15.1 = mean((test2$y - test_pred_15.1)^2)

test_err_3; test_err_15; test_err_3.1; test_err_15.1
```

```
## [1] 1.258852
```

```
## [1] 1.632571
```

```
## [1] 1.205183
```

```
## [1] 1.00663
```

What is the best possible prediction rule f in this case? Obtain the test MSE of the best prediction rule as well.

```
test_err_15.1
```

```
## [1] 1.00663
```

The best predicition rule is the one that, based on the training data, reecords the lowest MSEpred. In this case, it is the polynomial of degree 15 in the case of 10.000 training data. The value of the MSE is "1.00663" as one can see above.

Report the 4 obtained test mean squared error values (for degree 3 and 15 and for both training set sizes). Explain the obtained numbers using bias and variance. Hints: Start with the results for the large training set, taking the optimal MSE into account as well can help.

```
test_err_3; test_err_15; test_err_3.1; test_err_15.1
```

```
## [1] 1.258852
```

```
## [1] 1.632571
```

```
## [1] 1.205183
```

```
## [1] 1.00663
```

As mentioned before, the lower the MSE (which is determined by Bias and Variance), the better the predictor. In a perfect world, both Bias and Variance would be low, in reality though, there is usually a trade-off between these two. A good predictor should be flexible enough to capture the relationship in the data (low bias), but on the other side not too flexible, because than it might perform poorly on new data (high variance). For the large training set, the higher degree polynomial performs better, as the structure in the data is more complex and therefore a higher degree is needed. In the lower training set, however, we can see that the 3 degree

polynomial performs better. It seems to be sufficient to explain the structure in the training data, while being less flexible, which allows is to perform better on new data compared to the 15 degree polynomial.