K-Nearest Neighbor

Nate Wells

Math 243: Stat Learning

September 10th, 2021

Outline

In today's class, we will...

- Discuss the Bayes Classifier
- Implement KNN as estimate for Bayes Classifier

Section 1

The Bayes Classifier

The Task

Suppose Y is categorical response variable with several levels $A_1, \ldots, A_k$.

Goal: Build a model f to classify an observation into levels A or B based on the values of several predictors $X_1, X_2, \ldots, X_p$ (quantitative or categorical)

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \qquad \text{where } f, \epsilon \text{ take values in } \{A_1, \ldots, A_k\}$$

The Task

Suppose Y is categorical response variable with several levels $A_1, \ldots, A_k$.

Goal: Build a model f to classify an observation into levels A or B based on the values of several predictors $X_1, X_2, \ldots, X_p$ (quantitative or categorical)

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \qquad \text{where } f, \epsilon \text{ take values in } \{A_1, \ldots, A_k\}$$

How do we measure accuracy of our model?

The Task

Suppose Y is categorical response variable with several levels $A_1, \ldots, A_k$.

Goal: Build a model f to classify an observation into levels A or B based on the values of several predictors $X_1, X_2, \ldots, X_p$ (quantitative or categorical)

$$Y = f(X_1, X_2, \ldots, X_p) + \epsilon \qquad \text{where } f, \epsilon \text{ take values in } \{A_1, \ldots, A_k\}$$

How do we measure accuracy of our model?

- Training data: Compute error rate on observations in training data:

$$\text{Training Error} = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

where $I(y_i \neq \hat{y}_i)$ is the indicator variable that equals 1 if $y_i \neq \hat{y}_i$ and 0 otherwise.

The Task

Suppose Y is categorical response variable with several levels $A_1, \dots, A_k$.

Goal: Build a model f to classify an observation into levels A or B based on the values of several predictors $X_1, X_2, \dots, X_p$ (quantitative or categorical)

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon \qquad \text{where } f, \epsilon \text{ take values in } \{A_1, \dots, A_k\}$$

How do we measure accuracy of our model?

- Training data: Compute error rate on observations in training data:

$$\text{Training Error} = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

  where $I(y_i \neq \hat{y}_i)$ is the indicator variable that equals 1 if $y_i \neq \hat{y}_i$ and 0 otherwise.

- Test data: Compute average proportion of errors on test data

$$\text{Test Error} = \text{Avg. } I(y_0 \neq \hat{y}_0)$$

  where $\hat{y}_0$ is the predicted class for a test observation with predictor x_0.

The Best Possible Model

In general, the value of a response Y may depend on more than just the values of the predictors $X_1, \ldots, X_p$ in a model.

The Best Possible Model

In general, the value of a response Y may depend on more than just the values of the predictors $X_1, \dots, X_p$ in a model.

- That is, given the value of predictors x_0, the value of the response y_0 is random.

The Best Possible Model

In general, the value of a response Y may depend on more than just the values of the predictors $X_1, \ldots, X_p$ in a model.

- That is, given the value of predictors x_0, the value of the response y_0 is random.

We can show that the model which minimizes test error is

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

The Best Possible Model

In general, the value of a response Y may depend on more than just the values of the predictors $X_1, \ldots, X_p$ in a model.

- That is, given the value of predictors x_0, the value of the response y_0 is random.

We can show that the model which minimizes test error is

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

- A proof can be found on p. 18-22 of Elements of Statistical Learning (req. Math 391)

The Best Possible Model

In general, the value of a response Y may depend on more than just the values of the predictors $X_1, \ldots, X_p$ in a model.

- That is, given the value of predictors x_0, the value of the response y_0 is random.

We can show that the model which minimizes test error is

$$f(x_0) = \mathrm{argmax}_j P(Y = A_j \mid X = x_0)$$

- A proof can be found on p. 18-22 of Elements of Statistical Learning (req. Math 391)
- In practice, we cannot build this optimal model, since we don't know $P(Y = A_j \mid X = x_0)$

Simulation

Suppose Y takes two values A and B, and X_1 and X_2 are predictors taking values in $[0, 1]$.

Simulation

Suppose Y takes two values A and B, and X_1 and X_2 are predictors taking values in $[0, 1]$.

Moreover, suppose the probability $Y = A$ given $X_1 = x_1$ and $X_2 = x_2$ is $(x_1^2 + x_2^2)/2$

Simulation

Suppose Y takes two values A and B, and X_1 and X_2 are predictors taking values in $[0, 1]$.

Moreover, suppose the probability $Y = A$ given $X_1 = x_1$ and $X_2 = x_2$ is $(x_1^2 + x_2^2)/2$

```
set.seed(1)
n<-200
x1<-runif(n, 0,1 )
x2<-runif(n, 0,1)
p<-(x1^2 + x2^2)/2
```

Simulation

Suppose Y takes two values A and B, and X_1 and X_2 are predictors taking values in $[0,1]$.

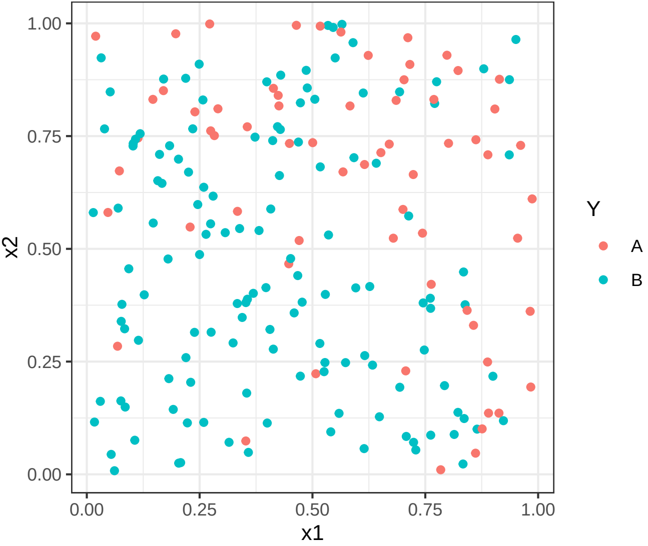
Moreover, suppose the probability $Y = A$ given $X_1 = x_1$ and $X_2 = x_2$ is $(x_1^2 + x_2^2)/2$

```
set.seed(1)
n<-200
x1<-runif(n, 0,1 )
x2<-runif(n, 0,1)
p<-(x1^2 + x2^2)/2
```

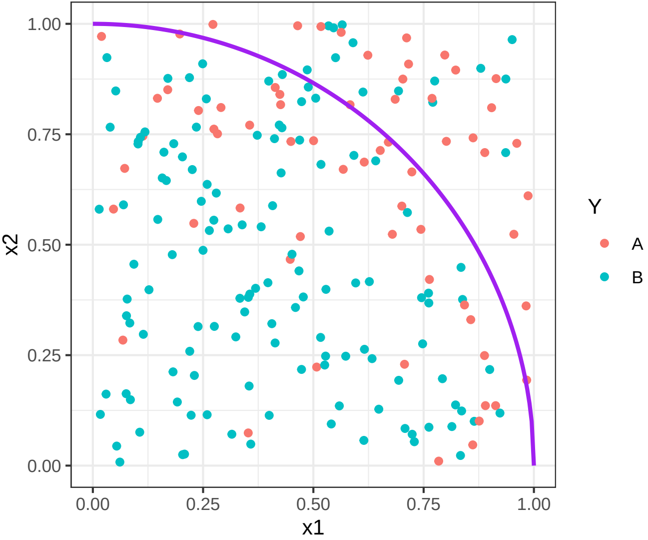
Then

$$f(x_0) = \operatorname{argmax}_j P(Y = A_j \mid X = x_0) = \begin{cases} A, & \text{if } x_1^2 + x_2^2 \geq 1 \\ B, & \text{if } x_1^2 + x_2^2 < 1 \end{cases}$$

Plot 1



Plot 2



Expected Error Rate

In general, using the Bayes Classifier produces an expected error rate of

$$1 - \text{Avg.}\left(\max_j \mathrm{P}(Y = A_j \mid X = x_0)\right)$$

Expected Error Rate

In general, using the Bayes Classifier produces an expected error rate of

$$1 - \mathrm{Avg.}\left(\max_j \mathrm{P}(Y = A_j \mid X = x_0)\right)$$

For our simulation, this gives an error of $1/3$.

Expected Error Rate

In general, using the Bayes Classifier produces an expected error rate of

$$1 - \text{Avg.}\left(\max_j \mathrm{P}(Y = A_j \mid X = x_0)\right)$$

For our simulation, this gives an error of $1/3$.

- Can verify using multivariate calculus or by sampling a large number of times.

Expected Error Rate

In general, using the Bayes Classifier produces an expected error rate of

$$1 - \mathrm{Avg.}\left(\max_j \mathrm{P}(Y = A_j \mid X = x_0)\right)$$

For our simulation, this gives an error of $1/3$.

- Can verify using multivariate calculus or by sampling a large number of times.

This is the theoretical lower bound on average error for a classification problem.

Section 2

K-Nearest Neighbors

From Bayes Classifier to KNN

In theory, the Bayes Classifier is our best model for classification.

From Bayes Classifier to KNN

In theory, the Bayes Classifier is our best model for classification.

- In practice, we don't know the conditional probability of Y given X, and so cannot build a Bayes Classifier model.

From Bayes Classifier to KNN

In theory, the Bayes Classifier is our best model for classification.

- In practice, we don't know the conditional probability of Y given X, and so cannot build a Bayes Classifier model.
- But given sufficient data, we can *estimate* the conditional probabilities (assuming they are generated by a continuous function).

From Bayes Classifier to KNN

In theory, the Bayes Classifier is our best model for classification.

- In practice, we don't know the conditional probability of Y given X, and so cannot build a Bayes Classifier model.
- But given sufficient data, we can *estimate* the conditional probabilities (assuming they are generated by a continuous function).

Given a positive integer K and a test observation x_0, let N_0 denote the K nearest training observations to x_0. Then

$$P(Y = A_j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

From Bayes Classifier to KNN

In theory, the Bayes Classifier is our best model for classification.

- In practice, we don't know the conditional probability of Y given X, and so cannot build a Bayes Classifier model.
- But given sufficient data, we can *estimate* the conditional probabilities (assuming they are generated by a continuous function).

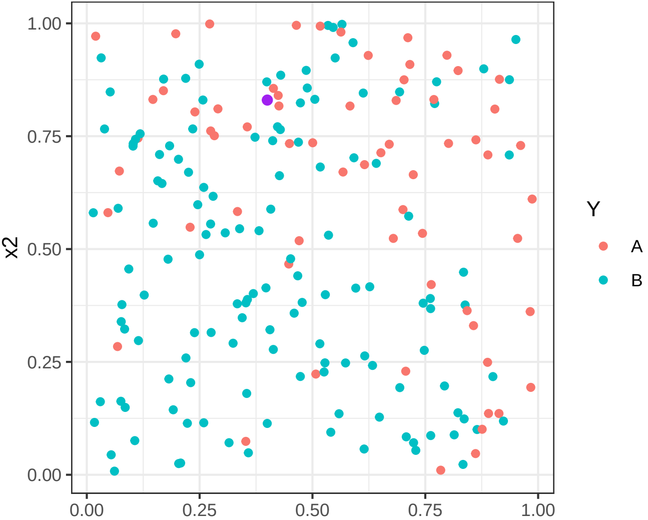
Given a positive integer K and a test observation x_0, let N_0 denote the K nearest training observations to x_0. Then

$$P(Y = A_j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$$

- Our model is therefore $f(x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$.

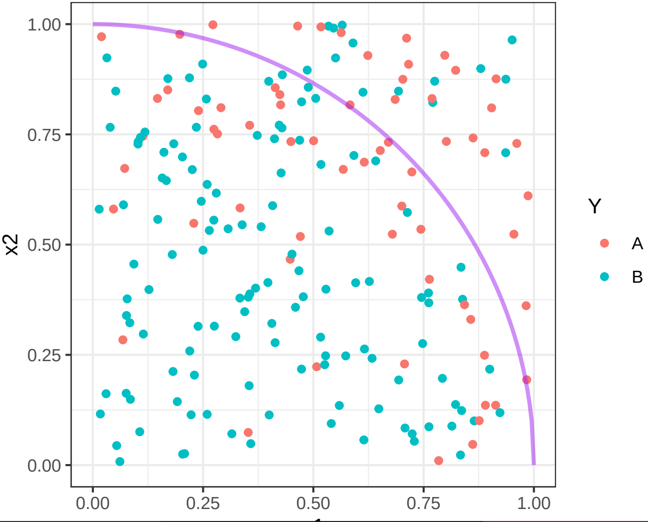
Simulation

Classify x_0 for $K = 1, 2, 3, 5, 10, 200$.



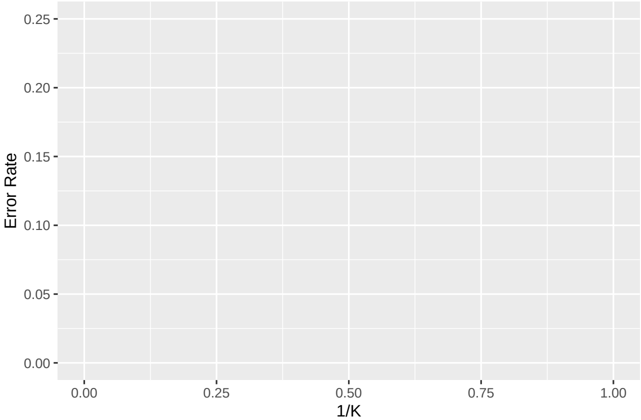
Simulation

Sketch the classification boundaries for a variety of values of K.



Error Rates

Sketch the graph of KNN error rates as function of K^{-1}



Extra Practice

Use the first part of the .Rmd file on the course website to generate 5 random points and form classification boundaries for $K = 1$ and $K = 2$ KNN.

Then use the second part of the .Rmd file to classify 5 randomly generated points.