<u>EE2211 Tutorial 7</u>

Question 1:

This question explores the use of Pearson's correlation as a feature selection metric. We are given the following training dataset.

	Datapoint 1	Datapoint 2	Datapoint 3	Datapoint 4	Datapoint 5
Feature 1	0.3510	2.1812	0.2415	-0.1096	0.1544
Feature 2	1.1796	2.1068	1.7753	1.2747	2.0851
Feature 3	-0.9852	1.3766	-1.3244	-0.6316	-0.8320
Target y	0.2758	1.4392	-0.4611	0.6154	1.0006

What are the top two features we should select if we use Pearson's correlation as a feature selection metric? Here's the definition of Pearson's correlation. Given N pairs of datapoints $\{(a_1, b_1), (a_2, b_2), \cdots, (a_N, b_N)\}$, the Pearson's correlation r is defined as $r = \frac{\frac{1}{N}\sum_{n=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\frac{1}{N}\sum_{n=1}^{N}(a_i - \bar{a})^2}\sqrt{\frac{1}{N}\sum_{n=1}^{N}(b_i - \bar{b})^2}}$, where $\bar{a} = \frac{1}{N}\sum_{n=1}^{N} a_n$ and $\bar{b} = \frac{1}{N}\sum_{n=1}^{N} b_n$ are the empirical means of a and b respectively. $\sigma_a = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(a_i - \bar{a})^2}$ and $\sigma_b = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(b_i - \bar{b}\right)^2}$ are referred to as the empirical standard deviation of a and b. $Cov(a, b) = \frac{1}{N}\sum_{n=1}^{N}(a_i - \bar{a})(b_i - \bar{b})$ is known as the empirical covariance between a and b

Question 2:

This question further explores linear regression and ridge regression. The following data pairs are used for training:

$$\{x = -10\} \rightarrow \{y = 4.18\}$$

$$\{x = -8\} \rightarrow \{y = 2.42\}$$

$$\{x = -3\} \rightarrow \{y = 0.22\}$$

$$\{x = -1\} \rightarrow \{y = 0.12\}$$

$$\{x = 2\} \rightarrow \{y = 0.25\}$$

$$\{x = 7\} \rightarrow \{y = 3.09\}$$

The data for testing are as follows:

$$\{x = -9\} \rightarrow \{y = 3\}$$

$$\{x = -7\} \rightarrow \{y = 1.81\}$$

$$\{x = -5\} \rightarrow \{y = 0.80\}$$

$$\{x = -4\} \rightarrow \{y = 0.25\}$$

$$\{x = -2\} \rightarrow \{y = -0.19\}$$

$$\{x = 1\} \rightarrow \{y = 0.4\}$$

$$\{x = 4\} \rightarrow \{y = 1.24\}$$

$$\{x = 5\} \rightarrow \{y = 1.68\}$$

$$\{x = 6\} \rightarrow \{y = 2.32\}$$

$$\{x = 9\} \rightarrow \{y = 5.05\}$$

(a) Use the polynomial model from orders 1 to 6 to train and test the data without regularization. Plot the Mean Squared Errors (MSE) over orders from 1 to 6 for both the training and the test sets. Which model order provides the best MSE in the training and test sets? Why? [Hint: the underlying data was generated using a quadratic function + noise]

(b) Use regularization (ridge regression) λ=1 for all orders and repeat the same analyses. Compare the plots of (a) and (b). What do you see? [Hint: the underlying data was generated using a quadratic function + noise]