Coding Assignment 1

Due Monday, Feb 15, 11:59 p.m. (Chicago Time)

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.,
>
> $$X|Y = k, Z = l \sim \mathcal{N}\big(\mathbf{m}_{kl}, s^2\mathbf{I}_2\big),$$
>
> 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
>
> $$\frac{1}{10}\sum_{l=1}^{10}\left(\frac{1}{\sqrt{2\pi s^2}}\right)^2 e^{-\|\mathbf{x}-\mathbf{m}_{kl}\|^2/(2s^2)}.$$

First, generate the twenty two-dimensional vectors as follows: for $l = 1, \ldots, 10$,

$$\begin{aligned}\mathbf{m}_{0l} \text{ i.i.d.} &\sim \mathcal{N}\big((0,1)^T, \sigma^2\mathbf{I}_2\big)\\ \mathbf{m}_{1l} \text{ i.i.d.} &\sim \mathcal{N}\big((1,0)^T, \sigma^2\mathbf{I}_2\big).\end{aligned}$$

Then, repeat the following simulation 20 times using the same set of centers generated above. 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[1])

for each the following **four** procedures:

- Linear regression with cut-off value[2] 0.5,
- quadratic regression with cut-off value 0.5,
- kNN classification with k chosen by 10-fold cross-validation, and
- the Bayes rule (assume your 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 —-

[1] For each sample, the incurred error is 1 if there is a mistake, and 0 otherwise.

[2] 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.

- Specify the values for s and σ. Suggest to choose a larger value for σ and a smaller value for s, e.g., pick a value for σ and then set $s^2 = \sigma^2/5$.

- 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.