# Naive Bayes

Prof Wells

STA 295: Stat Learning

April 11th, 2024

## Outline

- Review elements of probability theory
- Discuss Naive Bayes theory and motivation
- Implement Naive Bayes in R

Section 1

# Probability Theory

# Bayes Rule

**Bayes Rule**: For any two events $E$ and $B$,

$$P(E|B) = \frac{P(B|E)P(E)}{P(B)}$$

## Bayes Rule

**Bayes Rule**: For any two events $E$ and $B$,

$$P(E|B) = \frac{P(B|E)P(E)}{P(B)}$$

- $P(E)$ is called the *prior probability* of $E$ and represents our initial beliefs about the chances that event $E$ occurs.
- Suppose $B$ is an event that we observe occurring.
- $P(E|B)$ is called the *posterior probability* of $E$ and represents our updated beliefs about the chances that event $E$ occurs, knowing that event $B$ occurred.
- $P(B|E)/P(B)$ is called the *Bayes Factor* and represents the likelihood that $B$ occurs given $E$ occurred relative to the probability of $B$ occurring among all possible scenarios.

## Bayes Rule

**Bayes Rule**: For any two events $E$ and $B$,

$$P(E|B) = \frac{P(B|E)P(E)}{P(B)}$$

- $P(E)$ is called the *prior probability* of $E$ and represents our initial beliefs about the chances that event $E$ occurs.
- Suppose $B$ is an event that we observe occurring.
- $P(E|B)$ is called the *posterior probability* of $E$ and represents our updated beliefs about the chances that event $E$ occurs, knowing that event $B$ occurred.
- $P(B|E)/P(B)$ is called the *Bayes Factor* and represents the likelihood that $B$ occurs given $E$ occurred relative to the probability of $B$ occurring among all possible scenarios.
- Bayes Rule follows from the definition of conditional probability:

$$P(E|B) = \frac{P(E \text{ and } B)}{P(B)} \qquad P(B|E) = \frac{P(E \text{ and } B)}{P(E)}$$

# Law of Total Probability

Bayes Rule is most often combined with another powerful probability result:

Suppose $E_1, E_2, \ldots, E_k$ are a list of events that are:

- *mutually exclusive*: $P(E_i \text{ and } E_j) = 0$
- *exhaustive*: $P(E_1) + P(E_2) \cdots + P(E_k) = 1$
  - Example: Suppose we have two coins: one coin is double-headed, and the other coin is a fair coin. One coin is selected at random. Let $E_1$ be the event the double-headed coin is selected, and let $E_2$ be the event the fair coin is selected.

**Law of Total Probability**: For any event $B$,

$$P(B) = P(F|E_1)P(E_1) + P(F|E_2)P(E_2) + \cdots + P(F|E_k)P(E_k)$$

# Law of Total Probability

Bayes Rule is most often combined with another powerful probability result:

Suppose $E_1, E_2, \ldots, E_k$ are a list of events that are:

- *mutually exclusive*: $P(E_i \text{ and } E_j) = 0$
- *exhaustive*: $P(E_1) + P(E_2) \cdots + P(E_k) = 1$
  - Example: Suppose we have two coins: one coin is double-headed, and the other coin is a fair coin. One coin is selected at random. Let $E_1$ be the event the double-headed coin is selected, and let $E_2$ be the event the fair coin is selected.

**Law of Total Probability**: For any event $B$,

$$P(B) = P(F|E_1)P(E_1) + P(F|E_2)P(E_2) + \cdots + P(F|E_k)P(E_k)$$

## Example

Suppose we randomly select one of the two coins above. What is the probability the coin lands heads?

# Law of Total Probability

Bayes Rule is most often combined with another powerful probability result:

Suppose $E_1, E_2, \ldots, E_k$ are a list of events that are:

- *mutually exclusive*: $P(E_i \text{ and } E_j) = 0$
- *exhaustive*: $P(E_1) + P(E_2) \cdots + P(E_k) = 1$
  - Example: Suppose we have two coins: one coin is double-headed, and the other coin is a fair coin. One coin is selected at random. Let $E_1$ be the event the double-headed coin is selected, and let $E_2$ be the event the fair coin is selected.

**Law of Total Probability**: For any event $B$,

$$P(B) = P(F|E_1)P(E_1) + P(F|E_2)P(E_2) + \cdots + P(F|E_k)P(E_k)$$

## Example

Suppose we randomly select one of the two coins above. What is the probability the coin lands heads?

$$P(\text{Heads}) = P(\text{Heads}|E_1)P(E_1) + P(\text{Heads}|E_2)P(E_2) = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$$

# Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

# Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

- The prior probability of selecting the double-headed coin is $P(A_1) = \frac{1}{2}$.

# Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

- The prior probability of selecting the double-headed coin is $P(A_1) = \frac{1}{2}$.
  - Is it still reasonable to believe there is a 50% chance of having selected the double-headed coin, given we observed heads?

## Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

- The prior probability of selecting the double-headed coin is $P(A_1) = \frac{1}{2}$.
  - Is it still reasonable to believe there is a 50% chance of having selected the double-headed coin, given we observed heads?
- Observing a heads is *more consistent* with the scenario where we selected the double-headed coin, than it is in the scenario where we selected the fair coin.

## Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

- The prior probability of selecting the double-headed coin is $P(A_1) = \frac{1}{2}$.
  - Is it still reasonable to believe there is a 50% chance of having selected the double-headed coin, given we observed heads?
- Observing a heads is *more consistent* with the scenario where we selected the double-headed coin, than it is in the scenario where we selected the fair coin.
  - If the coin was double-headed, we would also flip heads. But if we had the fair coin, we would only flip heads 50 of the time.

# Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

- The prior probability of selecting the double-headed coin is $P(A_1) = \frac{1}{2}$.
  - Is it still reasonable to believe there is a 50% chance of having selected the double-headed coin, given we observed heads?
- Observing a heads is *more consistent* with the scenario where we selected the double-headed coin, than it is in the scenario where we selected the fair coin.
  - If the coin was double-headed, we would also flip heads. But if we had the fair coin, we would only flip heads 50 of the time.

Using Bayes Rule:

$$P(E_1|\mathrm{Heads}) = \frac{P(\mathrm{Heads}|E_1)P(E_1)}{P(\mathrm{Heads})} = \frac{P(\mathrm{Heads}|E_1)P(E_1)}{P(\mathrm{Heads}|E_1)P(E_1) + P(\mathrm{Heads}|E_2)P(A_2)} = \frac{1}{3/4} \cdot \frac{1}{2} = \frac{2}{3}$$

# Bayes Rule + Law of Total Probability

Suppose we randomly select on of the two coins, flip it, and observe that the coin lands heads. What is the probability that the selected coin was double-headed?

- The prior probability of selecting the double-headed coin is $P(A_1) = \frac{1}{2}$.
  - Is it still reasonable to believe there is a 50% chance of having selected the double-headed coin, given we observed heads?
- Observing a heads is *more consistent* with the scenario where we selected the double-headed coin, than it is in the scenario where we selected the fair coin.
  - If the coin was double-headed, we would also flip heads. But if we had the fair coin, we would only flip heads 50 of the time.

Using Bayes Rule:

$$P(E_1|\text{Heads}) = \frac{P(\text{Heads}|E_1)P(E_1)}{P(\text{Heads})} = \frac{P(\text{Heads}|E_1)P(E_1)}{P(\text{Heads}|E_1)P(E_1) + P(\text{Heads}|E_2)P(A_2)} = \frac{1}{3/4} \cdot \frac{1}{2} = \frac{2}{3}$$

- That is, the posterior probability $P(E_1|\text{Heads}) = \frac{2}{3}$ is larger than the prior probability $P(A_1) = \frac{1}{2}$.

# The Multiplication Rule

- Events $E_1, \ldots, E_k$ are independent if knowing that one occurred does not make it more or less likely that any of the others occurred.

# The Multiplication Rule

- Events $E_1, \ldots, E_k$ are independent if knowing that one occurred does not make it more or less likely that any of the others occurred.
  - Suppose we roll 2 dice. Let $E_1$ be the event that the first is a 6, and let $E_2$ be the event that the second is an even number. Then $E_1, E_2$ are independent, since the roll of the first die has no influence on the roll of the second die.

# The Multiplication Rule

- Events $E_1, \ldots, E_k$ are independent if knowing that one occurred does not make it more or less likely that any of the others occurred.
    - Suppose we roll 2 dice. Let $E_1$ be the event that the first is a 6, and let $E_2$ be the event that the second is an even number. Then $E_1, E_2$ are independent, since the roll of the first die has no influence on the roll of the second die.

**Multiplication Rule**: If events $E_1, \ldots, E_k$ are independent, then

$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_k) = P(E_1) \cdot P(E_2) \cdots P(E_k)$$

## The Multiplication Rule

- Events $E_1, \ldots, E_k$ are independent if knowing that one occurred does not make it more or less likely that any of the others occurred.
  - Suppose we roll 2 dice. Let $E_1$ be the event that the first is a 6, and let $E_2$ be the event that the second is an even number. Then $E_1, E_2$ are independent, since the roll of the first die has no influence on the roll of the second die.

**Multiplication Rule**: If events $E_1, \ldots, E_k$ are independent, then

$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_k) = P(E_1) \cdot P(E_2) \cdots P(E_k)$$

- In the dice example above, the probability that a 6 is rolled on the first die and an odd number is rolled on the second is

$$P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

## The Multiplication Rule

- Events $E_1, \ldots, E_k$ are independent if knowing that one occurred does not make it more or less likely that any of the others occurred.
  - Suppose we roll 2 dice. Let $E_1$ be the event that the first is a 6, and let $E_2$ be the event that the second is an even number. Then $E_1, E_2$ are independent, since the roll of the first die has no influence on the roll of the second die.

**Multiplication Rule**: If events $E_1, \ldots, E_k$ are independent, then

$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_k) = P(E_1) \cdot P(E_2) \cdots P(E_k)$$

- In the dice example above, the probability that a 6 is rolled on the first die and an odd number is rolled on the second is

$$P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

- If events are **dependent** then the multiplication rule *does not* hold. At best,

$$P(E_1 \text{ and } E_2) = P(E_1 | E_2) \cdot P(E_2)$$

# The Multiplication Rule

- Events $E_1, \ldots, E_k$ are independent if knowing that one occurred does not make it more or less likely that any of the others occurred.
  - Suppose we roll 2 dice. Let $E_1$ be the event that the first is a 6, and let $E_2$ be the event that the second is an even number. Then $E_1, E_2$ are independent, since the roll of the first die has no influence on the roll of the second die.

**Multiplication Rule**: If events $E_1, \ldots, E_k$ are independent, then

$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_k) = P(E_1) \cdot P(E_2) \cdots P(E_k)$$

- In the dice example above, the probability that a 6 is rolled on the first die and an odd number is rolled on the second is

$$P(E_1 \text{ and } E_2) = P(E_1) \cdot P(E_2) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

- If events are **dependent** then the multiplication rule *does not* hold. At best,

$$P(E_1 \text{ and } E_2) = P(E_1|E_2) \cdot P(E_2)$$

  - In order to calculate the probability both occur, we need to known about the relationship between the two events.

Section 2

# Generative Models

# Probability Models

For classification problem, average test error rate is minimized using the Bayes' classifier:

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

- i.e. predict the class that has the greatest conditional probability, given the data.

## Probability Models

For classification problem, average test error rate is minimized using the Bayes' classifier:

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

- i.e. predict the class that has the greatest conditional probability, given the data.

Both KNN and Logistic regression attempt to directly estimate the conditional probability $P(Y = A_j \mid X)$:

# Probability Models

For classification problem, average test error rate is minimized using the Bayes' classifier:

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

- i.e. predict the class that has the greatest conditional probability, given the data.

Both KNN and Logistic regression attempt to directly estimate the conditional probability $P(Y = A_j \mid X)$:

- Logistic regression:

$$P(Y = A_j \mid X) \approx \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

# Probability Models

For classification problem, average test error rate is minimized using the Bayes' classifier:

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

- i.e. predict the class that has the greatest conditional probability, given the data.

Both KNN and Logistic regression attempt to directly estimate the conditional probability $P(Y = A_j \mid X)$:

- Logistic regression:

$$P(Y = A_j \mid X) \approx \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- KNN:

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

# Probability Models

For classification problem, average test error rate is minimized using the Bayes' classifier:

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

- i.e. predict the class that has the greatest conditional probability, given the data.

Both KNN and Logistic regression attempt to directly estimate the conditional probability $P(Y = A_j \mid X)$:

- Logistic regression:

$$P(Y = A_j \mid X) \approx \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

- KNN:

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

Alternatively, we might instead model the **opposite** conditional probability: $P(X|Y = A_j)$

- This is the distribution of the predictors, within each class of the response.

## Probability Models

For classification problem, average test error rate is minimized using the Bayes' classifier:

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

- i.e. predict the class that has the greatest conditional probability, given the data.

Both KNN and Logistic regression attempt to directly estimate the conditional probability $P(Y = A_j \mid X)$:

- Logistic regression:

$$P(Y = A_j \mid X) \approx \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}$$

- KNN:

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

Alternatively, we might instead model the **opposite** conditional probability: $P(X|Y = A_j)$

- This is the distribution of the predictors, within each class of the response.
- Our goal would then be to reverse this probability to get $P(Y = A_i|X)$.

## Bayes Rule and Law of Total Probability, Again

**Goal**: Estimate $P(Y = A_j|X)$.

**Method**: estimate $P(X|Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

## Bayes Rule and Law of Total Probability, Again

**Goal**: Estimate $P(Y = A_j|X)$.

**Method**: estimate $P(X|Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

$$P(Y = A_j|X) = \frac{P(X|Y = A_j)}{P(X)} = \frac{P(X|Y = A_j)}{\sum_i P(X|Y = A_i)P(Y = A_i)}$$

# Bayes Rule and Law of Total Probability, Again

**Goal**: Estimate $P(Y = A_j|X)$.

**Method**: estimate $P(X|Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

$$P(Y = A_j|X) = \frac{P(X|Y = A_j)}{P(X)} = \frac{P(X|Y = A_j)}{\sum_i P(X|Y = A_i)P(Y = A_i)}$$

- Suppose $X$ represents a single predictor. One model assumes that $X$ is normally distributed within each class of $Y$.

# Bayes Rule and Law of Total Probability, Again

**Goal**: Estimate $P(Y = A_j|X)$.

**Method**: estimate $P(X|Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

$$P(Y = A_j|X) = \frac{P(X|Y = A_j)}{P(X)} = \frac{P(X|Y = A_j)}{\sum_i P(X|Y = A_i)P(Y = A_i)}$$

- Suppose $X$ represents a single predictor. One model assumes that $X$ is normally distributed within each class of $Y$.
  - To estimate $P(X|Y = A_i)$, we compute the mean and standard deviation of $X$ within each level of $Y$

# Bayes Rule and Law of Total Probability, Again

**Goal**: Estimate $P(Y = A_j|X)$.

**Method**: estimate $P(X|Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

$$P(Y = A_j|X) = \frac{P(X|Y = A_j)}{P(X)} = \frac{P(X|Y = A_j)}{\sum_i P(X|Y = A_i)P(Y = A_i)}$$

- Suppose $X$ represents a single predictor. One model assumes that $X$ is normally distributed within each class of $Y$.
  - To estimate $P(X|Y = A_i)$, we compute the mean and standard deviation of $X$ within each level of $Y$
  - Then we use the formula for probabilities from the Normal distribution (of the estimated mean and variance) to calculate $P(X|Y = A_i)$ for each $A_i$.

# Bayes Rule and Law of Total Probability, Again

**Goal**: Estimate $P(Y = A_j|X)$.

**Method**: estimate $P(X|Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

$$P(Y = A_j|X) = \frac{P(X|Y = A_j)}{P(X)} = \frac{P(X|Y = A_j)}{\sum_i P(X|Y = A_i)P(Y = A_i)}$$

- Suppose $X$ represents a single predictor. One model assumes that $X$ is normally distributed within each class of $Y$.
  - To estimate $P(X|Y = A_i)$, we compute the mean and standard deviation of $X$ within each level of $Y$
  - Then we use the formula for probabilities from the Normal distribution (of the estimated mean and variance) to calculate $P(X|Y = A_i)$ for each $A_i$.
- We also estimate the prior probabilities $P(Y = A_i)$ using the proportion of observations in each class of $Y$ (ignoring the predictor $X$).

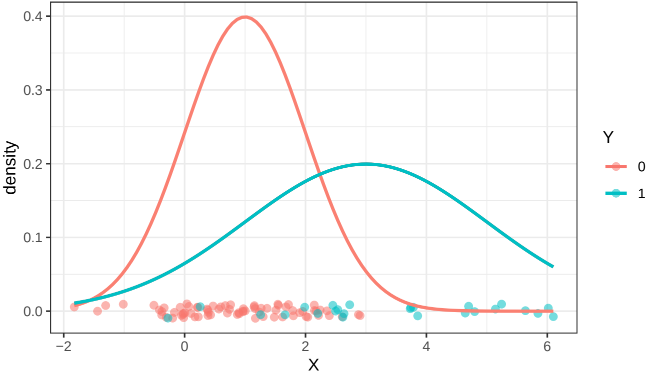
# Simulation

Consider a binary numeric response variable $Y$ and a single quantitative predictor $X$.

# Simulation

Consider a binary numeric response variable $Y$ and a single quantitative predictor $X$.

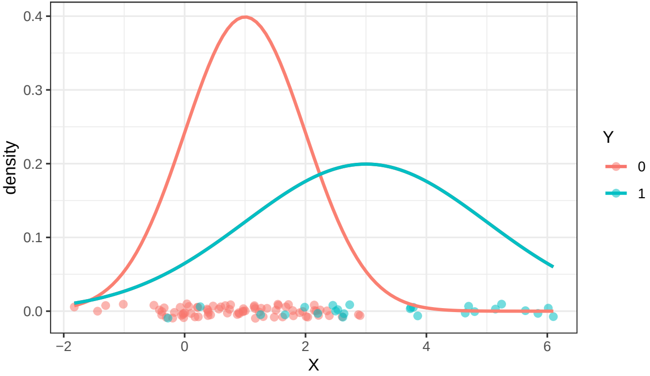
- Suppose if $Y = 0$, then $X \sim N(1, 1)$ and if $Y = 1$, then $X \sim N(3, 2)$
    - Additionally, suppose $P(Y = 0) = .75$ and $P(Y = 1) = .25$.



# Simulation

Consider a binary numeric response variable $Y$ and a single quantitative predictor $X$.

- Suppose if $Y = 0$, then $X \sim N(1, 1)$ and if $Y = 1$, then $X \sim N(3, 2)$
  - Additionally, suppose $P(Y = 0) = .75$ and $P(Y = 1) = .25$.




- What feature of the graph shows that $P(Y = 0) = .75$ and $P(Y = 1) = .25$?

# Fit Model

We calculate estimates for the mean and standard deviation of $X$, within each level of $Y$, along with the proportion of observations within each level of $Y$:

# Fit Model

We calculate estimates for the mean and standard deviation of $X$, within each level of $Y$, along with the proportion of observations within each level of $Y$:

```
sim_data %>% group_by(Y) %>% summarize(mean = mean(X), sd = sd(X), n_obs = n()) %>%
  mutate(prop = n_obs/sum(n_obs))
```

```
## # A tibble: 2 x 5
##   Y      mean    sd n_obs  prop
##   <chr> <dbl> <dbl> <int> <dbl>
## 1 0     0.828  1.03    75  0.75
## 2 1     3.43   1.78    25  0.25
```

# Fit Model

We calculate estimates for the mean and standard deviation of $X$, within each level of $Y$, along with the proportion of observations within each level of $Y$:

```
sim_data %>% group_by(Y) %>% summarize(mean = mean(X), sd = sd(X), n_obs = n()) %>%
  mutate(prop = n_obs/sum(n_obs))
```

```
## # A tibble: 2 x 5
##   Y      mean    sd n_obs  prop
##   <chr> <dbl> <dbl> <int> <dbl>
## 1 0     0.828  1.03    75  0.75
## 2 1     3.43   1.78    25  0.25
```

- The Normal density function for data with mean $\mu$ and standard deviation $\sigma$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

# Fit Model

We calculate estimates for the mean and standard deviation of $X$, within each level of $Y$, along with the proportion of observations within each level of $Y$:

```
sim_data %>% group_by(Y) %>% summarize(mean = mean(X), sd = sd(X), n_obs = n()) %>%
  mutate(prop = n_obs/sum(n_obs))
```

```
## # A tibble: 2 x 5
##   Y      mean    sd n_obs  prop
##   <chr> <dbl> <dbl> <int> <dbl>
## 1 0     0.828  1.03    75  0.75
## 2 1     3.43   1.78    25  0.25
```

- The Normal density function for data with mean $\mu$ and standard deviation $\sigma$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- We can use this density formula, along with our estimates of $\mu$, $\sigma$ and $P(Y = A_j)$, to calculate

$$P(X|Y = A_j) \cdot P(Y = A_j)$$

# Fit Model

We calculate estimates for the mean and standard deviation of $X$, within each level of $Y$, along with the proportion of observations within each level of $Y$:

```
sim_data %>% group_by(Y) %>% summarize(mean = mean(X), sd = sd(X), n_obs = n()) %>%
  mutate(prop = n_obs/sum(n_obs))
```

```
## # A tibble: 2 x 5
##   Y      mean    sd n_obs  prop
##   <chr> <dbl> <dbl> <int> <dbl>
## 1 0     0.828  1.03    75  0.75
## 2 1     3.43   1.78    25  0.25
```

- The Normal density function for data with mean $\mu$ and standard deviation $\sigma$ is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

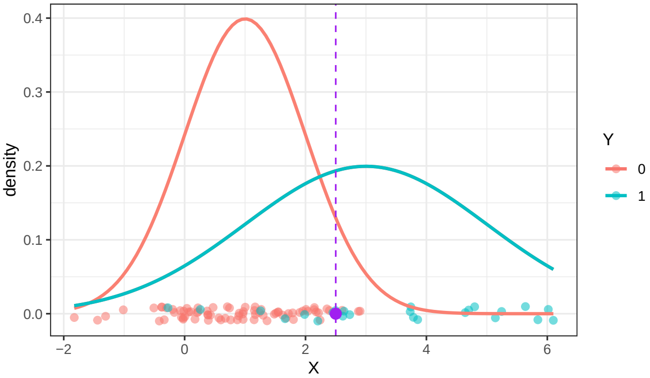
- We can use this density formula, along with our estimates of $\mu$, $\sigma$ and $P(Y = A_j)$, to calculate

$$P(X|Y = A_j) \cdot P(Y = A_j)$$

- And from this, using Bayes Rule, we can calculate $P(Y = A_j|X)$.

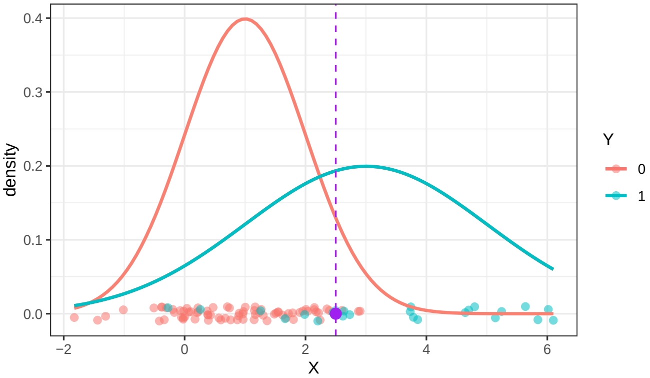
# Prediction

- Suppose we wish to classify a test point with $X = 2.5$



# Prediction

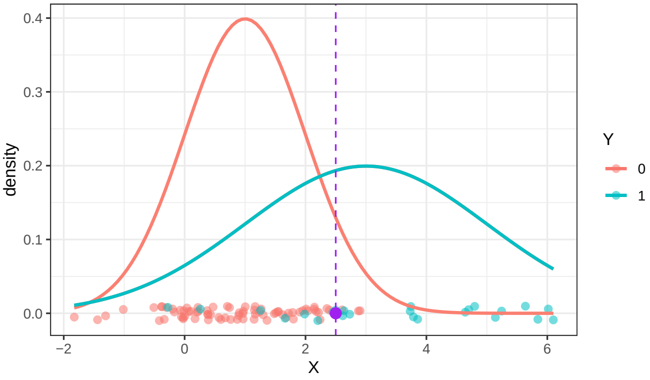
- Suppose we wish to classify a test point with $X = 2.5$




- On the one hand, $X = 2.5$ is more likely when $Y = 1$ than when $Y = 0$.

# Prediction

- Suppose we wish to classify a test point with $X = 2.5$




- On the one hand, $X = 2.5$ is more likely when $Y = 1$ than when $Y = 0$.
- But on the other hand, in general, $Y = 1$ occurs much more frequently than \$Y = 0 \$.

# Estimating Density

As $X$ is a continuous variable, we can't compute $P(X = 2.5)$. But we can compute the density functions at $X = 2.5$, which is the rate of generating data near $x = 2.5$.

## Estimating Density

As $X$ is a continuous variable, we can't compute $P(X = 2.5)$. But we can compute the density functions at $X = 2.5$, which is the rate of generating data near $x = 2.5$.

- If $Y = 1$, then $\mu_1 = 3.43$ and $\sigma_1 = 1.78$ and so

$$f_1(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.78^2}} \exp\left(-\frac{(2.5 - 3.43)^2}{2 \cdot 1.78^2}\right) = 0.196$$

# Estimating Density

As $X$ is a continuous variable, we can't compute $P(X = 2.5)$. But we can compute the density functions at $X = 2.5$, which is the rate of generating data near $x = 2.5$.

- If $Y = 1$, then $\mu_1 = 3.43$ and $\sigma_1 = 1.78$ and so

$$f_1(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.78^2}} \exp\left(-\frac{(2.5 - 3.43)^2}{2 \cdot 1.78^2}\right) = 0.196$$

- If $Y = 0$, then $\mu_0 = 0.83$ and $\sigma_0 = 1.03$ and so

$$f_0(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.03^2}} \exp\left(-\frac{(2.5 - 0.83)^2}{2 \cdot 1.03^2}\right) = 0.104$$

# Estimating Density

As $X$ is a continuous variable, we can't compute $P(X = 2.5)$. But we can compute the density functions at $X = 2.5$, which is the rate of generating data near $x = 2.5$.

- If $Y = 1$, then $\mu_1 = 3.43$ and $\sigma_1 = 1.78$ and so

$$f_1(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.78^2}} \exp\left(-\frac{(2.5 - 3.43)^2}{2 \cdot 1.78^2}\right) = 0.196$$

- If $Y = 0$, then $\mu_0 = 0.83$ and $\sigma_0 = 1.03$ and so

$$f_0(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.03^2}} \exp\left(-\frac{(2.5 - 0.83)^2}{2 \cdot 1.03^2}\right) = 0.104$$

- We are more likely to see data near $X = 2.5$ when $Y = 1$ than when $Y = 0$. However, we also need to take into account the overall chance that $Y = 1$:

# Estimating Density

As $X$ is a continuous variable, we can't compute $P(X = 2.5)$. But we can compute the density functions at $X = 2.5$, which is the rate of generating data near $x = 2.5$.

- If $Y = 1$, then $\mu_1 = 3.43$ and $\sigma_1 = 1.78$ and so

$$f_1(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.78^2}} \exp\left(-\frac{(2.5 - 3.43)^2}{2 \cdot 1.78^2}\right) = 0.196$$

- If $Y = 0$, then $\mu_0 = 0.83$ and $\sigma_0 = 1.03$ and so

$$f_0(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.03^2}} \exp\left(-\frac{(2.5 - 0.83)^2}{2 \cdot 1.03^2}\right) = 0.104$$

- We are more likely to see data near $X = 2.5$ when $Y = 1$ than when $Y = 0$. However, we also need to take into account the overall chance that $Y = 1$:

$$f_1(2.5) \cdot P(Y = 1) = 0.196 \cdot 0.25 = 0.049 \qquad f_0(2.5) \cdot P(Y = 0) = 0.104 \cdot 0.75 = 0.078$$

# Estimating Density

As $X$ is a continuous variable, we can't compute $P(X = 2.5)$. But we can compute the density functions at $X = 2.5$, which is the rate of generating data near $x = 2.5$.

- If $Y = 1$, then $\mu_1 = 3.43$ and $\sigma_1 = 1.78$ and so

$$f_1(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.78^2}} \exp\left(-\frac{(2.5 - 3.43)^2}{2 \cdot 1.78^2}\right) = 0.196$$

- If $Y = 0$, then $\mu_0 = 0.83$ and $\sigma_0 = 1.03$ and so

$$f_0(2.5) = \frac{1}{\sqrt{2\pi \cdot 1.03^2}} \exp\left(-\frac{(2.5 - 0.83)^2}{2 \cdot 1.03^2}\right) = 0.104$$

- We are more likely to see data near $X = 2.5$ when $Y = 1$ than when $Y = 0$. However, we also need to take into account the overall chance that $Y = 1$:

$$f_1(2.5) \cdot P(Y = 1) = 0.196 \cdot 0.25 = 0.049 \qquad f_0(2.5) \cdot P(Y = 0) = 0.104 \cdot 0.75 = 0.078$$

- Therefore, $P(Y = 0|X = 2.5) > P(Y = 1|X = 2.5)$ since

$$\frac{f_0(2.5)P(Y = 0)}{f_0(2.5)P(Y = 0) + f_1(2.5)P(Y = 1)} > \frac{f_1(2.5) \cdot P(Y = 0)}{f_0(2.5)P(Y = 0) + f_1(2.5)P(Y = 1)}$$

# Extending to Multiple Predictors

The previous method works great with a single predictor. However, we run into difficulties if we want to use multiple predictors.

# Extending to Multiple Predictors

The previous method works great with a single predictor. However, we run into difficulties if we want to use multiple predictors.

- Estimating $P(X_1, X_2, \ldots, X_p | Y = A_j)$ can require immense amounts of data:
  - We need to estimate not only the individual distributions of each $X_i$, but also estimate all of the $\approx 2^p$ relationships between the $X$'s

# Extending to Multiple Predictors

The previous method works great with a single predictor. However, we run into difficulties if we want to use multiple predictors.

- Estimating $P(X_1, X_2, \ldots, X_p | Y = A_j)$ can require immense amounts of data:
  - We need to estimate not only the individual distributions of each $X_i$, but also estimate all of the $\approx 2^p$ relationships between the $X$'s
- There are a few methods for overcoming this challenge:
  - **Discriminant Analysis** (LDA / QDA) assumes that the only noteworthy relationship between predictors is correlation. This reduces the problem to estimating $\approx p^2$ relationships

# Extending to Multiple Predictors

The previous method works great with a single predictor. However, we run into difficulties if we want to use multiple predictors.

- Estimating $P(X_1, X_2, \ldots, X_p | Y = A_j)$ can require immense amounts of data:
  - We need to estimate not only the individual distributions of each $X_i$, but also estimate all of the $\approx 2^p$ relationships between the $X$'s
- There are a few methods for overcoming this challenge:
  - **Discriminant Analysis** (LDA / QDA) assumes that the only noteworthy relationship between predictors is correlation. This reduces the problem to estimating $\approx p^2$ relationships
- **Naive Bayes** assumes that there are *no* noteworthy relationships among predictors. We only need to estimate individual distributions for each predictor.

# Extending to Multiple Predictors

The previous method works great with a single predictor. However, we run into difficulties if we want to use multiple predictors.

- Estimating $P(X_1, X_2, \ldots, X_p | Y = A_j)$ can require immense amounts of data:
  - We need to estimate not only the individual distributions of each $X_i$, but also estimate all of the $\approx 2^p$ relationships between the $X$'s
- There are a few methods for overcoming this challenge:
  - **Discriminant Analysis** (LDA / QDA) assumes that the only noteworthy relationship between predictors is correlation. This reduces the problem to estimating $\approx p^2$ relationships
- **Naive Bayes** assumes that there are *no* noteworthy relationships among predictors. We only need to estimate individual distributions for each predictor.
- We investigate only the latter. It turns out that the former produces models that are *very* comparable to logistic regression.

Section 3

# Naive Bayes

# Naive Bayes?

**Goal**: Estimate $P(Y = A_j | X_1, X_2, \ldots, X_p)$.

**Method**: estimate $P(X_1, \ldots, X_p | Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

# Naive Bayes?

**Goal**: Estimate $P(Y = A_j | X_1, X_2, \ldots, X_p)$.

**Method**: estimate $P(X_1, \ldots, X_p | Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

- The Naive Bayes model assumes that $X_1, \ldots, X_p$ are **independent**, and so by the multiplication rule:

$$P(X_1, \ldots, X_p | Y = A_j) = P(X_1 | Y = A_j) \cdot P(X_2 | Y = A_j) \cdots P(X_p | Y = A_j)$$

# Naive Bayes?

**Goal**: Estimate $P(Y = A_j | X_1, X_2, \ldots, X_p)$.

**Method**: estimate $P(X_1, \ldots, X_p | Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

- The Naive Bayes model assumes that $X_1, \ldots, X_p$ are **independent**, and so by the multiplication rule:

$$P(X_1, \ldots, X_p | Y = A_j) = P(X_1 | Y = A_j) \cdot P(X_2 | Y = A_j) \cdots P(X_p | Y = A_j)$$

- Each term $P(X_i | Y = A_j)$ can be estimated individually:

# Naive Bayes?

**Goal**: Estimate $P(Y = A_j | X_1, X_2, \ldots, X_p)$.

**Method**: estimate $P(X_1, \ldots, X_p | Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

- The Naive Bayes model assumes that $X_1, \ldots, X_p$ are **independent**, and so by the multiplication rule:

$$P(X_1, \ldots, X_p | Y = A_j) = P(X_1 | Y = A_j) \cdot P(X_2 | Y = A_j) \cdots P(X_p | Y = A_j)$$

- Each term $P(X_i | Y = A_j)$ can be estimated individually:
  - If $X_i$ is continuous, we estimate $P(X_i | A_j)$ using a normal distribution model (as before)

# Naive Bayes?

**Goal**: Estimate $P(Y = A_j | X_1, X_2, \ldots, X_p)$.

**Method**: estimate $P(X_1, \ldots, X_p | Y = A_j)$ for all levels of $A_j$, and combine them using Bayes Rule and Law of Total Probability:

- The Naive Bayes model assumes that $X_1, \ldots, X_p$ are **independent**, and so by the multiplication rule:

$$P(X_1, \ldots, X_p | Y = A_j) = P(X_1 | Y = A_j) \cdot P(X_2 | Y = A_j) \cdots P(X_p | Y = A_j)$$

- Each term $P(X_i | Y = A_j)$ can be estimated individually:
  - If $X_i$ is continuous, we estimate $P(X_i | A_j)$ using a normal distribution model (as before)
  - If $X_i$ categorical, we estimate $P(X_i | A_j)$ by computing the proportion of observations in each level of $X_i$, among all observations with $Y = A_j$.

## Why Naive Bayes

Why might we make such an unreasonable (Naive?) assumption about independence?

# Why Naive Bayes

Why might we make such an unreasonable (Naive?) assumption about independence?

- All models are wrong. But some are useful.

# Why Naive Bayes

Why might we make such an unreasonable (Naive?) assumption about independence?

- All models are wrong. But some are useful.
- When we have many variables but few observations per variable, we often do not have luxury of estimating a large number of relationships.
  - We need simplifying assumptions (high bias, low variance)

# Why Naive Bayes

Why might we make such an unreasonable (Naive?) assumption about independence?

- All models are wrong. But some are useful.
- When we have many variables but few observations per variable, we often do not have luxury of estimating a large number of relationships.
  - We need simplifying assumptions (high bias, low variance)
- Naive Bayes can provide non-linear decision boundaries (trading one flexibility for another)

# Why Naive Bayes

Why might we make such an unreasonable (Naive?) assumption about independence?

- All models are wrong. But some are useful.
- When we have many variables but few observations per variable, we often do not have luxury of estimating a large number of relationships.
  - We need simplifying assumptions (high bias, low variance)
- Naive Bayes can provide non-linear decision boundaries (trading one flexibility for another)
- For model accuracy, the goal is correctly predicting the class of $Y$, not necessarily estimating the probability that $Y$ is in that class:
  - Naive Bayes tends to produce woefully incorrect estimates of $P(Y = A_j|X)$.
  - But usually concurs with the **prediction** that would be made by the true probability model

# Why Naive Bayes

Why might we make such an unreasonable (Naive?) assumption about independence?

- All models are wrong. But some are useful.
- When we have many variables but few observations per variable, we often do not have luxury of estimating a large number of relationships.
  - We need simplifying assumptions (high bias, low variance)
- Naive Bayes can provide non-linear decision boundaries (trading one flexibility for another)
- For model accuracy, the goal is correctly predicting the class of $Y$, not necessarily estimating the probability that $Y$ is in that class:
  - Naive Bayes tends to produce woefully incorrect estimates of $P(Y = A_j|X)$.
  - But usually concurs with the **prediction** that would be made by the true probability model
- Sometimes dependence among variables can "cancel out" in aggregate. I.e. error in estimating $P(X_1|X_2)$ can be cancelled by error in estimating $P(X_2|X_3)$ and $P(X_1|X_3)$.

# Naive Bayes in R

- We fit a Naive Bayes model using the `naiveBayes` function in the `e1071` package:

## Naive Bayes in R

- We fit a Naive Bayes model using the `naiveBayes` function in the `e1071` package:

```
library(e1071)
nb_mod <- naiveBayes(Y ~ X1 + X2, data = training_data)
```

# Naive Bayes in R

- We fit a Naive Bayes model using the `naiveBayes` function in the `e1071` package:

```
library(e1071)
nb_mod <- naiveBayes(Y ~ X1 + X2, data = training_data)
```

- We make predictions for class using `predict`

```
my_preds <- predict(nb_mod, data = test_data)
```

# Naive Bayes in R

- We fit a Naive Bayes model using the `naiveBayes` function in the `e1071` package:

```
library(e1071)
nb_mod <- naiveBayes(Y ~ X1 + X2, data = training_data)
```

- We make predictions for class using `predict`

```
my_preds <- predict(nb_mod, data = test_data)
```

- And we can obtain the naive bayes estimates for probabilities using:

```
my_probs<- predict(nb_mod, data = test_data, type = "raw")
```

# Titanic Again

How does Naive Bayes do on the Titanic data set explored previously?

# Titanic Again

How does Naive Bayes do on the Titanic data set explored previously?

- We look at some of the variables:

```
library(dplyr)
glimpse(Titanic)
```

```
## Rows: 1,313
## Columns: 10
## $ pclass    <chr> "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st~
## $ survived  <fct> 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, ~
## $ name      <chr> "Allen, Miss Elisabeth Walton", "Allison, Miss Helen Loraine~
## $ age       <dbl> 29.0000, 2.0000, 30.0000, 25.0000, 0.9167, 47.0000, 63.0000,~
## $ embarked  <chr> "Southampton", "Southampton", "Southampton", "Southampton", ~
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "Montreal~
## $ room      <chr> "B-5", "C26", "C26", "C26", "C22", "E-12", "D-7", "A-36", "C~
## $ ticket    <chr> "24160 L221", NA, NA, NA, NA, NA, "13502 L77", NA, NA, NA, "~
## $ boat      <chr> "2", NA, "(135)", NA, "11", "3", "10", NA, "2", "(22)", "(12~
## $ sex       <chr> "female", "female", "male", "female", "male", "male", "femal~
```

# Titanic Again

How does Naive Bayes do on the Titanic data set explored previously?

- We look at some of the variables:

```
library(dplyr)
glimpse(Titanic)
```

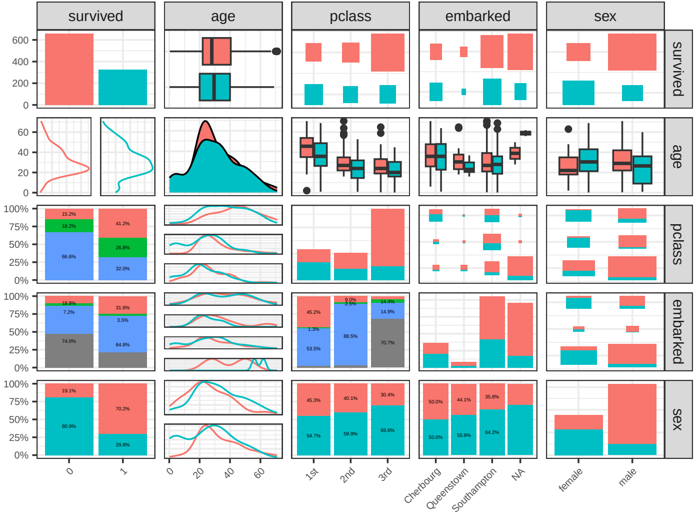
```
## Rows: 1,313
## Columns: 10
## $ pclass    <chr> "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st~
## $ survived  <fct> 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, ~
## $ name      <chr> "Allen, Miss Elisabeth Walton", "Allison, Miss Helen Loraine~
## $ age       <dbl> 29.0000, 2.0000, 30.0000, 25.0000, 0.9167, 47.0000, 63.0000,~
## $ embarked  <chr> "Southampton", "Southampton", "Southampton", "Southampton", ~
## $ home.dest <chr> "St Louis, MO", "Montreal, PQ / Chesterville, ON", "Montreal~
## $ room      <chr> "B-5", "C26", "C26", "C26", "C22", "E-12", "D-7", "A-36", "C~
## $ ticket    <chr> "24160 L221", NA, NA, NA, NA, NA, "13502 L77", NA, NA, NA, "~
## $ boat      <chr> "2", NA, "(135)", NA, "11", "3", "10", NA, "2", "(22)", "(12~
## $ sex       <chr> "female", "female", "male", "female", "male", "male", "femal~
```

- And break our data into test/training sets:

```
library(rsample)
set.seed(10)
Titanic_split <- initial_split(Titanic)
Titanic_train <- training(Titanic_split)
Titanic_test <- testing(Titanic_split)
```

# Data Visualization

```r
library(GGally)
Titanic_train %>% select(survived, age, pclass, embarked, sex) %>% ggpairs(aes(color = survived))
```



# Exploratory Analysis

- What trends are apparent among variables?
- Does it seem like predictors are independent, given values of the response?

# Fitting the Naive Bayes Model

- We first fit the model using age, pcclass, embarked and sex

```
nb_fit <- naiveBayes(survived ~ age + pclass + embarked + sex, data = Titanic_train)
nb_fit$tables
```

```
##    age
## Y       [,1]     [,2]
##   0 31.73908 14.29293
##   1 30.15109 15.62311
##    pclass
## Y         1st       2nd       3rd
##   0 0.1517451 0.1820941 0.6661608
##   1 0.4123077 0.2676923 0.3200000
```

```
##    embarked
## Y    Cherbourg Queenstown Southampton
##   0 0.18786127 0.07225434  0.73988439
##   1 0.31640625 0.03515625  0.64843750
##    sex
## Y       female      male
##   0 0.1911988 0.8088012
##   1 0.7015385 0.2984615
```

# Fitting the Naive Bayes Model

- We first fit the model using age, pcclass, embarked and sex

```
nb_fit <- naiveBayes(survived ~ age + pclass + embarked + sex, data = Titanic_train)
nb_fit$tables
```

```
##    age
## Y       [,1]     [,2]
##   0 31.73908 14.29293
##   1 30.15109 15.62311
##    pclass
## Y         1st       2nd       3rd
##   0 0.1517451 0.1820941 0.6661608
##   1 0.4123077 0.2676923 0.3200000
```

```
##    embarked
## Y    Cherbourg Queenstown Southampton
##   0 0.18786127 0.07225434  0.73988439
##   1 0.31640625 0.03515625  0.64843750
##    sex
## Y       female      male
##   0 0.1911988 0.8088012
##   1 0.7015385 0.2984615
```

- For quantitative variables, the first column is the predictor mean and the second is the predictor standard deviation, within each response class.

# Fitting the Naive Bayes Model

- We first fit the model using age, pcclass, embarked and sex

```
nb_fit <- naiveBayes(survived ~ age + pclass + embarked + sex, data = Titanic_train)
nb_fit$tables
```

```
##    age
## Y        [,1]     [,2]
##   0 31.73908 14.29293
##   1 30.15109 15.62311
##    pclass
## Y         1st       2nd       3rd
##   0 0.1517451 0.1820941 0.6661608
##   1 0.4123077 0.2676923 0.3200000
```

```
##    embarked
## Y    Cherbourg Queenstown Southampton
##   0 0.18786127 0.07225434  0.73988439
##   1 0.31640625 0.03515625  0.64843750
##    sex
## Y      female      male
##   0 0.1911988 0.8088012
##   1 0.7015385 0.2984615
```

- For quantitative variables, the first column is the predictor mean and the second is the predictor standard deviation, within each response class.
- For categorical variables, the columns correspond to the proportions of that variable within each response class.

# Predicting with Naive Bayes

Now, we make class predictions

# Predicting with Naive Bayes

Now, we make class predictions

```
my_preds <- predict(nb_fit, Titanic_test)
head(my_preds)
```

```
## [1] 0 0 0 0 0 1
## Levels: 0 1
```

# Predicting with Naive Bayes

Now, we make class predictions

```
my_preds <- predict(nb_fit, Titanic_test)
head(my_preds)
```

```
## [1] 0 0 0 0 0 1
## Levels: 0 1
```

```
my_probs <- predict(nb_fit, Titanic_test, type = "raw")
head(my_probs)
```

```
##                0         1
## [1,] 0.7184279 0.2815721
## [2,] 0.6976581 0.3023419
## [3,] 0.7110352 0.2889648
## [4,] 0.5752423 0.4247577
## [5,] 0.6976581 0.3023419
## [6,] 0.1192007 0.8807993
```

# Predicting with Naive Bayes

Now, we make class predictions

```
my_preds <- predict(nb_fit, Titanic_test)
head(my_preds)
```

```
## [1] 0 0 0 0 0 1
## Levels: 0 1
```

```
my_probs <- predict(nb_fit, Titanic_test, type = "raw")
head(my_probs)
```

```
##              0         1
## [1,] 0.7184279 0.2815721
## [2,] 0.6976581 0.3023419
## [3,] 0.7110352 0.2889648
## [4,] 0.5752423 0.4247577
## [5,] 0.6976581 0.3023419
## [6,] 0.1192007 0.8807993
```

And create a results data frame

```
nb_results <- data.frame(obs = Titanic_test$survived, preds = my_preds, probs = my_probs)
```

# Model Assessment

Compute accuracy, sensitivity and specificity:

```
library(yardstick)
my_metrics <- metric_set(accuracy, sensitivity, specificity)
my_metrics(nb_results, truth = obs, estimate = preds)
```

```
## # A tibble: 3 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 accuracy    binary         0.799
## 2 sensitivity binary         0.980
## 3 specificity binary         0.5
```

# Model Assessment

Compute accuracy, sensitivity and specificity:

```
library(yardstick)
my_metrics <- metric_set(accuracy, sensitivity, specificity)
my_metrics(nb_results, truth = obs, estimate = preds)
```

```
## # A tibble: 3 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 accuracy    binary         0.799
## 2 sensitivity binary         0.980
## 3 specificity binary         0.5
```

- Overall, the model was moderately accurate

# Model Assessment

Compute accuracy, sensitivity and specificity:

```
library(yardstick)
my_metrics <- metric_set(accuracy, sensitivity, specificity)
my_metrics(nb_results, truth = obs, estimate = preds)
```

```
## # A tibble: 3 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 accuracy    binary         0.799
## 2 sensitivity binary         0.980
## 3 specificity binary         0.5
```

- Overall, the model was moderately accurate
  - The model was very good at correctly identifying true survivors (high sensitivity)

# Model Assessment

Compute accuracy, sensitivity and specificity:

```
library(yardstick)
my_metrics <- metric_set(accuracy, sensitivity, specificity)
my_metrics(nb_results, truth = obs, estimate = preds)
```

```
## # A tibble: 3 x 3
##   .metric     .estimator .estimate
##   <chr>       <chr>          <dbl>
## 1 accuracy    binary         0.799
## 2 sensitivity binary         0.980
## 3 specificity binary         0.5
```

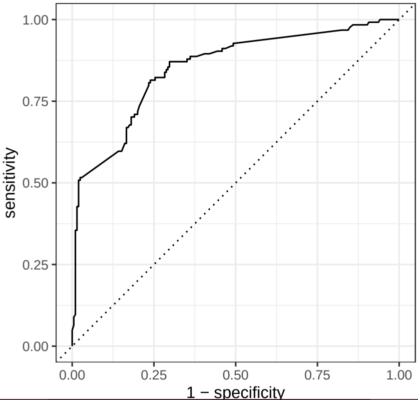
- Overall, the model was moderately accurate
  - The model was very good at correctly identifying true survivors (high sensitivity)
  - But was not as good at correctly identifying true non-survivors (mediocre specificity)

# ROC and AUC

```
roc_auc(nb_results, truth = obs, probs.1, event_level = "second")
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.850
```

```
autoplot( roc_curve(nb_results, truth = obs, probs.1, event_level = "second") )
```



## Comparison

How does Naive Bayes compare to logistic regression?

# Comparison

How does Naive Bayes compare to logistic regression?

```
my_glm <- glm(survived ~ age + pclass + embarked + sex, data = Titanic_train, family = "binomial
glm_probs <- predict(my_glm, newdata = Titanic_test, type = "response")
glm_preds <- as.factor( ifelse(glm_probs > 0.5, 1, 0))
glm_results <- data.frame(obs = Titanic_test$survived, preds = glm_preds, probs = glm_probs)
```

```
## # A tibble: 8 x 4
##   .metric     .estimator .estimate model
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.813 logistic
## 2 sensitivity binary         0.929 logistic
## 3 specificity binary         0.691 logistic
## 4 roc_auc     binary         0.897 logistic
## 5 accuracy    binary         0.799 Naive Bayes
## 6 sensitivity binary         0.980 Naive Bayes
## 7 specificity binary         0.5   Naive Bayes
## 8 roc_auc     binary         0.850 Naive Bayes
```

# Comparison

How does Naive Bayes compare to logistic regression?

```
my_glm <- glm(survived ~ age + pclass + embarked + sex, data = Titanic_train, family = "binomial
glm_probs <- predict(my_glm, newdata = Titanic_test, type = "response")
glm_preds <- as.factor( ifelse(glm_probs > 0.5, 1, 0))
glm_results <- data.frame(obs = Titanic_test$survived, preds = glm_preds, probs = glm_probs)
```

```
## # A tibble: 8 x 4
##   .metric     .estimator .estimate model
##   <chr>       <chr>          <dbl> <chr>
## 1 accuracy    binary         0.813 logistic
## 2 sensitivity binary         0.929 logistic
## 3 specificity binary         0.691 logistic
## 4 roc_auc     binary         0.897 logistic
## 5 accuracy    binary         0.799 Naive Bayes
## 6 sensitivity binary         0.980 Naive Bayes
## 7 specificity binary         0.5   Naive Bayes
## 8 roc_auc     binary         0.850 Naive Bayes
```

- Logistic regression beats Naive Bayes (except on sensitivity)

# Comparative ROC Curves