STAT 542/CS 598 Fall 2018

Coding Assignment 1

Due Thursday, Sept. 13, 11:30 p.m. (PT)

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

I have also discussed this example in class; please check notes from this week on "kNN vs. LinearRegression" and the related Rcode.

Following the data generating process, generate a training sample of size 200 and a test sample of size 10,000.

Evaluate the performance (the averaged 0/1 error[1]) for the following three procedures:

- Linear regression with cut-off value 0.5,
- kNN classification with $k = 1, 3, 5, 7, 11, 21, 31, 45, 69, 101, 151$, and
- the Bayes rule (assume your know the values of $\mathbf{m}_{kl}$'s and s).

Summarize your result graphically. Design your graph so that it shows the **test** and **training** errors for linear regression and kNN, and **test** error for the Bayes classifier. Check Figure 2.4 of ESL and figures from the notes.

Write R/Python code to simulate the data, compute the errors, and produce a PDF file of your graph.

(Continue on the next page ⟶)

[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 the R/Python code that produces the PDF file. If we source your R file or run your Python script, we should see a PDF file generated in the same directory.

- Name your R/Python file 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.R`

  You can add whatever characters after your netID.

- Name the PDF file similarly, starting with

  `AssignmentOutput_1_xxxx_netID`

  where "xxxx" is the last 4-dig of your University ID. Name your submitted PDF file the same as the file that will be generated by your code.

- 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 PDF file.