Assignment 2

Machine Learning and Big Data for Economics and Finance

For each exercise, provide the `R` code, the `R` output and comment both the codes and the results.

Exercise 1. (50 points)

The probit model is a popular alternative to logistic regression. It is also learned by maximizing the likelihood of the model, but the main difference is in the way the conditional probability of the output variable given the inputs is computed. For example, when the model has one input, then that probability is given by

$$\Pr(Y=1|X=x)=\Phi(\beta_0+\beta_1 x),$$

where Φ is the cumulative distribution function of a standard normal random variable $N(0,1)$

$$\Phi(x)=\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}t^2}\mathrm{d}t.$$

Write an `R` function that, given a dataset of one input and one output variables, computes the maximum likelihood estimate of (β_0,β_1) along with their bootstrapped standard errors. Also let that `R` function return a prediction of outputs Y if an optional input argument of new values of $(x_1,...,x_k)$ for X is given. Finally apply the `R` function to the dataset in the file `LR2.csv`.

For this exercise, neither the function `glm` nor any `R` package can be used.

Exercise 2. (50 points)

Consider the following model. Let X and U be two independent uniformly distributed random variables and let Y be given by the following equation

$$Y=I\bigg(U\leqslant\frac{1}{1+\exp(-\beta_0-\beta_1X-\beta_2X^3-\beta_3\log(X))}\bigg),$$

where I is the indicator function and $\beta_0,\beta_1,\beta_2,\beta_3$ are fixed parameters.

1. Assuming one observes a sample of size n of the variables X and Y, comment the underlying model for that dataset while explaining its structure in plain English.
2. Write an `R` function that generates a sample of size n of the variables X and Y.
3. Given the parameters $\beta_0=-4$, $\beta_1=2$, $\beta_2=5$ and $\beta_3=4$, generate 1000 samples of size $n=1000$ from the model. Construct a single figure where box and whisker plots of the test error rates of the following 10 models are plotted: Linear probability model, logistic regression, linear discriminant analysis, quadratic discriminant analysis and nearest neighbor classification with the number of nearest neighbors $K=1,...,6$. Comment the results extensively.

For this exercise, except for `class` and `MASS`, no `R` packages can be used.