INF554 - Machine Learning I

Lab 1: SOLUTIONS

Question 1

In this answer we derive the ordinary least squares estimator of our linear model parameters. We begin by re-expressing the mean squared error (MSE) in its matrix form.

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

Note about the dimensions:
$Z \in \mathbb{R}^{n\times d}$ (consequently, $Z_i \in \mathbb{R}^{1\times d}$) and $\beta \in \mathbb{R}^{d\times 1}$

Now we take the derivative with respect to β,

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

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

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

Question 2

Note that in order to invert the matrix Z^TZ in Equation (1) we have to assume that Z is full rank, i.e., that the columns of Z are linearly independent. This is known as the full rank assumption of linear models. If our design matrix Z 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 Z is a linear combination of other columns of Z 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 $2Z^TZ$, this is a semi positive definite matrix if Z is full rank and hence we are able to 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 the figure 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.

Question 5

The MSE improved; using more data is one (easy, but sometimes expensive, depending on how easy is to obtain labelled data) way to improve the results of a model.