Machine Learning I: Statistical Learning Theory
Homework 1, Due January 22

Problem 1 - Matrix Calculus: Let $X, \beta \in \mathbb{R}^p$, $\mathbf{X}$ be a $N \times p$ matrix, and let $Y \in \mathbb{R}^N$. Using the denominator layout notation conventions from Lecture 2,

a) Prove that $\frac{\partial(X^T\beta)}{\partial X} = \beta$.

b) Assume as above that Y is an N vector but assume that Y depends on X and X depends on some $Z \in \mathbb{R}^q$. Show that

$$\frac{\partial Y}{\partial Z} = \frac{\partial X}{\partial Z}\frac{\partial Y}{\partial X}.$$

Does the order matter?

c) Let a be a scalar that depends on X and let Y be a N-vector that depends on X. Prove that

$$\frac{\partial}{\partial X}(aY) = a\frac{\partial Y}{\partial X} + \frac{\partial a}{\partial X}Y^T$$

Problem 2 - Loss Functions:

a) In Lecture 1, we showed that the residual sum of square can be written

$$RSS(\beta) := (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta).$$

Prove that $0 = \frac{\partial}{\partial\beta}RSS(\beta)$ when

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\beta) = 0,$$

and use it to show that $RSS(\beta)$ has an critical point when $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.

b) Ridge regression changes the loss function to add in a term penalizing the β if they get too large: For any$\lambda \in \mathbb{R}^+$,

$$\text{Ridge}_\lambda(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda^2\beta^T\beta.$$

Find an expression for the location of the critical point/s of $\text{Ridge}_\lambda(\beta)$.

Problem 3 - Computing Linear Regression: Consider the points

X	1.2	3.2	5.1	3.5	2.6
Y	7.8	1.2	6.4	2.6	8.1

1. Fit a linear function to this dataset when the loss is RSS. You may use a computer to solve the matrix equation but you should report the best fit function.

2. Fit a linear function to this dataset when the loss is the Ridge Loss from (2.b) with $\lambda = 1$ and with $\lambda = 10$. What specifically explains the difference in values between the three fits?