Assignment 1

Machine Learning and Big Data for Economics and Finance

Exercise 1. Nonlinear regression learning exercise

Consider the model

$$Y = e^{\beta} X + \varepsilon,$$

where $\beta = 0.6$, X, Y and ε are three random variables such that $X \sim N(0,1)$, $\varepsilon \sim N(0,1)$ and X and ε are independent of each other.

1. Write `R` code to generate a sample of size $n = 1000$ from this model. Print summary statitics of the variables generated.
2. Assuming now that one only observes X and Y in that sample, let us conduct the supervised learning exercise where the objective is to predict Y given X.

   Try to learn the function f in $Y \approx f(X)$ by the three different models

   - Linear regression.
   - Quadratic regression.
   - Cubic regression.

   For each of the three models,

   a. Write `R` code.

   b. Show `R` output.

   c. Plot the residuals and discuss the residual plots.

   d. Plot on a single figure the data and the in-sample predicted values for all three models.
3. Assuming again that one only observes X and Y in that sample, let us further assume that we know that f takes the functional form $f(x) = e^{\beta} x$.

   We will try to obtain an estimate for β by minimizing the sum of the squares of the residulas

   $$Q(b) = \sum_{i=1}^{n} (y_i - e^{b} x_i)^2$$

   where x_i and y_i are obviously the points in the sample.

   a. Plot Q for b taking values on a grid of points of size 100 in the interval $[-1.5, 1.5]$.

   b. Deduce an estimate $\hat{\beta}$.

   c. Plot a figure showing the in-sample predicted values vs. the actual data.

   d. Would it have been possible to estimate this model using the `lm()` function in `R`?
4. In this last question, we will compare the fit for all 4 models (linear, quadratic, cubic and actual model).

   a. Compute the training mean squared error for all four models.

   b. Generate a new sample of size $m = 100$ from the true model and compute the test MSE for all four models using that new sample.

   c. Based on the results, which model do you choose? Discuss.

Exercise 2. Consider the following sample of the three random variables X_1, X_2 and Y:

Obs.	X_1	X_2	Y
1	1	2	0
2	1	3	0
3	-3	1	0
4	2	2	1
5	3	2	1
6	4	1	1
7	4	3	1

Table 1.

1. Enter the data into `R`.
2. Given an input of the form (a, b), write an `R` function `predicty(a,b)` that outputs a prediction of $\Pr\{Y = 1 | X_1 = a, X_2 = b\}$ based on 1-nearest neighbor classification and based on the training sample in the table. Test your function on 3 random points.

Notes: These are some helpful additional hints.

- An `R` function `sum2` that takes two inputs `u` and `v` and return their sum could be written as follows

```
sum2 = function(u,v)
{
  s = u + v
  return(s)
}
```

  and then could be used in the console by simply typing `sum2(3,4)` in order to add 3 and 4.

- The `R` function `seq` is useful for generating a sequence (list) of numbers.