Homework 1 Solution

Due date: 9.19.2018

• Problem 1

1. Problem 1.2 from textbook Predictive Learning

**Solution:** We generate 12 data samples \((x,y)\). The \(x\)-coordinate is uniformly distributed in \([0,1]\). The \(y\)-coordinate has Gaussian distribution with zero mean and variance 0.5. We have to specify the standard deviation and use "`randn`" function to generate the noise. The polynomial fitting models are shown below along with the data samples. The mean-squared error(MSE) for the linear, quadratic, and degree 6 models are 0.25678, 0.25369, 0.065321, respectively.

![Graph showing data samples and polynomial models](image)

A polynomial of degree 6 gives the best(smallest) fitting error. However, it would not give the best prediction accuracy for future data.

• Problem 2

1. Problem 1.4 from textbook Predictive Learning

**Solution:** The probability of receiving no response to a single spam message is 1-0.004=0.996. So the probability if receiving a favorable response(at least one sale) to 250 spam messages is 

\[ 1-0.996^{250} = 0.62. \]

• Problem 3
1. Daily prices of Vanguard Total Stock Market ETF (Symbol VTI Analysis)

**Solution:**

- (1) The Histogram for the observation $X(t)$ is shown below:

```
Figure 1: Histogram of $X(t)$, estimated mean and standard deviation and empirical distribution
```

And the mean and standard deviation are estimated using normfit function shown in Figure 1.

- (2) The Histogram for the 4-days moving average $MA(t)$ is shown below:

```
Figure 2: Histogram of $MA(t)$, estimated mean and standard deviation and empirical distribution
```

And the mean and standard deviation are estimated using normfit function shown in Figure 2.

- (3) The mean of the observations $X(t)$ and moving average $MA(t)$ are approximately the same based on historical data and the standard deviation of $MA(t)$ is approximately half of that of $X(t)$.

**Analyze the Derivation:**

MA(t) = \frac{(X(t) + X(t-1) + X(t-2) + X(t-3))}{4}. Assuming that $X(t)$ are statistically independent,
the mean $\overline{MA(t)}$ is the same as $\overline{X(t)}$. The standard deviation of $MA(t)$ can be calculated as $\text{VAR}(MA) = 4 \times \text{VAR}(\frac{1}{4}X)$ where $\text{VAR}(\frac{1}{4}X) = (\frac{1}{16}) \times \text{VAR}(X)$, so that $\text{VAR}(MA) = 4 \times (\frac{1}{16}) \times \text{VAR}(X)$. So, $\sigma_{MA}^2 = \frac{1}{4} \sigma_x^2$, so that the $\sigma_{MA} = \frac{1}{2} \sigma_x$. Hence, the standard deviation of $MA(t)$ estimated from data should be half of $\sigma_x$. And this relationship approximately holds for Year 2017 data, that is $\sigma_x = 0.44$ and $\sigma_{MA} = 0.19 \approx 0.5 \times \sigma_x$. 