NNTI: Exercise Sheet 4

Team #25: Camilo Martínez 7057573, Honglu Ma 7055053

November 30, 2023

Problem 4.1 - Bias and Variance

(a)

If $\hat{f}$ is the predicted function by our linear regression model, then the variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\hat{f}$. But ideally the estimate for f should not vary too much between training sets. In other words, high variance implies that small changes in the training data can result in large changes in our predicted function $\hat{f}$. In general, more flexible statistical methods have higher variance. On the other hand, bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model. For example, linear regression assumes that there is a linear relationship between Y and a set of predictors X_i. One could say that the model is biased to consider that given a set of x_i is proportional or linearly related to a set of outputs y_i. Generally, more flexible methods result in less bias. More flexible methods result in an increase in variance and a decrease in bias. Bias indicates how well the model predicts and Variance is a measurement of how similar the models trained with different training set behaves. A complex model results in a low bias and a high variance and a simple model results in a high bias and a low variance [1].

(b)

Considering the concepts introduced in the explanation above, overfitting means low bias but high variance. On the other hand, underfitting means high bias but low variance.

(c)

To show that the given equation holds, we have to consider the mathematical definition of bias and variance of a function f, $Bias(f)$ and $Var(f)$ [2] [3]:

$$Bias(f(x)) = E[f(x)] - f(x)$$

$$Var(f(x)) = E[(f(x) - E[f(x)])^2]$$

With these definitions and using the properties of the expected value of a function, represented by $E[f(x)]$ we can perform the following derivation:

$$\begin{aligned}
MSE(y, \hat{f}) &= E[(y - \hat{f}(x_0))^2)] \\
&= E[y^2 - 2y\hat{f} + \hat{f}^2] \\
&= E[y^2] - 2E[y\hat{f}] + E[\hat{f}^2] \\
&= E[(f + \varepsilon)^2] - 2E[(f + \varepsilon)\hat{f}] + E[\hat{f}^2] \\
&= E[f^2] + 2E[f]E[\varepsilon] + E[\varepsilon^2] - 2(E[f\hat{f}] + E[\varepsilon]E[\hat{f}]) + E[\hat{f}^2] \\
&= f^2 + 2f \cdot 0 + Var(\varepsilon) - 2E[f]E[\hat{f}] - 2 \cdot 0 \cdot E[\hat{f}] + E[\hat{f}^2] \\
&= f^2 + Var(\varepsilon) - 2fE[\hat{f}] + E[\hat{f}^2] \\
&= f^2 + Var(\varepsilon) - 2fE[\hat{f}] + E[\hat{f}^2] + E[\hat{f}]^2 - E[\hat{f}]^2 \\
&= (f^2 - 2fE[\hat{f}] + E[\hat{f}]^2) + (E[\hat{f}^2] - E[\hat{f}]^2) + Var(\varepsilon) \\
&= (f - E[\hat{f}])^2 + (E[\hat{f}^2] - E[\hat{f}]^2) + Var(\varepsilon) \\
&= Bias(\hat{f})^2 + Var(\hat{f}) + Var(\varepsilon)
\end{aligned}$$

That is, the MSE can be decomposed into the sum of squared bias, variance, and irreducible error. This decomposition is the bias-variance tradeoff. It means that in order to minimize the MSE, we need to select a statistical learning method that simultaneously achieves that our $\hat{f}$ has low variance and low bias (it is a tradeoff ultimately). Note that variance is inherently a nonnegative quantity, and squared bias is also nonnegative [1].

(d)

When the training set size goes up, the variance goes down. This can be intuitively derived, since more data generally means that our models can get more robust, since the model now has more data to work with and derive the intricacies in-between. Mathematically speaking, the variance is inversely proportional to the training size N. On the other hand, bias will remain the same even if the training set size increases. This is because bias is inherently correlated with the model's degrees of freedom or complexity and is therefore not related to the training size. We see this in the plot of the final exercise of the practical problem. Nevertheless, practically speaking, it is still true that a small training set limits the optimal capacity of our models, meaning we cannot train more complex models, even if we would like to do so. This was seen in the slide 30/50 of Chapter 4.

Problem 4.2 - Maximum Likelihood Estimate (MLE)

(a)

Let the output variable $y = y_1, ..., y_m$ consist of m i.i.d. normal variables and has likelihood

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) = \prod_{m=1}^{M} \mathcal{N}(\boldsymbol{y_m}|\boldsymbol{w}^\mathsf{T}\boldsymbol{x_m}, \sigma^2) \tag{1}$$

Where

$$\mathcal{N}(\boldsymbol{y_m}|\boldsymbol{w}^\mathsf{T}\boldsymbol{x_m}, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2\right)$$

To maximize (1), we can take the log which makes it simpler to work with. This is called the log-likelihood and is defined as follows:

$$\log\left(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})\right) = \log\left(\prod_{m=1}^{M} \mathcal{N}(\boldsymbol{y_m}|\boldsymbol{w}^\mathsf{T}\boldsymbol{x_m}, \sigma^2)\right) \tag{2}$$

Problem 4.2 continued on next page...

From there, we take advantage of the properties of the log function, mainly $\log(ab) = \log a + \log b$

$$\begin{aligned}
\log\left(p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})\right) &= \sum_{m=1}^{M} \log\left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2\right)\right) \\
&= \sum_{m=1}^{M} \log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{m=1}^{M} \log\left(\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2\right)\right) \\
&= M\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \sum_{m=1}^{M} -\frac{1}{2\sigma^2}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2 \\
&= -M\log\left(\sigma\sqrt{2\pi}\right) - \frac{1}{2\sigma^2}\sum_{m=1}^{M}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2
\end{aligned}$$

Finally, we arrive at the following expression

$$\log\left(p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{w})\right) = -M\log(\sigma) - \frac{M}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{m=1}^{M}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2 \tag{3}$$

Since $-M\log(\sigma) - \frac{M}{2}\log(2\pi)$ as well as $\frac{1}{2\sigma^2}$ are just constants, we only need to maximize the last term that involves the sum. Moreover, maximizing a negative value is the same as minimizing its positive counterpart. Thus, the entire problem becomes minimizing the following expression

$$\sum_{m=1}^{M}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2$$

From there, we can introduce a new constant $\frac{1}{M}$ which does not affect the minimization process at all, and we get an expression that is the Mean Squared Error, for which we derived on Assignment 3 Exercise 3.1.c the optimal weight vector $\boldsymbol{w}$

$$\text{MSE} = \frac{1}{M}\sum_{m=1}^{M}(\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2$$

This proves that a linear regression procedure that consists of minimizing the MSE can be justified as a Maximum-Likelihood procedure, which was our starting point.

(b)

Assuming a standard normal prior on the weights $\boldsymbol{w}$ of the form $\mathcal{N}(\boldsymbol{w}\,|\,0, \frac{1}{\lambda}\boldsymbol{I})$, where λ is just a constant that defines the precision of the distribution, we can express the likelihood function as follows

$$p(\boldsymbol{w}\,|\,\lambda) = \prod_{m=1}^{M}\mathcal{N}\left(\boldsymbol{w}\,|\,0, \frac{1}{\lambda}\boldsymbol{I}\right) = \left(\frac{\lambda}{2\pi}\right)^{M/2}\exp\left(-\frac{\lambda}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{w}\right)$$

Where M is the total number of elements in the $\boldsymbol{w}$ vector. Using Bayes' theorem, the posterior distribution for $\boldsymbol{w}$ is proportional to the product of the prior distribution and the likelihood function, that is

$$p(\boldsymbol{w}\,|\,\boldsymbol{X},\boldsymbol{y},\lambda) \propto p(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{w},\lambda)\,p(\boldsymbol{w}\,|\,\lambda) \tag{4}$$

Where $p(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{w},\lambda)$ is the same one we introduced in the previous exercise, with an additional parameter λ. Now, to determine the weight vector $\boldsymbol{w}$, we need to find the most probable value of $\boldsymbol{w}$ given the data, in other words by maximizing the posterior distribution given by (4). This technique is called maximum posterior, or simply MAP. As we have already explained in previous exercises, we can approach this problem by taking the log function and maximize that instead. That is, our expression to maximize becomes

$$\log\left(p(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{w},\lambda)\,p(\boldsymbol{w}\,|\,\lambda)\right) = \log p(\boldsymbol{y}\,|\,\boldsymbol{X},\boldsymbol{w},\lambda) + \log p(\boldsymbol{w}\,|\,\lambda) \tag{5}$$

Problem 4.2 continued on next page...

In (5), we can identify that the first term is the same as the one in the previous exercise. And we already proved that maximizing that term is the same as minimizing $\sum_{m=1}^{M} (\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2$ or any constant multiplied by this term. Note that this expression does not depend on σ, which we defined as our precision parameter λ. On the other hand, maximizing the second term $\log p(\boldsymbol{w} \mid \lambda)$ is the same as maximizing $-\frac{\lambda}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{w}$, since $(\frac{\lambda}{2\pi})^{M/2}$ is just a constant. And that in turn is the same as minimizing its negative counterpart. In summary, the entire problem becomes minimizing the following expression

$$\mathcal{L}(\boldsymbol{w}) = \frac{1}{2}\sum_{m=1}^{M} (\boldsymbol{y_m} - \boldsymbol{w}^\mathsf{T}\boldsymbol{x_m})^2 + \frac{\lambda}{2}\boldsymbol{w}^\mathsf{T}\boldsymbol{w} \tag{6}$$

Note that we multiplied the first term by $1/2$ which does not have any impact on the minimization process, since it is just a constant. Finally, we recognize that the resulting expression in (6) is the L_2-regularized least squares loss. Thus we see that maximizing the posterior distribution or MAP procedure is equivalent to minimizing the least squares criterion with (L_2-)regularization, also known as ridge regression in the Statistics literature, with a regularization parameter given by λ, by assuming a standard normal prior on the weights.

Problem 4.3 - Bias-Variance Trade-Off Exploration

See attached .ipynb solution in .zip file.

References

[1] James Gareth et al. "2.2.2 The Bias-Variance Trade-Off". In: *An Introduction to Statistical Learning with Applications in Python*. Springer, 2023, pp. 31–34.

[2] M. Jordan, J. Kleinberg, and B. Schölkopf. "Expectation and Covariances". In: *Pattern Recognition and Machine Learning*. 2006. Chap. 1.2.2, pp. 19–20.

[3] M. Jordan, J. Kleinberg, and B. Schölkopf. "The Bias-Variance Decomposition". In: *Pattern Recognition and Machine Learning*. 2006. Chap. 3.2, pp. 147–152.