Lab0

Ryan Morreale, Emery Berry, Jacob Krol

Lab0: About Your Team, About Individuals, and Team Collaboration

Due: 11:59PM Sunday, January 25 on Canvas as a knitted pdf file of a Quarto document

1. You will determine your team's name and goals.
2. You will add to the document your individual sections.
3. You will practice collaborating on an applied statistical learning "project."

Instructions for Lab0

1. Using Quarto, you will complete the first team section "About Team Teamname." See further instructions below.
2. Each team member will add their own individual section to the team Quarto document. How do you want to collaborate on your document this semester? Github? Google Colab? Something else? See instructions for individual sections below.
3. As a team, you will complete an applied "project". There will be some individual components (so that everyone gets practice with implementing the stat learning methods) and team components.

About Your Team

In this team section, include a team photo (all of you), your team name, and your team's main goal for this semester for this course (i.e., what does your team want to accomplish by the end of the semester?). Throughout the semester you will be giving feedback to your teammates about how well they are helping your team accomplish its goal(s).

Individual Sections

Each individual must complete their own subsection, which must include:

- a photo of yourself with a caption explaining the context
- at least one (non-statistics) question you would like to know the answer to that could potentially be answered by applying statistical learning methods to data
- what you would love to be doing six months after graduation and then five years after that (what would make you excited to be doing?)
- what you hope your greatest career accomplishment will be
- and given these hopes and goals, what you are hoping to learn/accomplish/do in this course.
- You must also include something of your own choosing not described above. Anything. Be creative!

Team Applied Section

Find the misclassification rate for K-nearest neighbors (KNN) for a range of k values, given a complicated true generating model. Do this individually. As a team, compare answers and then plot the decision boundary for the team's "best" value of k.

Also, individually practice walking through the steps of Q1, Q2, and Q3 for this "project". Within this section will be a team section (what is the "best" k and what is the decision boundary) and individual sections. In your individual section, describe a plausible story for Q1, Q2, and Q3. Specifically, for Q1: What is Y, X1, and X2, and why should we care?. Make up something plausible that you actually care about. Y could be whether a kitten is adopted from a shelter given X1=age and X2=health of the cat. That's just one of countless plausible examples.

For Q2: Make predictions given X1 and X2 using KNN and k=(1 and 5) or (2 and 6) or (3 and 7) or (4 and 8). That is, each teammate fits KNN for two different values of k. Compute the misclassification rates for your values of k.

For Q3: Using this model and your plausible Q1 scenario, what is the answer for X1=0.5 and X2=0.5? For your scenario, what are some ethical implications of your stat learning modeling?

Generating Model

We have a logistic regression generating model. Given $x_1 \in [0,1]$ and $x_2 \in [0,1], Y \sim Ber(p)$, where p is related to x_1 and x_2 through the logistic link function: $\log(\frac{p}{(1-p)}) = x_1 - 2x_2 - 2{x_1}^2 + {x_2}^2 + 3x_1x_2$, where log is the natural log, sometimes written as ln.

The code for this is below.

```
library(class)
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.6
v forcats   1.0.1     v stringr   1.6.0
v ggplot2   4.0.1     v tibble    3.3.1
v lubridate 1.9.4     v tidyr     1.3.2
v purrr     1.2.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
#Generative model
set.seed(116) #setting a random seed so that we can reproduce everything exactly if we want t

generate_y <- function(x1,x2) { #two input parameters to generate the output y
  logit <- x1 -2*x2 -2*x1^2 + x2^2 + 3*x1*x2
  p <- exp(logit)/(1+exp(logit)) #apply the inverse logit function
  y <- rbinom(1,1,p) #y becomes a 0 (with prob 1-p) or a 1 with probability p
}
```

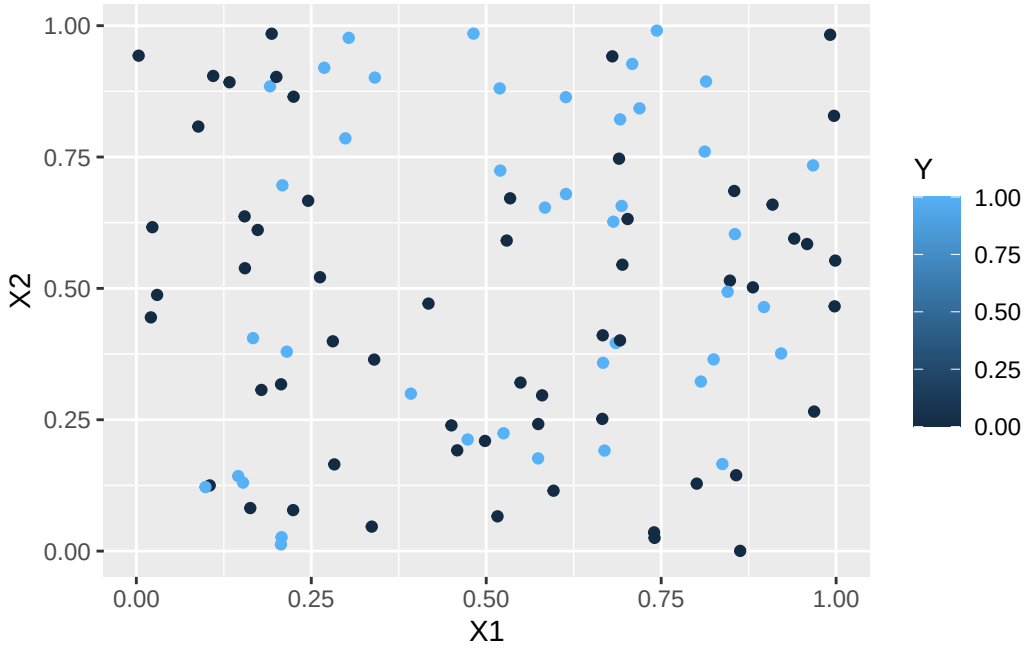
Example code

We are going to use our generating model to create a test set of 100 predictors (x1, x2), and then 100 outcomes. Then we plot all three variables just to see what the generating model is doing.

```
# Generate a dataset with 100 points
set.seed(116)
n = 100
X1 <- runif(n,0,1)
X2 <- runif(n,0,1)

#I'm going to use a for loop to generate 100 y's
Y <- rep(0,n) #initializing my Y to be a vector of 0's
for (i in 1:n) {
  Y[i] <- generate_y(X1[i],X2[i])
}

sum(Y) #How many 0's and 1's were predicted? In this training set, 42% were 1's. However, alr
```

```
[1] 42
```

```
training <- cbind(X1,X2,Y) #combining all of my variables into a training dataset
ggplot(data=training, aes(x=X1, y=X2, color=Y)) +
  geom_point()
```



This generating function seems to produce Y's with some spatial pattern in the X1, X2 parameter space, but the regions of 0's and 1's are not very well separated. How well will KNN do to classify/predict Y given new x1 and x2 values?

Test dataset

We are going to generate a new set of 100 predictors (x1, x2) and outcomes (y) that we will use as our "ground truth".

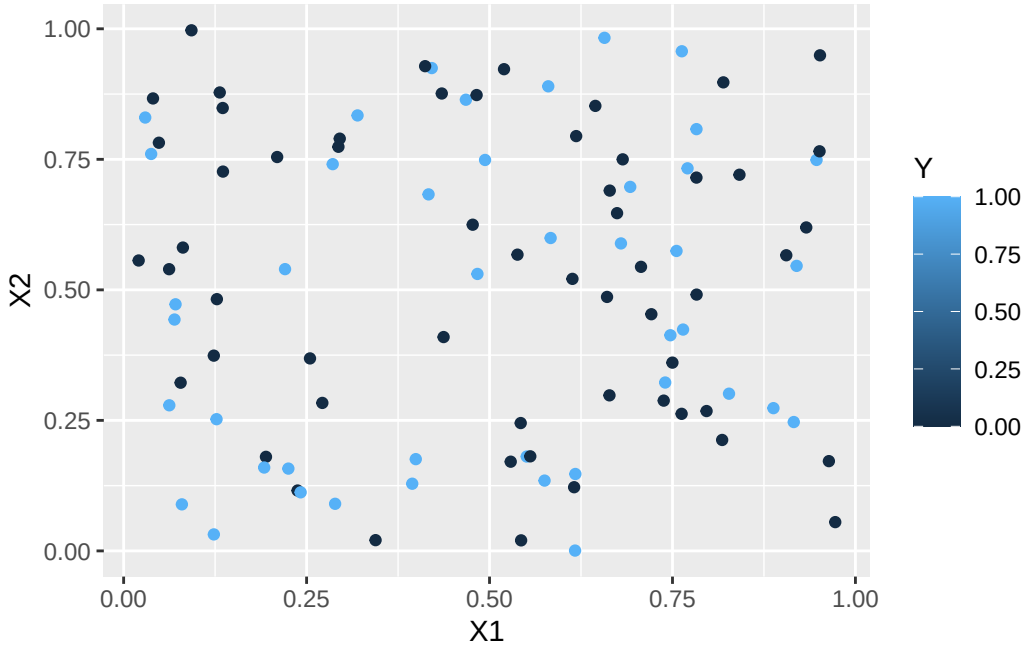
So, create the test dataset (using random seed=121) first.

```
# Create the training dataset as above using seed=116
# Create a testing dataset using seed=121

set.seed(121)
n = 100
X1 <- runif(n,0,1)
```

```r
X2 <- runif(n,0,1)

#I'm going to use a for loop to generate 100 y's
Y <- rep(0,n) #initializing my Y to be a vector of 0's
for (i in 1:n) {
  Y[i] <- generate_y(X1[i],X2[i])
}

sum(Y) #43 1's, which is much closer to the 51.5% true rate
```

```
[1] 43
```

```r
testing <- cbind(X1,X2,Y)

#Let's plot the test set. Does it look like the training set? Yeah, looks similar.
ggplot(data=testing, aes(x=X1, y=X2, color=Y)) +
  geom_point()
```



What individuals need to do

1. Given the training set (seed=116) and the testing set (seed=121), fit KNN on two different values of k.

2. Calculate the misclassification rate for each k. If you don't know how to do this, ask a teammate or the professor.
3. If possible, plot the decision boundaries for your k values.
4. Summarize the Q1, Q2, and Q3 aspects of this "project." Use your imagination. Everyone should have a different scenario.
5. Think about and discuss with your team some of the bonus questions below.

What teams need to do

1. Find the best k. Plot the decision boundary for that k.
2. Answer as many of the bonus questions as you can.

Bonus questions

Ultimately, we would like to know more about this problem, but we don't necessarily have all the tools yet to answer all of our questions. Here are some bonus questions:

- What is the misclassification rate for k=1…2j, where j is the "optimal" k. That is, for a whole range of k values?
- What is the Bayes decision boundary for the optimal k?
- How is the misclassification rate broken down into variance, bias, and irreducible error of the KNN estimator? Related to these questions are how the misclassification rate changes when we use different training or test sets.
- What happens to the misclassification rate if we go outside the $[0, 1]^2$ parameter space? I.e., if x_1 and x_2 are < 0 or > 1?

Note: the intended audience for your team document is your teammates, the professor, and your future self.

Some intended outcomes from this assignment:

- You and your team will learn how to effectively edit a document collaboratively
- You will think about what you want to accomplish in life and how this course relates to that
- Your teammates, the professor, and the TA will learn something more about you
- You will get more experience applying statistical learning methods
- You will gain experience collaborating with your teammates on an applied problem
- Get practice evaluating a method, specifically KNN
- Practice thinking about the whole problem (i.e., Q1Q2Q3, not just the Q2 quantitative parts)

Individual answers

Jacob Krol

Non-technical answers section (Jake)

Photo of myself (Jake)

What non-statistic question would I like to answer using statistical learning? (Jake)

The non-statistics question I would like to answer is ...

Are there genefusion mutations driving human cancers that have yet to be discovered, and can we find them using sequencing data from thousands of people?

What would I like to do after graduation? (Jake)

Six months after graduation, I would love to have both stable income, start acquiring more financial assets, and start a family with my partner.

Five years after graduation, I would love to own a home, and I hope to balance my family and work time well.

What is my most desired career accomplishment? (Jake)

My most desired career accomplishment would be impactful publish papers or patents that directly improve diagnosis/treatment of genetic diseases.

How do my hopes and goals relate to the course? (Jake)

I am hoping to better understand the underlying math of statistical learning methods. Particularly, I am interested in models with high interpretability that I can apply to my Ph.D. research focused on human genetic diseases.

Additional facts about me (Jake)

I love to cook, but I am not very good at it.

Technical answers section (Jake)

Plausible story for KNN data (Jake)

- The outcome, Y, is a binary indicator of whether a patient has a disease. 0 means the patient does not have the disease, and 1 means the patient does have the disease.
- The variable X1 is the outcome of a diagnostic test (e.g., bloodwork) for a disease. The test score ranges from 0 to 1. 0 is the lowest confidence that the patient has the disease. 1 implies the highest confidence that the patient has the disease.
- The variable X2 is doctor's belief that the patient has the disease. Belief also ranges from 0 to 1 with 0 being the lowest confidence and 1 being the highest confidence.

A hosptial is considering using their historical electronic health records (EHR) data on

Q1,Q2,Q3 for story (Jake)

Q1 (Jake)

The domain of the problem is healthcare, specifically disease diagnosis. The goal of the study is to develop an accurate, automated approach to predict whether a patient has a disease given the diagnostic test outcome and the doctor's belief about their disease state. An important context is that diseases often have serious health implications that could change the patient's life. Both type 1 and type 2 error of diagnosis could cause major issues. False positives (Type I error) may lead to a harsh treatment being assigned to a patient without a disease. While for false negatives (Type II error), a patient's disease state may worsen because they do not receive timely treatment. The data about test results, doctor's beliefs, and disease states are collected from a hosptial's electronic health records (EHR) system. Breifly, the plan of the study is to use historical EHR from the past few years to evaluate how diagnostic tests and doctor belief states are at predicting the true disease state of a patient. There are many ethical considerations in this study. Here, I discuss only a few. Will doctors and patients know their EHR data is being used to develop a diagnostic model? Will both doctors and patients know when a model is being used to assist in diagnosis? What is an acceptable accuracy of a diagnostic model? Who will take the blame for incorrect model predictions? For which diseases is it acceptable to use a model for diagnosis?

Q2 (Jake)

Hypothetically, the data would be carefully extracted form the EHR system and processed into a clean 3 column table with columns X1, X2, and Y. 50% of the data would be used for training and 50% for testing. A scatterplot visualization with axes X1 and X2 colored by Y would be used to explore the data. For modeling, a KNN model would be fit to the data with k=3 and k=7. The misclassification rate would be calculated for both k values. Finally, we are interested in inferring the disease state for a new patient where both the test and doctor belief are in disease is 0.5. In addition to the ethical considerations for Q1, another ethical consideration is the choice of KNN which is has no parametric interpretability. All data processing, model fitting, evaluation, and inference would be contained in a single script for reproducibility.

Q3 (Jake)

Interpreted findings include the misclassification rates for k=3 and k=7, the decision boundary plots for both k values, and the predicted disease state for a new patient with test and doctor belief both equal to 0.5. Conclusions will be drawn by considering the misclassification rates and decision boundaries. If the misclassification rates are low, then KNN may be a good choice for disease diagnosis. If the decision boundaries are smooth and reasonable, then KNN may be a good choice for disease diagnosis. The predicted disease state for the new patient will be used to illustrate how the model can be used for inference. A key ethical implications is the potential consequences of incorrect model predictions on patient health. Furthermore, if

misclassificaiton rate is low, then we may recommend using KNN as a diagnostic tool. A call to action is only necessary if th model proves to be accurate enough for clinical use. A reasonable implementation may be adapting this quarto document into a modular pipeline to automate integration of new EHR data, model fitting, evaluation, and inference.

KNN misclassification rates for k=3 and k=7 (Jake)

Train data.

```
library(class)
library(tidyverse)
#Generative model
set.seed(116) #setting a random seed so that we can reproduce everything exactly if we want t

generate_y <- function(x1,x2) { #two input parameters to generate the output y
  logit <- x1 -2*x2 -2*x1^2 + x2^2 + 3*x1*x2
  p <- exp(logit)/(1+exp(logit)) #apply the inverse logit function
  y <- rbinom(1,1,p) #y becomes a 0 (with prob 1-p) or a 1 with probability p
}
# Generate a dataset with 100 points
set.seed(116)
n = 100
X1_tr <- runif(n,0,1)
X2_tr <- runif(n,0,1)

#I'm going to use a for loop to generate 100 y's
Y_tr <- rep(0,n) #initializing my Y to be a vector of 0's
for (i in 1:n) {
  Y_tr[i] <- generate_y(X1_tr[i],X2_tr[i])
}

training <- cbind(X1_tr,X2_tr,Y_tr) #combining all of my variables into a training dataset
```

Test data

```
set.seed(121)
n = 100
X1_te <- runif(n,0,1)
X2_te <- runif(n,0,1)

Y_te <- rep(0,n) #initializing my Y to be a vector of 0's
for (i in 1:n) {
```

```
  Y_te[i] <- generate_y(X1_te[i],X2_te[i])
}

sum(Y) #43 1's, which is much closer to the 51.5% true rate
```

```
[1] 43
```

```
testing <- cbind(X1_te,X2_te,Y_te)
```

Evaluate KNN for k=3 and k=7.

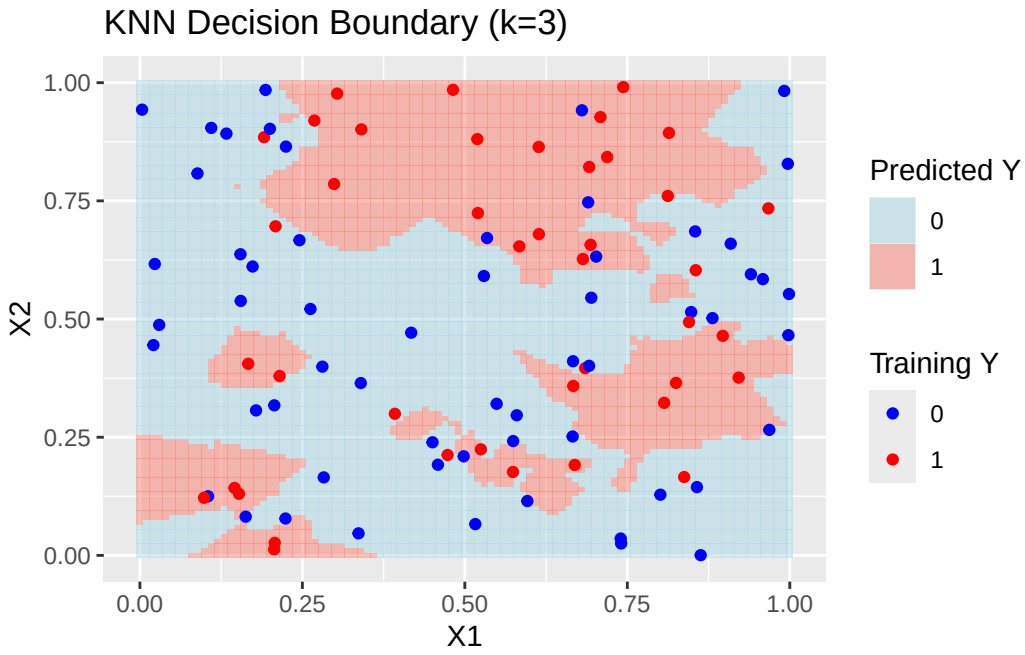
```
ks=c(3,7)
mcrate <- c()

n_test <- nrow(testing)
for (k in ks) {
  y_pred <- knn(train = training[,1:2], test = testing[,1:2],
                        cl = training[,3], k = k)

  misclass_rate <- sum(y_pred != testing[,3]) / n_test
  mcrate <- c(mcrate, misclass_rate)
}
# print misclassification rates
for (i in 1:length(ks)) {
  cat("k =", ks[i], "misclassification rate:", mcrate[i], "\n")
}
```

```
k = 3 misclassification rate: 0.47
k = 7 misclassification rate: 0.43
```

KNN decision boundary for k=3 and k=7

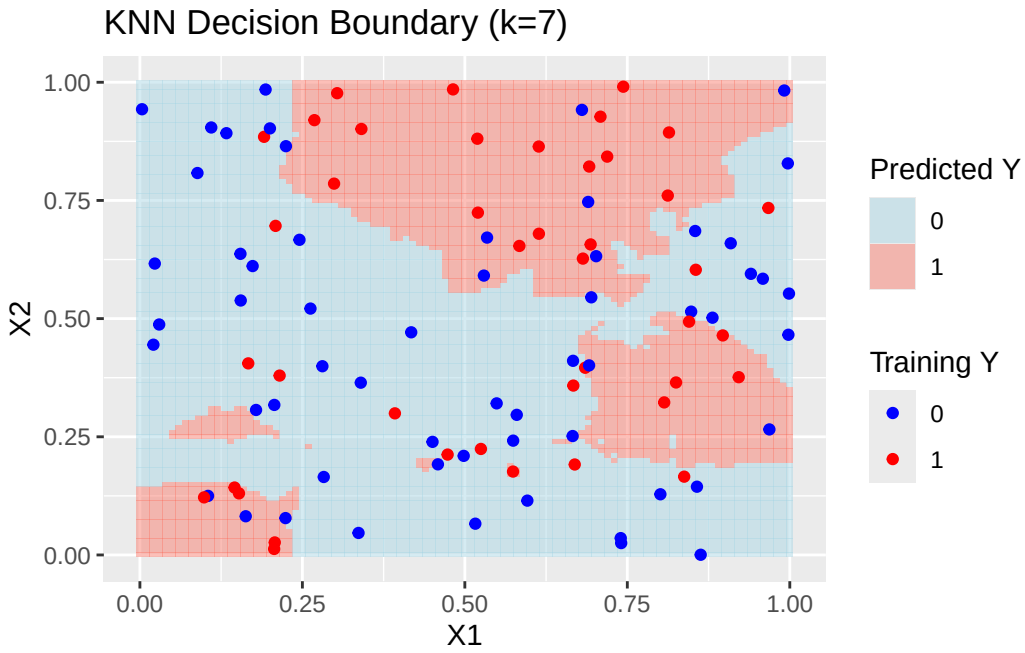
```
# grid unit square
test_grid <- expand.grid(
  X1 = seq(0, 1, by = 0.01),
  X2 = seq(0, 1, by = 0.01)
)

test_grid$y_pred_k3 <- knn(
  train= training[, 1:2],
```

```
  test= test_grid,
  cl= training[, 3],
  k=3
)
test_grid$y_pred_k7 <- knn(
  train= training[, 1:2],
  test= test_grid[, 1:2],
  cl= training[, 3],
  k=7
)

# plot db for k=3
ggplot() +
  geom_tile(data = test_grid, aes(x = X1, y = X2, fill = y_pred_k3), alpha = 0.5) +
  geom_point(data = as.data.frame(training), aes(x = X1_tr, y = X2_tr, color = as.factor(Y_tr
  labs(title = "KNN Decision Boundary (k=3)", fill = "Predicted Y", color = "Training Y") +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "salmon")) +
  scale_color_manual(values = c("0" = "blue", "1" = "red"))
```




```
# plot db for k=7
ggplot() +
  geom_tile(data = test_grid, aes(x = X1, y = X2, fill = y_pred_k7), alpha = 0.5) +
```

```
  geom_point(data = as.data.frame(training), aes(x = X1_tr, y = X2_tr, color = as.factor(Y_tr
  labs(title = "KNN Decision Boundary (k=7)", fill = "Predicted Y", color = "Training Y") +
  scale_fill_manual(values = c("0" = "lightblue", "1" = "salmon")) +
  scale_color_manual(values = c("0" = "blue", "1" = "red"))
```



KNN prediction for new data point (X1=0.5, X2=0.5)

```
new_data <- data.frame(X1 = 0.5, X2 = 0.5)
predicted_y_k3 <- knn(train = training[, 1:2],
                      test = new_data,
                      cl = training[, 3],
                      k = 3)
predicted_y_k7 <- knn(train = training[, 1:2],
                      test = new_data,
                      cl = training[, 3],
                      k = 7)
cat("Predicted Y for (X1=0.5, X2=0.5) with k=3:", predicted_y_k3, "\n")
```

```
Predicted Y for (X1=0.5, X2=0.5) with k=3: 1
```

```r
cat("Predicted Y for (X1=0.5, X2=0.5) with k=7:", predicted_y_k7, "\n")
```

```
Predicted Y for (X1=0.5, X2=0.5) with k=7: 1
```

```r
# 1 for both k=3 and k=7
```

Cleanup (Jake)

```r
rm(list = ls())
```

Summary

I used KNN to predict a binary disease state based on a diagnostic test result and doctor's belief. Misclassification rates were 0.47 and 0.43 for k=3 and k=7, respectively. Since both classes' features were generated from a random uniform distribution, I did not expect much better performance than a random model: misclassification rate = 0.5. Qualitatively, decision boundary plots show that misclassification often occurs near a boundary. Finally, I evaluated a new patient test point with diagnostic test and doctor's belief both at 0.5, KNN predicted a disease state of 1 for both k values. This is bad because X1,X2=0.5 implies that both the test and the doctor were uncertain about the diagnosis, yet the model is forced to pick a definite outcome; in this case, the model chose to diagnose the disease. In my opinion, the misclassification performance being close to a baseline model is not sufficient for real clinical use. Furthermore, the ethical implications of incorrect model predictions on patient health are serious. Therefore, I would not recommend using KNN as a diagnostic tool in this scenario.