# Homework 2: Classification Methods

### Overview

Due Sunday by 11:59 pm.

### Fork the `problem-set-2` repository

## The Bayes Classifier

1. (20 points) For classification problems, the test error rate is minimized by a simple classifier that assigns each observation to the most likely class given its predictor values:

$$\Pr(Y = j|X = x_0)$$

where $x_0$ is the test observation and each possible class is represented by $J$. This is a **conditional probability** that $Y = j$, given the observed predictor vector $x_0$. This classifier is known as the **Bayes classifier**. If the response variable is binary (i.e. two classes), the Bayes classifier corresponds to predicting class one if $\Pr(Y = 1|X = x_0) > 0.5$, and class two otherwise.

Produce a graph illustrating this concept. Specifically, implement the following elements in your program:

   a. Set your random number generator seed.
   b. Simulate a dataset of $N = 200$ with $X_1, X_2$ where $X_1, X_2$ are random uniform variables between $[-1, 1]$.
   c. Calculate $Y = X_1 + X_1^2 + X_2 + X_2^2 + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = 0.25)$.
   d. $Y$ is defined in terms of the log-odds of success on the domain $[-\infty, +\infty]$. Calculate the probability of success bounded between $[0, 1]$.
   e. Plot each of the data points on a graph and use color to indicate if the observation was a success or a failure.
   f. Overlay the plot with Bayes decision boundary, calculated using $X_1, X_2$.
   g. Give your plot a meaningful title and axis labels.
   h. The colored background grid is optional.

## Exploring Simulated Differences between LDA and QDA

*Note: Unless otherwise specified, assume the number of observations $N = 1000$.*

2. (20 points) If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?
   a. Repeat the following process 1000 times.
      i. Simulate a dataset of 1000 observations with $X_1, X_2 \sim \text{Uniform}(-1, +1)$. $Y$ is a binary response variable defined by a Bayes decision boundary of $f(X) = X_1 + X_2$, where values 0 or greater are coded `TRUE` and values less than 0 or coded `FALSE`. Whereas your simulated $Y$ is a function of $X_1 + X_2 + \epsilon$ where $\epsilon \sim N(0, 1)$. That is, your simulated $Y$ is a function of the Bayes decision boundary plus some irreducible error.
      ii. Randomly split your dataset into 70/30% training/test sets.
      iii. Use the training dataset to estimate LDA and QDA models.
      iv. Calculate each model's training and test error rate.
   b. Summarize all the simulations' error rates and report the results in tabular and graphical form. Use this evidence to support your answer.

3. (20 points) If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?
    a. Repeat the following process 1000 times.
        i. Simulate a dataset of 1000 observations with $X_1, X_2 \sim \text{Uniform}(-1, +1)$. $Y$ is a binary response variable defined by a Bayes decision boundary of $f(X) = X_1 + X_1^2 + X_2 + X_2^2$, where values 0 or greater are coded `TRUE` and values less than 0 or coded `FALSE`. Whereas your simulated $Y$ is a function of $X_1 + X_1^2 + X_2 + X_2^2 + \epsilon$ where $\epsilon \sim N(0, 1)$. That is, your simulated $Y$ is a function of the Bayes decision boundary plus some irreducible error.
        ii. Randomly split your dataset into 70/30% training/test sets.
        iii. Use the training dataset to estimate LDA and QDA models.
        iv. Calculate each model's training and test error rate.
    b. Summarize all the simulations' error rates and report the results in tabular and graphical form. Use this evidence to support your answer.
4. (20 points) In general, as sample size $n$ increases, do we expect the test error rate of QDA relative to LDA to improve, decline, or be unchanged? Why?
    a. Use the non-linear Bayes decision boundary approach from part (2) and vary $n$ across your simulations (e.g., simulate 1000 times for `n = c(1e02, 1e03, 1e04, 1e05)`).
    b. Plot the test error rate for the LDA and QDA models as it changes over all of these values of $n$. *Use this graph to support your answer.*

# Modeling voter turnout

An important question in American politics is why do some people participate in the political process, while others do not? Participation has a direct impact on outcomes – if you fail to participate in politics, the government and political officials are less likely to respond to your concerns. Typical explanations focus on a resource model of participation – individuals with greater resources, such as time, money, and civic skills, are more likely to participate in politics. One area of importance is understanding voter turnout, or why people participate in elections. Using the resource model of participation as a guide, we can develop several expectations. First, women, who more frequently are the primary caregiver for children and earn a lower income, are less likely to participate in elections than men. Second, older Americans, who typically have more time and higher incomes available to participate in politics, should be more likely to participate in elections than younger Americans. Finally, individuals with more years of education, who are generally more interested in politics and understand the value and benefits of participating in politics, are more likely to participate in elections than individuals with fewer years of education.

While these explanations have been repeatedly tested by political scientists, an emerging theory assesses an individual's mental health and its effect on political participation.[1] Depression increases individuals' feelings of hopelessness and political efficacy, so depressed individuals will have less desire to participate in politics. More importantly to our resource model of participation, individuals with depression suffer physical ailments such as a lack of energy, headaches, and muscle soreness which drain an individual's energy and requires time and money to receive treatment. For these reasons, we should expect that individuals with depression are less likely to participate in election than those without symptoms of depression.

The 1998 General Social Survey included several questions about the respondent's mental health. `mental_health.csv` reports several important variables from this survey.

- `vote96` - 1 if the respondent voted in the 1996 presidential election, 0 otherwise
- `mhealth_sum` - index variable which assesses the respondent's mental health, ranging from 0 (an individual with no depressed mood) to 9 (an individual with the most severe depressed mood)[2]
- `age` - age of the respondent
- `educ` - Number of years of formal education completed by the respondent

[1] Ojeda, C. (2015). Depression and political participation. *Social Science Quarterly*, 96(5), 1226-1243.

[2] The variable is an index which combines responses to four different questions: "In the past 30 days, how often did you feel: 1) so sad nothing could cheer you up, 2) hopeless, 3) that everything was an effort, and 4) worthless?" Valid responses are none of the time, a little of the time, some of the time, most of the time, and all of the time.

- `black` - 1 if the respondent is black, 0 otherwise
- `female` - 1 if the respondent is female, 0 if male
- `married` - 1 if the respondent is currently married, 0 otherwise
- `inc10` - Family income, in $10,000s

5. (20 points) Building several classifiers and comparing output.
   a. Split the data into a training and test set (70/30).
   b. Using the training set and all important predictors, estimate the following models with `vote96` as the response variable:
      i. Logistic regression model
      ii. Linear discriminant model
      iii. Quadratic discriminant model
      iv. Naive Bayes (you can use the default hyperparameter settings)
      v. $K$-nearest neighbors with $K = 1, 2, \ldots, 10$ (that is, 10 separate models varying $K$) and *Euclidean* distance metrics
   c. Using the test set, calculate the following model performance metrics:
      i. Error rate
      ii. ROC curve(s) / Area under the curve (AUC)
   d. Which model performs the best? Be sure to define what you mean by "best" and identify supporting evidence to support your conclusion(s).