Math 680, Fall 2020
Homework 1 Due: Tuesday, 9/22/2020

General Instructions:

- Turn in all your HW through Canvas.
- All the HW files (except the R code) should be saved as a single PDF, and named in the form "Last-name_hw1.pdf".
- The code should be saved as "Last-name_hw1_code.r".
- Test your R code before submission to make sure it can be executed successfully by the "source()" function.

1. (**Optimal Decision Rule**)

   Suppose X is a continuous random variable with the pdf $f(x)$ and cdf $F(x)$. Show that

   (i) $\min_a E(X-a)^2 = E(X-EX)^2$.

   (ii) $\min_a E|X-a| = E|X-m|$, where m is the median of X.

   Here, "E" denote the expectation operator.

2. (**Bayes Estimator Under Squared Error Loss**)

   Let $X_1, \ldots, X_n$ be iid from Bernoulli(p), where $p \in (0,1)$ is the unknown parameter and n is the sample size. Assume the prior distribution on p is Beta(α, β), where the hyper-parameters $\alpha > 0$ and $\beta > 0$ are known. Consider the squared error loss for evaluating the estimator of p.

(i) Specify the posterior distribution of p.

(ii) Under the squared error loss, it is known that the Bayesian estimator of p, $\hat{p}_{\text{bayes}}$, is given by the posterior mean of p. Show that

$$\hat{p}_{\text{bayes}} = \frac{\sum_{i=1}^{n} X_i + \alpha}{\alpha + \beta + n}.$$

(iii) Show that the risk of $\hat{p}_{\text{bayes}}$ is given by

$$R(p, \hat{p}_{\text{bayes}}) = \frac{np(1-p)}{(\alpha+\beta+n)^2} + \left(\frac{np+\alpha}{\alpha+\beta+n} - p\right)^2.$$

This is also known as the MSE of $\hat{p}_{\text{bayes}}$.

(iv) Consider the special case $\alpha = \beta = \sqrt{n/4}$. Show that

$$\begin{aligned} \hat{p}_{\text{bayes}} &= \frac{\sum_{i=1}^{n} X_i + \sqrt{n/4}}{n + \sqrt{n}}, \\ R(p, \hat{p}_{\text{bayes}}) &= \frac{n}{4(n+\sqrt{n})^2}. \end{aligned}$$

3. (**Bayes Rule for Unequal Costs**)

Consider Example 1 in Lecture note 4 (see Page 24). If we change unequal costs as the following

$$C(0,1) = 3, \quad C(1,0) = 2.$$

(i) Derive the Bayes rule for this classification problem.

(ii) Write down the equation for the Bayes decision boundary.

(iii) Provide a numerical solution for the Bayes decision boundary.

4. (**Two-Class Classification Problem: Scenario 1**)

   This is a two-class problem, and we draw 100 points from each class in the following way.

   Generate 100 observations from a bivariate Gaussian distribution $N(\mu_1, \Sigma_1)$ with $\mu_1 = (2,1)^T$ and $\Sigma_1 = \mathbf{I}$ (the identity matrix), and label them as *Green*. Generate 100 observations from a bivariate Gaussian distribution $N(\mu_2, \Sigma_2)$ with $\mu_2 = (1,2)^T$ and $\Sigma_2 = \mathbf{I}$, and label them as *Red*.

   (i) Write R code to generate the training data. Set the seed with *set.seed(2020)* before calling the random number generation function.

   (ii) Draw the scatter plot of the training data, using different labels/colors for two classes.

   (iii) Generate a test set, with 500 observations from each class, using *set.seed(2019)*. Save the data set for Question 4.

   Submit your R code along with the scatter plot.

5. (**Bayes Classification Rule: Scenario 1**)

   Assume two classes in Scenario 1 described in Question 4 have the same prior probabilities.

   (i) Using the 0-1 loss, derive the Bayes classifier. Simplify the solution as much as you can.

   (ii) Add the Bayes decision boundary to the scatter plot drawn in Question 4.

(iii) Compute the training and test errors for the Bayes classifier, using the data generated in Question 4.