ML01 – Spring 2019
Lab 2: Exploratory data analysis, k nearest neighbor regression and classification, normal data generation

1 Exploratory data analysis and K nearest neighbor regression

The data in the file `prostate.data` come from a study by Stamey et al. (1989) that examined the correlation between the level of prostate specific antigen (PSA) and a number of clinical measures, in 97 men who were about to receive a radical prostatectomy. The goal is to predict the log of PSA (`lpsa`) from a number of measurements including log cancer volume (`lcavol`), log prostate weight `lweight`, `age`, log of benign prostatic hyperplasia amount `lbph`, seminal vesicle invasion `svi`, log of capsular penetration `lcp`, Gleason score `gleason`, and percent of Gleason scores 4 or 5 `pgg45`.

1. Read the data file. What are the different data types ? (Use the function `summary`).
2. Display the data using scatter plots (function `plot`) and boxplots (function `boxplot`). Which variables seem to explain the response variable `lpsa` ?
3. Predict `lpsa` from input variables `lcavol`, `lweight`, `age` and `lbph` using k nearest neighbor regression (function `knn.reg` in package `FNN`). (Use the partition between training and test data encoded in variable `train`. Normalize the input data using function `scale`).
4. Represent graphically the test mean-squared error as a function of K. Which value of K seems to be optimal ?

2 Normal data generation and k nearest neighbor regression

We consider a classification problem with $k = 3$ classes and $p = 2$ input variables. Let $Y \in \{1, 2, 3\}$ denote the class variables and $\mathbf{X} = (X_1, X_2)^T$

the feature vector. The marginal distribution of Y is defined by the following "prior probabilities" :

$$\mathbb{P}(Y = 1) = 0.3, \quad \mathbb{P}(Y = 2) = 0.2, \quad \mathbb{P}(Y = 3) = 0.5,$$

and the conditional densities of $\mathbf{X}$ given $Y = k$, $k = 1, 2, 3$ are multivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with

$$\boldsymbol{\mu}_1 = (0, 0)^T, \quad \boldsymbol{\mu}_2 = (0, 2)^T, \boldsymbol{\mu}_3 = (2, 0)^T,$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 2 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & -0.5 \\ -0.5 & 1 \end{pmatrix} \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

To generate an instance, we can first generate a realization of Y (using function `sample`), and then a realization or $\mathbf{X}$ from the condition distribution (using function `rmvnorm` of package `mvtnorm`).

1. Write a function `gen.data` that generates a data set of size N. Generate a training set of size $N = 100$ and a test set of size $N_t = 1000$. Plot the training data.
2. Classify the test set using the training set and the k nearest neighbor rule with $k = 5$ (function `knn` of package `FNN`). What is the test error rate (the proportion of misclassified data in the test set) ?
3. Repeat the previous question with different values of k and different values of N.
4. Plot the average test error rate as a function of k for $M = 10$ training sets of size $N = 100$, and $M = 10$ training sets of size $N = 500$. What do you observe ?