

## Coding Assignment 1

---

**Due Monday, Sep 14, 11:59 p.m.**

This assignment is related to the simulation study described in Section 2.3.1 (the so-called Scenario 2) of “Elements of Statistical Learning” (ESL).

**Scenario 2:** the two-dimensional data  $X \in \mathbf{R}^2$  in each class is generated from a mixture of 10 different bivariate Gaussian distributions with uncorrelated components and different means, i.e.,

$$X|Y = k, Z = l \sim \mathcal{N}(\mathbf{m}_{kl}, s^2 \mathbf{I}_2),$$

where  $k = 0, 1$ ,  $l = 1 : 10$ ,  $P(Y = k) = 1/2$ , and  $P(Z = 1) = 1/10$ . In other words, given  $Y = k$ ,  $X$  follows a mixture distribution with density function

$$\frac{1}{10} \sum_{l=1}^{10} \left( \frac{1}{\sqrt{2\pi s^2}} \right)^2 e^{-\|\mathbf{x} - \mathbf{m}_{kl}\|^2 / (2s^2)}.$$

You can choose your own values for  $s$  and the twenty 2-dim vectors  $\mathbf{m}_{kl}$ , or you can generate them from some distribution.

Repeat the following simulation 20 times. In each simulation,

1. follow the data generating process to generate a training sample of size 200 and a test sample of size 10,000, and
2. calculate the **training** and **test** errors (the averaged 0/1 error<sup>1</sup> )

for each the following **four** procedures:

- Linear regression with cut-off value<sup>2</sup> 0.5,
- quadratic regression with cut-off value 0.5,
- $k$ NN classification with  $k$  chosen by 10-fold cross-validation, and
- the Bayes rule (assume you know the values of  $\mathbf{m}_{kl}$ 's and  $s$ ).

Summarize your results on training errors and test errors graphically, e.g., using boxplot or stripchart. Also report the mean and standard error for the chosen  $k$  values.

*Continue on the next page —*

---

<sup>1</sup>For each sample, the incurred error is 1 if there is a mistake, and 0 otherwise.

<sup>2</sup>predict  $Y$  to be 1 if the returned estimate is bigger than the cut-off value, and 0 otherwise.

### What you need to submit?

An R Markdown file in HTML format.

- You are only allowed to use **two** packages: `class` and `ggplot2`. In other words, you have to write your own function to select the optimal K value based on 10-fold CV.
- Set the seed at the beginning of your code to be the last 4-dig of your University ID. So once we run your code, we can get the same result.
- Name your file starting with

`Assignment_1_xxxx_netID`

where “xxxx” is the last 4-dig of your University ID and make sure the same 4-dig is used as the seed in your code.

For example, the submission for Max Chen with UID 672757127 and netID mychen12 should be named as

`Assignment_1_7127_mychen12_MaxChen.html`

You can add whatever characters after your netID.