STAT 542 Spring 2019

Coding Assignment 1

Due Thursday, Jan 31, 3:00 a.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.,
>
> $$X|Y = k, Z = l \ \sim \ \mathcal{N}\big(\mathbf{m}_{kl}, s^2\mathbf{I}_2\big),$$
>
> 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
>
> $$\frac{1}{10}\sum_{l=1}^{10}\left(\frac{1}{\sqrt{2\pi s^2}}\right)^2 e^{-\|\mathbf{x}-\mathbf{m}_{kl}\|^2/(2s^2)}.$$

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, following the data generating process,

1. 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 the following four procedures:

- Linear regression with cut-off value 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 selected k values.

R packages you are allowed to use are `class` (for kNN) and `ggplot2` (for graphs).

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

What you need to submit?
A PDF file and an R Markdown file that produces the PDF file.

- Name your files starting with

  `Assignment_1_xxxx_netID`

  where "`xxxx`" is the last 4-dig of your University ID.

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

  `Assignment_1_7127_mychen12_MaxChen.Rmd/.pdf`

  You can add whatever characters after your netID.

- 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.