Homework 3 - Machine Learning

Felix Nguyen

February 2019

1 Exercise 1

- The best estimator is not necessarily unbiased, since this depends on various factors, such as the assumptions about population, the availability of data, unbiased estimator is difficult to compute, or biased estimator gives lower loss function value... Ideally, an unbiased estimator would be preferable, but the data we usually work with is rarely ideal. For example, standard deviation of a population is normally biased since it goes through non-linear transformation from variance, and the unbiased estimator in this case is difficult to compute and doesn't worth the hassle.
- Explain the relationship between MSE, bias, variance, and irreducible error of a prediction: While all of these terms describe errors in prediction, each describe different components of errors. The bias is an error from erroneous assumptions in the learning algorithm, while the variance is an error from sensitivity to small fluctuations in the training set. MSE is a measurement that represents both of these errors, which we can call reducible error. Irreducible error is the remaing part of error which is not accounted for by MSE.

2 Exercise 2

We have:

$$\begin{aligned}\sum_{i=1}^{n} Cov(\hat{y_i}, y_i) &= trace(Cov(\hat{y}, y)) = trace(Cov(X\hat{\beta}, y)) \\ &= trace(Cov(X(X^TX)^{-1}X^Ty, y) \\ &= trace(X(X^TX)^{-1}X^TCov(y, y)) \\ &= trace(X(X^TX)^{-1}X^TVar(y))\end{aligned}$$

Since $X(X^TX)^{-1}X^T = I$, with I is a K x K identity matrix, thus $trace(I) = K$, and $Var(y) = Var(\epsilon) = \sigma^2$, we then have:

$$\begin{aligned}\sum_{i=1}^{n} Cov(\hat{y_i}, y_i) &= \sigma^2 trace(I) \\ &= K\sigma^2\end{aligned}$$

3 Exercise 3

For $X < \xi_1$, we have $f(X) = \sum_{j=0}^{3} \beta_j X^j$. However, for natural cubic splines, this part must be linear, so we can say that $\beta_2 = \beta_3 = 0$.

Similarly, for $X > \xi_K$, the function form is:

$$
\begin{aligned}
f(X) &= \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X - \xi_k)^3 \\
&= \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \theta_k (X^3 - 3X^2\xi_k + 3X\xi_k^2 - \xi_k^3) \\
&= X(\beta_1 + 3\sum_{k=1}^{K} \theta_k \xi_k^2) + X^2(\beta_2 - 3\sum_{k=1}^{K} \theta_k \xi_k) + X^3(\beta_3 + \sum_{k=1}^{K} \theta_k) + \beta_0 - \sum_{k=1}^{K} \theta_k \xi_k^3
\end{aligned}
$$

This part must also be linear, so we can say that $\beta_2 - 3\sum_{k=1}^{K} \theta_k \xi_k$ and $\beta_3 + \sum_{k=1}^{K} \theta_k$ is 0. Since $\beta_2 = \beta_3 = 0$, this implies $\sum_{k=1}^{K} \theta_k \xi_k = 0$ and $\sum_{k=1}^{K} \theta_k = 0$.

We can see that β_0 and β_1 are the coefficients of N_1 and N_2. For N_{k+2}, we have:

$$
\begin{aligned}
\sum_{k=1}^{K} \theta_k (X - \xi_k)_+^3 &= \sum_{k=1}^{K-2} \theta_k (X - \xi_k)_+^3 + \theta_K (X - \xi_K)_+^3 + \theta_{K-1}(X - \xi_{K-1})_+^3 \\
= \sum_{k=1}^{K-2} \theta_k (X - \xi_k)_+^3 + (\theta_K + \theta_{K-1})(X - \xi_K)_+^3 &- \frac{-\theta_K \xi_{K-1} + \theta_{K-1}\xi_{K-1}}{\xi_K - \xi_{K-1}} [(X - \xi_{K-1})_+^3 - (X - \xi_K)_+^3)]
\end{aligned}
$$

From the constraints, we have: $\theta_K + \theta_{K-1} = -\sum_{k=1}^{K-2} \theta_k$ and $\theta_K \xi_K + \theta_{K-1}\xi_{K-1} = -\sum_{k=1}^{K-2} \theta_k \xi_k$. We can transform: $\theta_K \xi_K + \theta_{K-1}\xi_{K-1} = \theta_K \xi_K + \theta_K \xi_{K-1} - \theta_K \xi_{K-1} + \theta_{K-1}\xi_{K-1} = -\sum_{k=1}^{K-2} \theta_k \xi_K - \theta_K \xi_{K-1} + \theta_{K-1}\xi_{K-1}$. So we have the transformed function:

$$
\begin{aligned}
\sum_{k=1}^{K-2} \theta_k (X - \xi_k)_+^3 - \sum_{k=1}^{K-2} \theta_k (X - \xi_K)_+^3 &+ \frac{\sum_{k=1}^{K-2} \theta_k \xi_K - \sum_{k=1}^{K-2} \theta_k \xi_k}{\xi_K - \xi_{K-1}} [(X - \xi_{K-1})_+^3 - (X - \xi_K)_+^3)] \\
&= \sum_{k=1}^{K-2} (\xi_K - \xi_k)\theta_k \Big(\frac{(X - \xi_k)_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_k} - \frac{(X - \xi_{K-1})_+^3 - (X - \xi_K)_+^3}{\xi_K - \xi_{K-1}} \Big)
\end{aligned}
$$

So, we can say that for N_{k+2} the corresponding coefficient is $(\xi_K - \xi_k)\theta_k$ with the basis $N_{k+2} = d_k(X) - d_{K-1}(X)$ as given.