EE2211 Tutorial 7

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

$$r = \frac{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(b_i - \bar{b})^2}},$$

where $\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_i$ and $\bar{b} = \frac{1}{N}\sum_{i=1}^{N} b_i$ are the empirical means of a and b respectively. $\sigma_a = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(a_i - \bar{a})^2}$ and $\sigma_b = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(b_i - \bar{b}\right)^2}$ 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]

Question 3:

Suppose we randomly sample a training set D from some unknown distribution. For each training set D we sample, we train a regression model to predict y from x. We repeat this process 10 times resulting in 10 trained models. A new test sample (x, y) = (5, 10) is sampled from the same distribution that generated the training sets. Recall that y = f(x) + ε, where epsilon has mean 0 and variance $\sigma^2 > 0$. In this instance, the realization of the noise ε is 0.5 and hence, f(x = 5) = 9.5. Suppose the predictions of the new test sample based on the 10 trained models are 9, 11, 23, 6, 8, 12, 10, 4, 13, 7. Based on this 10 trials, what is the Bias^2 and Variance of our regression model?

Question 4:

Suppose we randomly sample a training set D from some unknown distribution. For each training set D we sample, we train a regression model to predict y from x. We repeat this process 10 times resulting in 10 trained models. A new test sample (x, y) = (3, 7) is sampled from the same distribution that generated the training sets. Recall that y = f(x) + ε, where epsilon has mean 0 and variance σ^2. Suppose the predictions of the new test sample based on the 10 trained models are 6, 8, 9, 5, 10, 5, 4, 8, 9, 3. Suppose the algorithm has a bias of 0, what is an estimate of σ^2 using the bias-variance decomposition theorem?

Question 5:

In this question, we will try to numerically verify the bias-variance decomposition theorem. Suppose $y = 0.01 + 0.1x + 0.05x^2 + \varepsilon$, where ε is Gaussian distributed with mean of 0 and standard deviation of 0.2. Write the following code:

i. Create a test set by randomly sampling 200 pairs of (x, y). To achieve this, you can
   i. Sample x uniformly between the interval from -10 to 10 (using numpy.random.uniform)
   ii. Sample y using the equation $y = f(x) + \varepsilon = 0.01 + 0.1x + 0.05x^2 + \varepsilon$ (using numpy.random.normal)

ii. Create a training set by randomly sampling 20 pairs of (x, y) using the same procedure in (i)

iii. Use the training set from (ii) to train polynomial regression (no regularization) for polynomial orders 1, 2, 3, 4 and 5. Used the trained models to perform prediction in the test set

iv. Repeat steps (ii) and (iii) 1000 times. Note that across these 1000 trials, the test set remains the same. Only the training set changes from trial to trial

Compute Bias^2, Variance, Mean Squared Error (MSE) for polynomial orders 1 to 5 based on the 1000 trials and averaged across the 200 test samples. More specifically, the test set comprises 200 samples (x_i, y_i), where $i = 1, \cdots, 200$. For a particular polynomial and for the j-th trial ($j = 1, \cdots, 1000$), let the prediction of the i-th test sample be $\hat{y}_{ij}$. The average prediction of the i-th test sample across the 1000 trials is defined as $E(\hat{y}_i) = \frac{1}{1000}\sum_{j=1}^{1000} \hat{y}_{ij}$

- Bias^2 can be computed as $\frac{1}{200}\sum_{i=1}^{200}(E(\hat{y}_i) - f(x_i))^2$
- Variance can be computed as $\frac{1}{200}\sum_{i=1}^{200}\frac{1}{1000}\sum_{j=1}^{1000}\left(\hat{y}_{ij} - E(\hat{y}_i)\right)^2$
- MSE can be computed as $\frac{1}{200}\sum_{i=1}^{200}\frac{1}{1000}\sum_{j=1}^{1000}(\hat{y}_{ij} - y_i)^2$

What do you observe?