MA2823: Foundations of Machine Learning **Homework 5**

Due October 31, 2016.

---

1 Homework Problem

Question 1

Derive the solution of the ridge regression estimator $\hat{\beta}_{ridge} = (X^TX + \lambda I)^{-1}X^Ty$, if $(X^TX + \lambda I)$ invertible.

Solution:

$$\begin{aligned}
\hat{\beta}_{ridge} &= \arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \\
&= (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta \\
&= y^Ty - \beta^TX^Ty - y^TX\beta + \beta^TX^TX\beta + \lambda\beta^T\beta \\
&= y^Ty - \beta^TX^Ty - y^TX\beta + \beta^T(X^TX + \lambda I)\beta
\end{aligned}$$

In order to find the minimum of the previous cost we set the first derivative in respect to β to zero:

$$\begin{aligned}
\frac{\partial}{\partial\beta} = 0 &\Rightarrow -X^Ty - X^Ty + ((X^TX + \lambda I) + (X^TX + \lambda I))\beta = 0 \\
&\Rightarrow (X^TX + \lambda I)\beta = X^Ty
\end{aligned}$$

which gives the solution of the ridge regression estimator $\hat{\beta}_{ridge} = (X^TX+\lambda I)^{-1}X^Ty$, if $(X^TX + \lambda I)$ invertible.

Question 2

Compare the different regression techniques (linear, ridge, lasso, elastic net) in respect to amplitude of coefficients and sparcity, i.e. number of non-zero coefficients.

Solution: The linear regression has no regularization term, thus the calculated coefficients can have large amplitude and be many in number. In the case of ridge regression, due to the L_2-norm regularization, the coefficients are smaller but all the predictors are kept in the model. The lasso, due to the L_1-penalty, does both continuous shrinkage and automatic variable selection simultaneously. Elastic Net balances between variable selection like the lasso and shrinkage of coefficients like the ridge regression, therefore it results in more non-zero coefficients than lasso, but with smaller amplitudes.

Question 3

Why do we seek for a prediction model with as few variables as possible?

Solution: A simpler model can be easier interpreted because it puts more light on the relationship between the response and covariates. Parsimony is especially an important issue when the number of predictors is large.