INF554 - Machine Learning I

Lab 1: SOLUTIONS

Question 1

In this answer we derive the ordinary least squares estimatorof our linear model parameters. We begin by reexpressing the mean squared error (MSE) in its matrix form.

$$\begin{aligned}
\text{MSE}(\beta) &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - X_i\beta)^2 \\
&= \frac{1}{N}(y - X\beta)^T(y - X\beta) \\
&= \frac{1}{N}\left(y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta\right).
\end{aligned}$$

Now we take the derivate with respect to β,

$$\frac{\partial}{\partial\beta}(MSE(\beta)) = \frac{1}{N}\left(-X^Ty - X^Ty + (X^TX + X^TX)\beta\right).$$

In order to find the optimum we set this derivative to zero and solve for $\hat{\beta}$.

$$\begin{aligned}
& 0 = \frac{1}{N}\left(-2X^Ty + 2X^TX\hat{\beta}\right). \\
\Rightarrow \quad & \hat{\beta} = (X^TX)^{-1}X^Ty. \qquad (1)
\end{aligned}$$

Question 2

Note that in order to invert the matrix X^TX in Equation (1) we have to assume that X is full rank, i.e., that the columns of X are linearly independent. This is known as the full rank assumption of linear models. If our design matrix X violates the full rank assumption then we are unable to invert it for the calculation of $\hat{\beta}$. This can happen either when we have, what is known as, perfect multicolinearity, where a column of X is a linear combination of other columns of X or if we have less data points N than features d.

In fact, the full rank assumption is also necessary to confirm that we indeed have a minimum at the optimum, which we have derived by taking the first derivative. If we take the second derivative of $MSE(\beta)$ we obtain $2X^TX$, this is a semi positive definite matrix if X is full rank and hence we are able confirm that $\hat{\beta}$ indeed minimises the mean squared error.

Question 3

Since in the mean squared error we are considering squared distances outliers which are far away from their predicted value have a large impact on the MSE. If we were to use a different loss function such as the mean absolute error, where absolute differences are used instead of the squared distances, then the impact of outliers would be reduced.

Question 4

Based on Figure 2 in lab1.pdf it seems best to use a polynomial of degree 4. For degrees larger than 4 we observe that the test MSE is larger than the training MSE, which indicates that our linear model is overfitting the training data.