# 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 $A_1, \ldots, A_k$, and that $X_1, \ldots, X_p$ are predictors (either categorical or quantitative).

- Assume that the level of $Y$ is not completely determined by the values of $X_1, \ldots, X_p$.

# The Task

Suppose $Y$ is categorical response variable with several levels $A_1, \ldots, A_k$, and that $X_1, \ldots, X_p$ are predictors (either categorical or quantitative).

- Assume that the level of $Y$ is not completely determined by the values of $X_1, \ldots, X_p$.

Goal: Build a model $\hat{g}(X_1, \ldots, X_p)$ that takes values in $\{A_1, \ldots, A_p\}$ that can be used to predict the class of $Y$ based on $X_1, \ldots, X_p$.

## The Task

Suppose $Y$ is categorical response variable with several levels $A_1, \ldots, A_k$, and that $X_1, \ldots, X_p$ are predictors (either categorical or quantitative).

- Assume that the level of $Y$ is not completely determined by the values of $X_1, \ldots, X_p$.

Goal: Build a model $\hat{g}(X_1, \ldots, X_p)$ that takes values in $\{A_1, \ldots, A_p\}$ that can be used to predict the class of $Y$ based on $X_1, \ldots, X_p$.

- How do we measure accuracy of our model? - Why not MSE?

# The Task

Suppose $Y$ is categorical response variable with several levels $A_1, \ldots, A_k$, and that $X_1, \ldots, X_p$ are predictors (either categorical or quantitative).

- Assume that the level of $Y$ is not completely determined by the values of $X_1, \ldots, X_p$.

Goal: Build a model $\hat{g}(X_1, \ldots, X_p)$ that takes values in $\{A_1, \ldots, A_p\}$ that can be used to predict the class of $Y$ based on $X_1, \ldots, X_p$.

- How do we measure accuracy of our model? - Why not MSE?
- Training data: Compute error rate on observations in training data:

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

where $I\big(y_i \neq \hat{g}(x_i)\big)$ equals 1 if $y_i \neq \hat{g}(x_i)$ and 0 otherwise.

## The Task

Suppose $Y$ is categorical response variable with several levels $A_1, \ldots, A_k$, and that $X_1, \ldots, X_p$ are predictors (either categorical or quantitative).

- Assume that the level of $Y$ is not completely determined by the values of $X_1, \ldots, X_p$.

Goal: Build a model $\hat{g}(X_1, \ldots, X_p)$ that takes values in $\{A_1, \ldots, A_p\}$ that can be used to predict the class of $Y$ based on $X_1, \ldots, X_p$.

- How do we measure accuracy of our model? - Why not MSE?
- Training data: Compute error rate on observations in training data:

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

  where $I\big(y_i \neq \hat{g}(x_i)\big)$ equals 1 if $y_i \neq \hat{g}(x_i)$ and 0 otherwise.

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

$$\text{Test Error} = \text{Avg. } I\big(y_i \neq \hat{g}(x_0)\big)$$

  with the average taken across many test observations $x_0$.

## The Best Possible Model

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

The model (called the **Bayes Classifier**) which minimizes test error is

$$g(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

# 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.

The model (called the **Bayes Classifier**) which minimizes test error is

$$g(x_0) = \mathrm{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

- This model assigns $Y$ to the most likely class, given the value of $x_0$.
- A proof can be found on p. 18-22 of Elements of Statistical Learning (uses tools from adv. probability)

## 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.

The model (called the **Bayes Classifier**) which minimizes test error is

$$g(x_0) = \text{argmax}_{A_j} P(Y = A_j \mid X = x_0)$$

- This model assigns $Y$ to the most likely class, given the value of $x_0$.
- A proof can be found on p. 18-22 of Elements of Statistical Learning (uses tools from adv. probability)
- In practice, we cannot build this optimal model, since we don't know know the formula for $P(Y = A_j \mid X = x_0)$

# Simulation

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

## Simulation

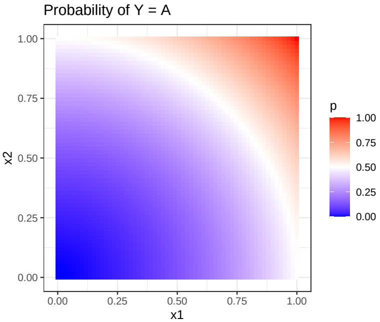
- Suppose $Y$ takes values $A$ or $B$, and $X_1$ and $X_2$ are predictors taking values in $[0, 1]$.
- Additionally, suppose that if $X_1 = x_1$ and $X_2 = x_2$, then $Y = A$ with probability

$$p = (x_1^2 + x_2^2)/2$$

## Simulation

- Suppose $Y$ takes values $A$ or $B$, and $X_1$ and $X_2$ are predictors taking values in $[0, 1]$.
- Additionally, suppose that if $X_1 = x_1$ and $X_2 = x_2$, then $Y = A$ with probability

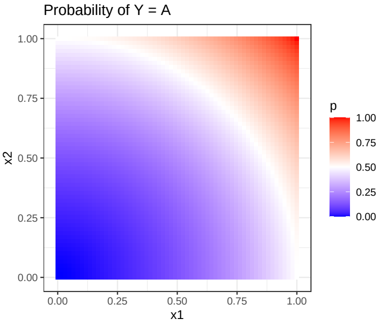
$$p = (x_1^2 + x_2^2)/2$$



# Simulation

- Suppose $Y$ takes values $A$ or $B$, and $X_1$ and $X_2$ are predictors taking values in $[0, 1]$.
- Additionally, suppose that if $X_1 = x_1$ and $X_2 = x_2$, then $Y = A$ with probability

$$p = (x_1^2 + x_2^2)/2$$

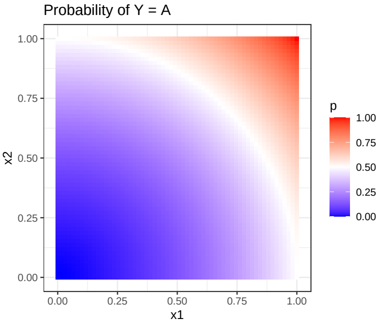



- What is the Bayes Classifier $g$?

# Simulation

- Suppose $Y$ takes values $A$ or $B$, and $X_1$ and $X_2$ are predictors taking values in $[0,1]$.
- Additionally, suppose that if $X_1 = x_1$ and $X_2 = x_2$, then $Y = A$ with probability

$$p = (x_1^2 + x_2^2)/2$$

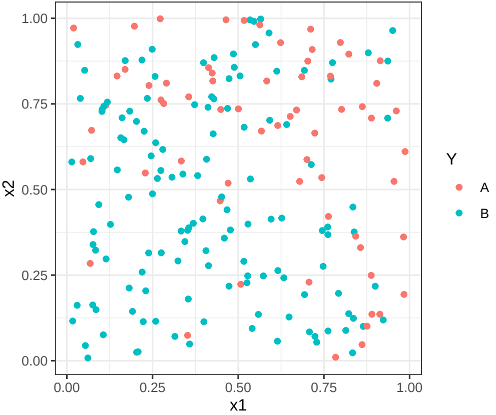



- What is the Bayes Classifier $g$?

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

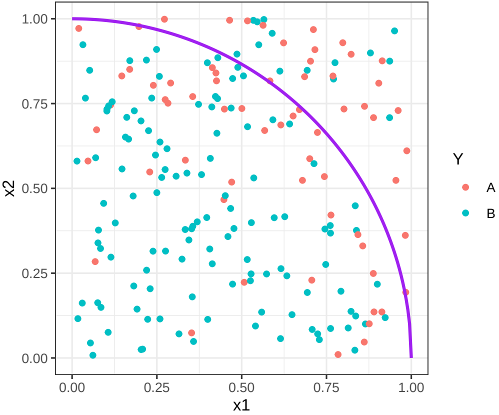
# Simulate Data

Let's simualte 200 data points from this model.



# The Bayes Classifier

The purple arc represents the Bayes Classifier boundary



# Expected Error Rate

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

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

## 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 $\frac{2}{3} - \frac{\pi}{8} \approx 0.274$.

## 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 $\frac{2}{3} - \frac{\pi}{8} \approx 0.274$.
    - 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 $\frac{2}{3} - \frac{\pi}{8} \approx 0.274$.
    - Can verify using multivariate calculus or by sampling a large number of times.
- This is the theoretical lower bound on average test error for this classification problem.

# 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 $\frac{2}{3} - \frac{\pi}{8} \approx 0.274$.
    - Can verify using multivariate calculus or by sampling a large number of times.
- This is the theoretical lower bound on average test error for this classification problem.
    - This is analogous to the irreducible error in regression problems

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 for $P$ is therefore $\hat{P}_j(x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$.

## 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).

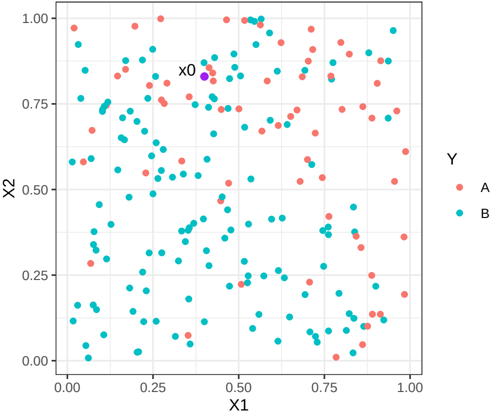
Given a positive integer $K$ and a test observation $x_0$, let $N_0$ denote the $K$ nearest training observations to $x_0$. Then

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

- Our model for $P$ is therefore $\hat{P}_j(x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = A_j)$.
- And our classifier model is $\hat{g}(x_0) = \mathrm{argmax}_{A_j} \hat{P}_j(x_0)$

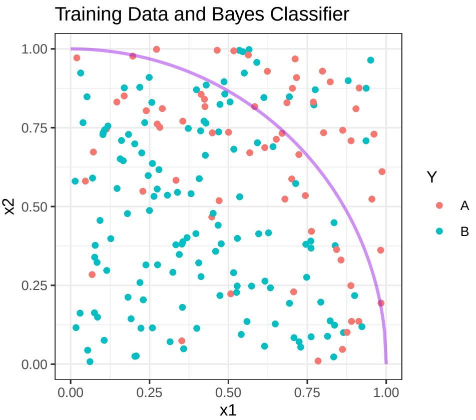
# Classify Points

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



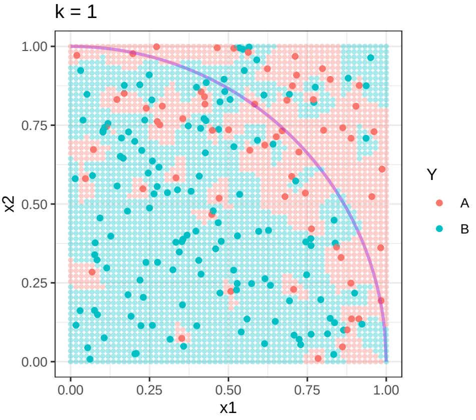
# Classification Boundaries

Here are the classification boundaries for a variety of values of $K$.



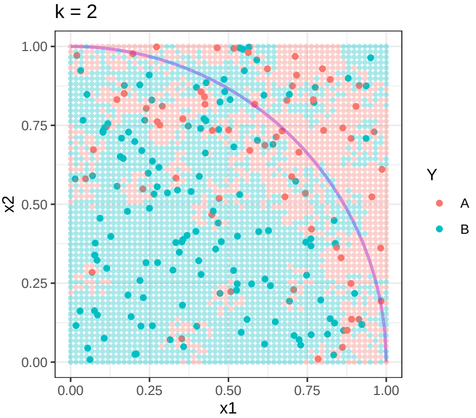
# k=1

Here are the classification boundaries for a variety of values of $K$.



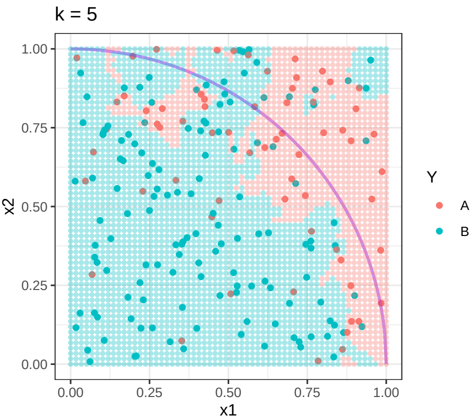
# k=2

Here are the classification boundaries for a variety of values of $K$.



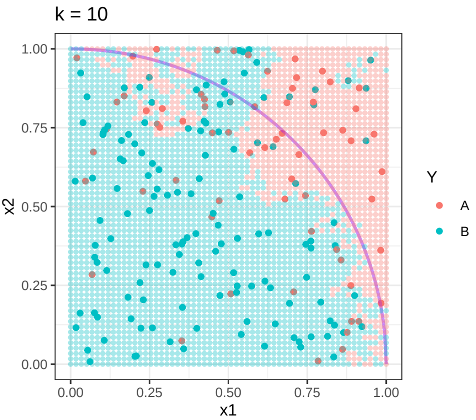
# k=5

Here are the classification boundaries for a variety of values of $K$.



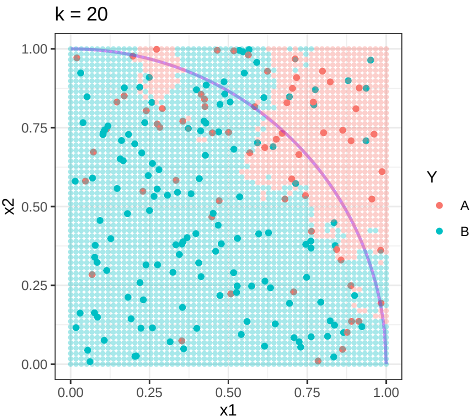
# k=10

Here are the classification boundaries for a variety of values of $K$.



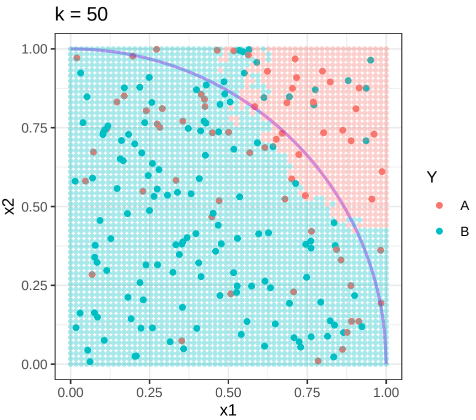
# k=20

Here are the classification boundaries for a variety of values of $K$.



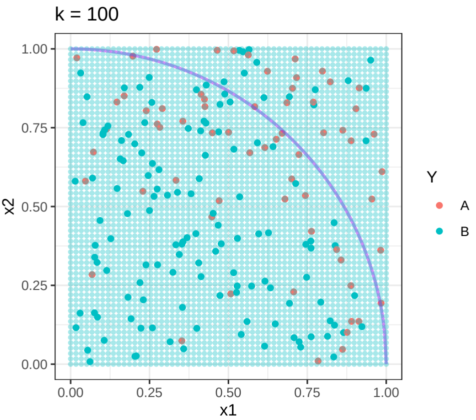
# k=50

Here are the classification boundaries for a variety of values of $K$.



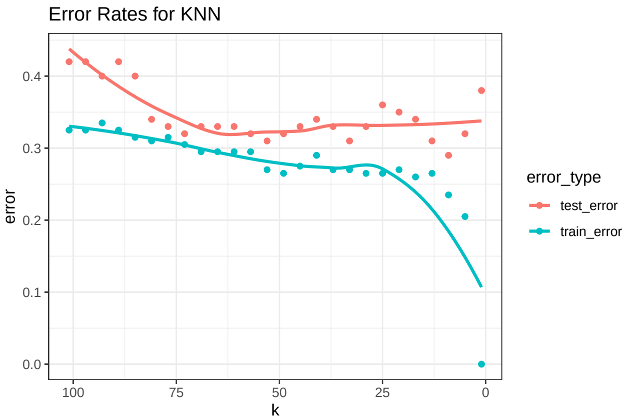
# k=100

Here are the classification boundaries for a variety of values of $K$.



# Error Rates

The graph below shows error rates for the training set, as well as a test set of 100 points.



## Extra Practice

1. Use the first part of the .Rmd file on the course website to generate 4 random points and form classification boundaries for $K = 1$ and $K = 2$ KNN.
2. Then use the second part of the .Rmd file to classify 5 randomly generated points.