

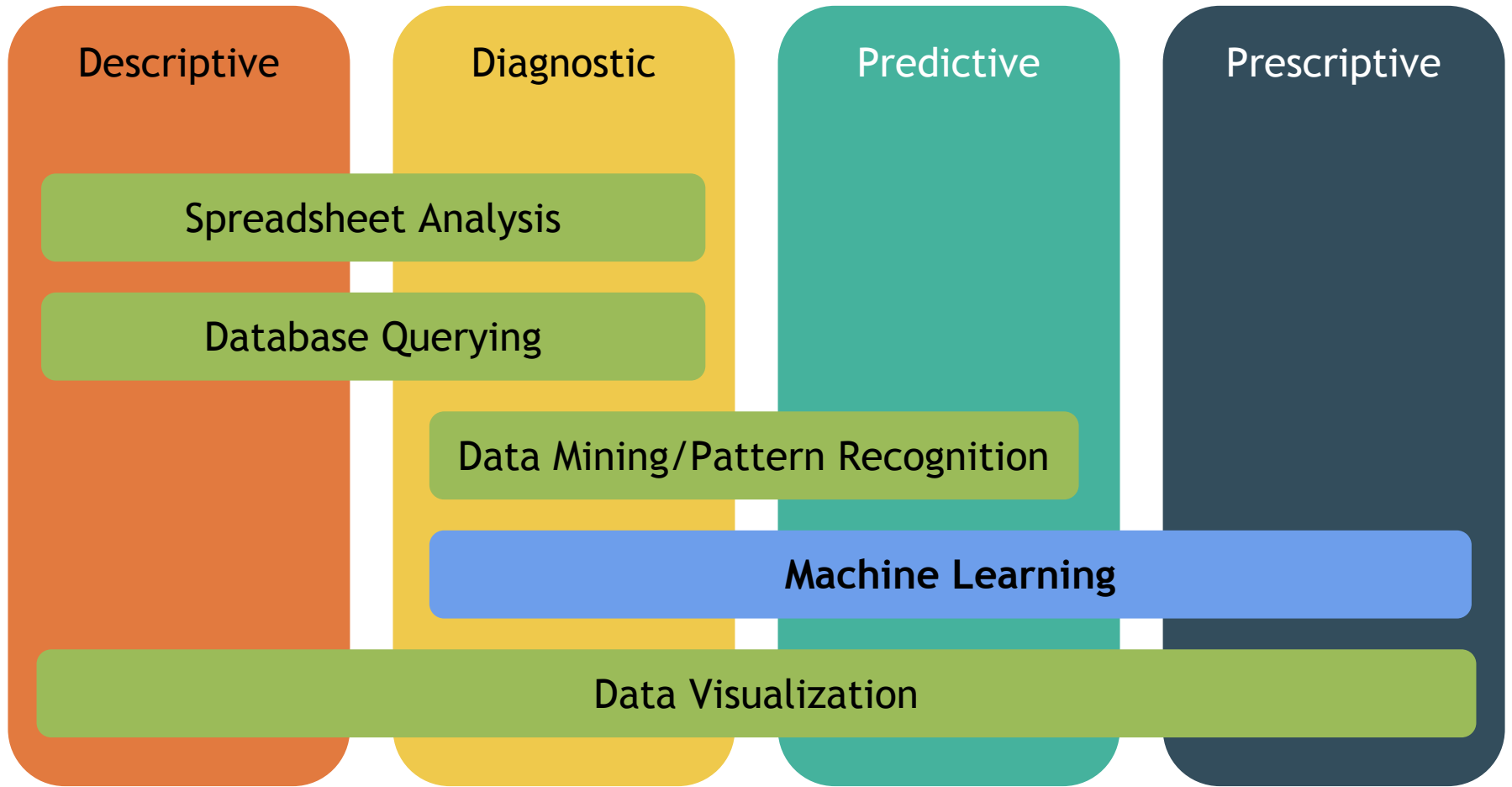
# CS102: Big Data

## Tools and Techniques, Discoveries and Pitfalls

Spring 2017  
Ethan Chan, Lisa Wang

*Lecture 11: Classification Part 2*

# Tools & Techniques



# Announcements

- Midterm

- Midterm Problem 3

- Answer: Al, Ben, Cob, Dan, Fay
    - If you gave that answer but were deducted points, see me (Ethan) after class for regrade

- Stats

- Mean: 36.03

- Solutions are posted on course website

- Assignment 4

- Part A released, Part B released next week
  - Both parts A and B are due Tue May 23

# Using ML to identify Skin Cancer

JANUARY 25, 2017

## Deep learning algorithm does as well as dermatologists in identifying skin cancer

*In hopes of creating better access to medical care, Stanford researchers have trained an algorithm to diagnose skin cancer.*



BY TAYLOR KUBOTA

It's scary enough making a doctor's appointment to see if a strange mole could be cancerous. Imagine, then, that you were in that situation while also living far away from the nearest doctor, unable to take time off work and unsure you had the money to cover the cost of the visit. In a scenario like this, an option to receive a diagnosis through your smartphone could be lifesaving.

Universal access to health care was on the minds of computer scientists at Stanford when they set out to create an artificially intelligent diagnosis algorithm for skin cancer. They made a database of nearly 130,000 skin disease images and trained their algorithm to visually diagnose potential cancer. From the very first test, it performed with inspiring accuracy.

"We realized it was feasible, not just to do something well, but as well as a human dermatologist," said



>130,000 images of skin lesions  
>2,000 different diseases

<http://news.stanford.edu/2017/01/25/artificial-intelligence-used-identify-skin-cancer/>

# Last Lecture

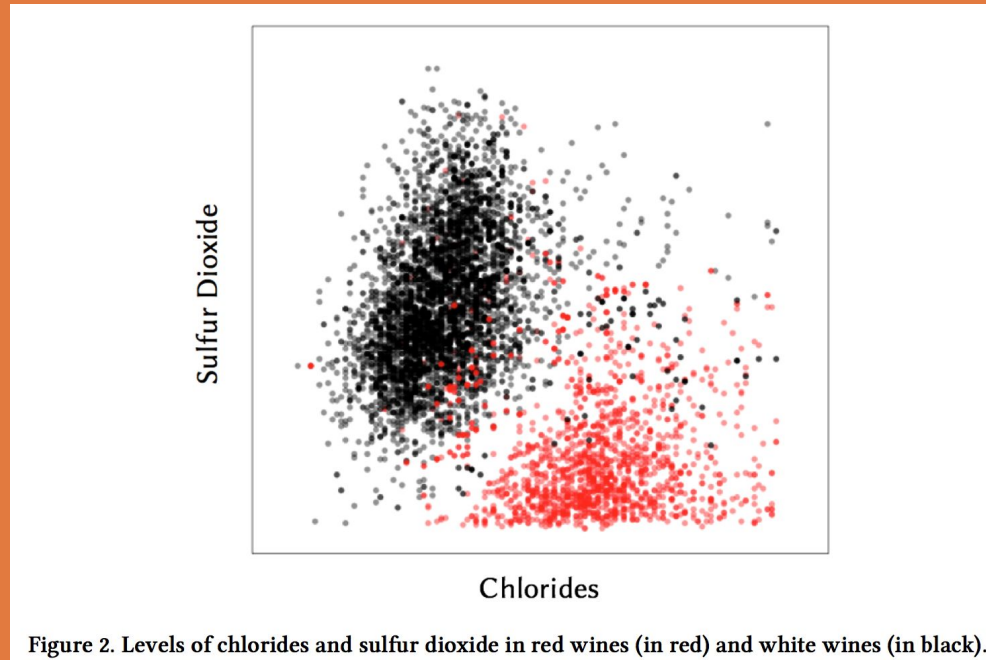
- Polynomial Regression
- Classification Part 1
  - K-Nearest Neighbors
  - Decision Trees / Random Forests
  - Logistic Regression

# Plan for Today

- Review Classification Part 1
  - More Examples
  - Apply classification algorithms from last lecture with python, pandas and sklearn
- Classification Part 2
  - Naive Bayes
  - Support Vector Machines

# *Review:* K-Nearest Neighbors

# KNN to Predict the Color of Wine

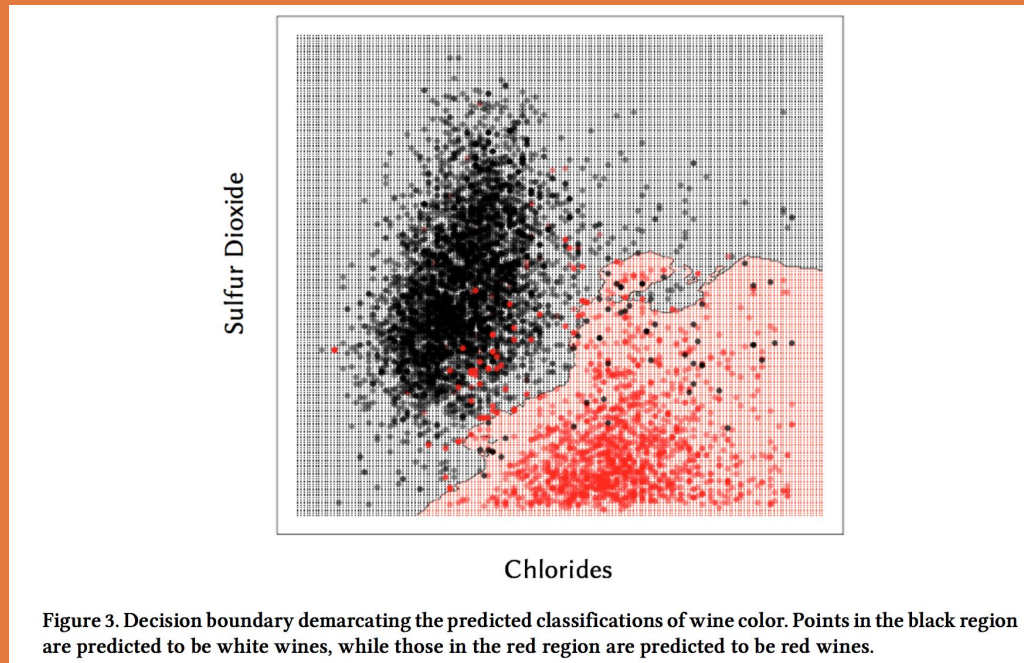


Source: “Numsense - Data Science for the Layman” by Ng & Soo.

- What are the input features?
- What are the target labels?
- How many classes?



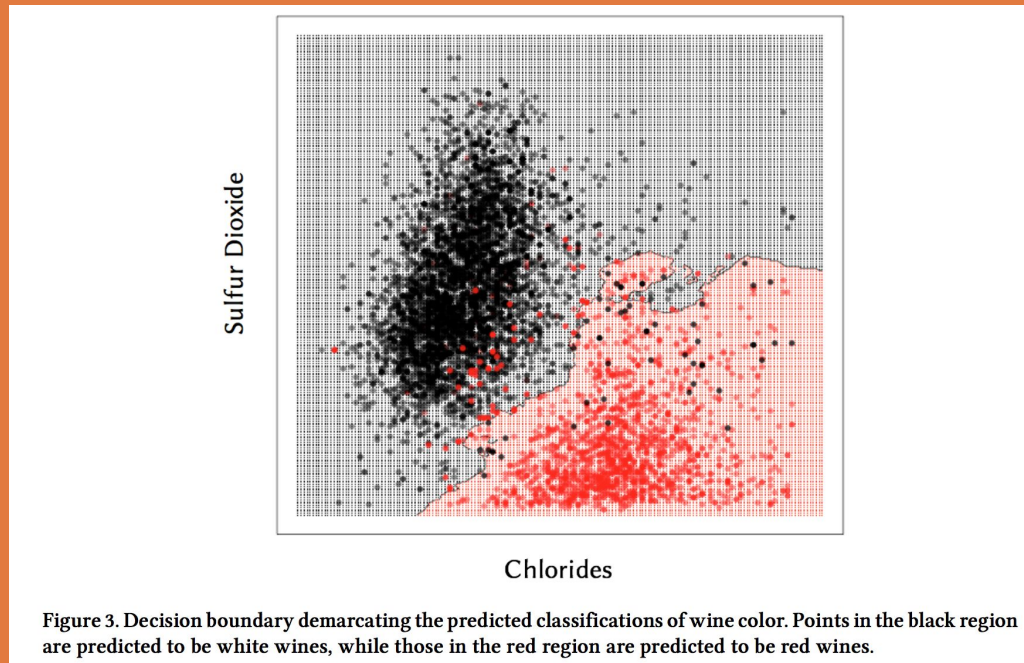
# KNN to Predict the Color of Wine



Source: “Numsense - Data Science for the Layman” by Ng & Soo.

- What does the decision boundary tell you about predictions?
- How can you compute the decision boundary for a fixed  $k$ ?

# KNN to Predict the Color of Wine



Source: “Numsense - Data Science for the Layman” by Ng & Soo.

- Explain “decision boundary” to your neighbor
- What does it tell you about predictions?

# KNN to Predict the Color of Wine

$k$  = number  
of neighbors

Test  
Accuracy

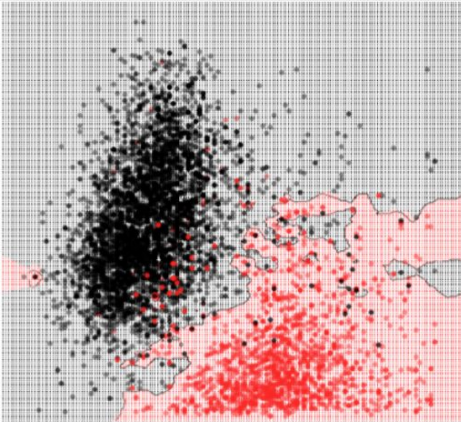
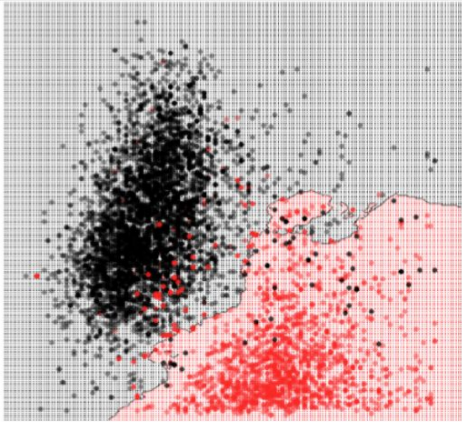
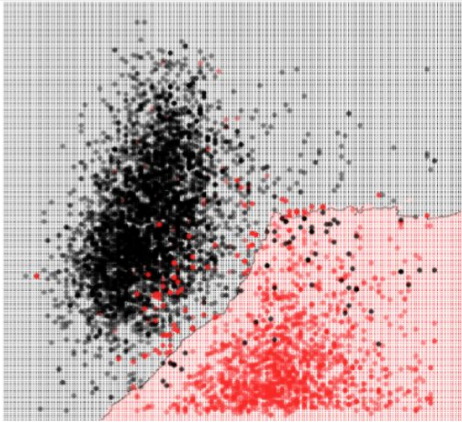
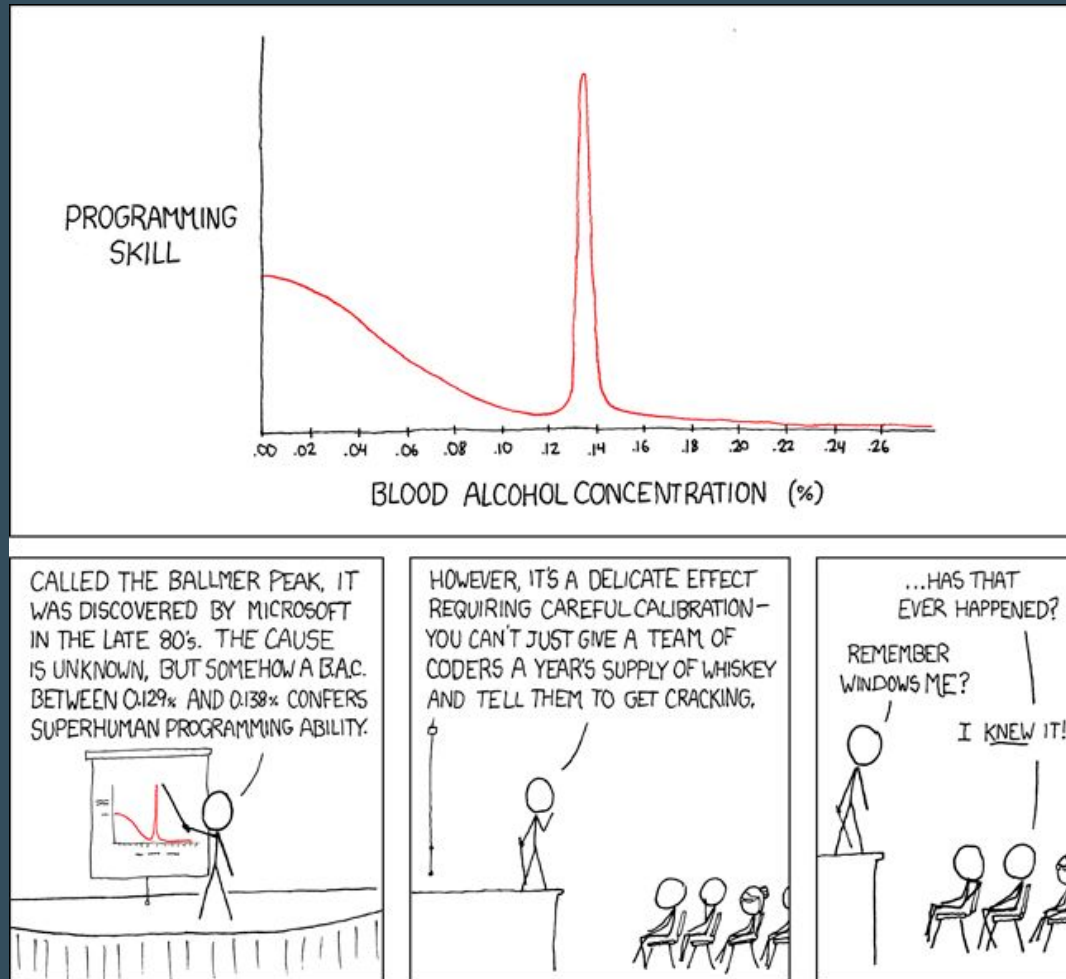
$k = 3$	$k = 17$	$k = 50$
		
98.2% accuracy	98.6% accuracy	97.8% accuracy
Overfit	Ideal fit	Underfit

Table 1. Comparison of model fit using varying values of  $k$ .

How does  $k$  influence the behavior  
underfitting / overfitting?

# Speaking of Wine...

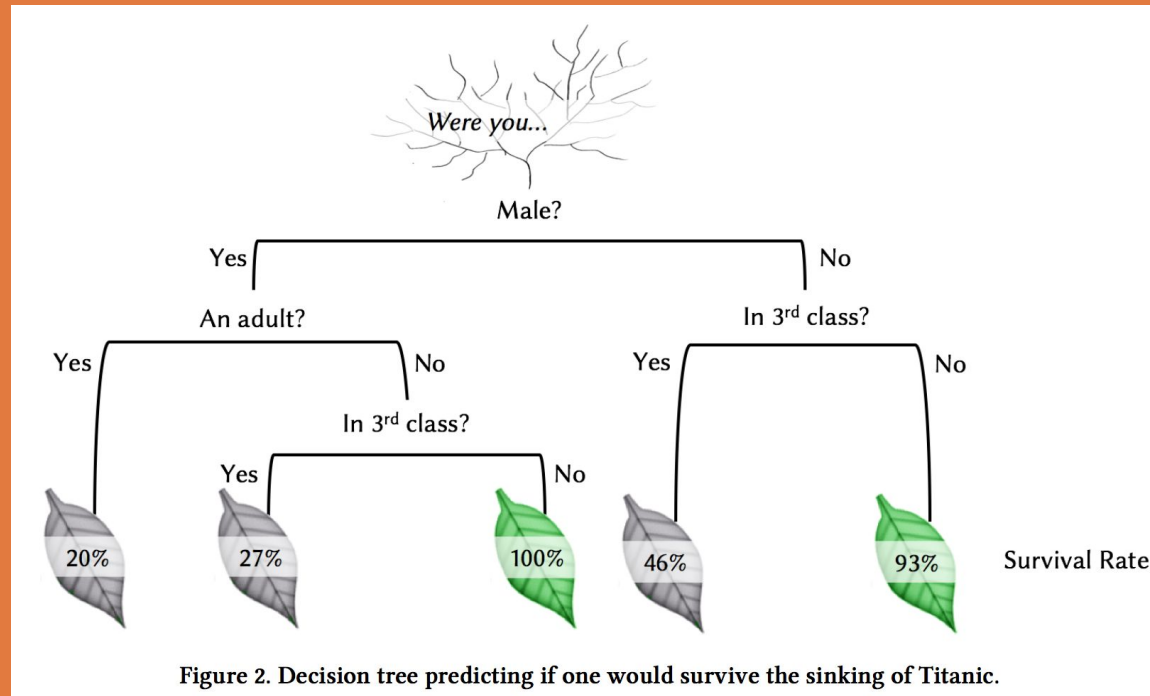


<https://xkcd.com/323/>

*Review:*

# Decision Trees/ Random Forests

# Decision Trees on the Titanic

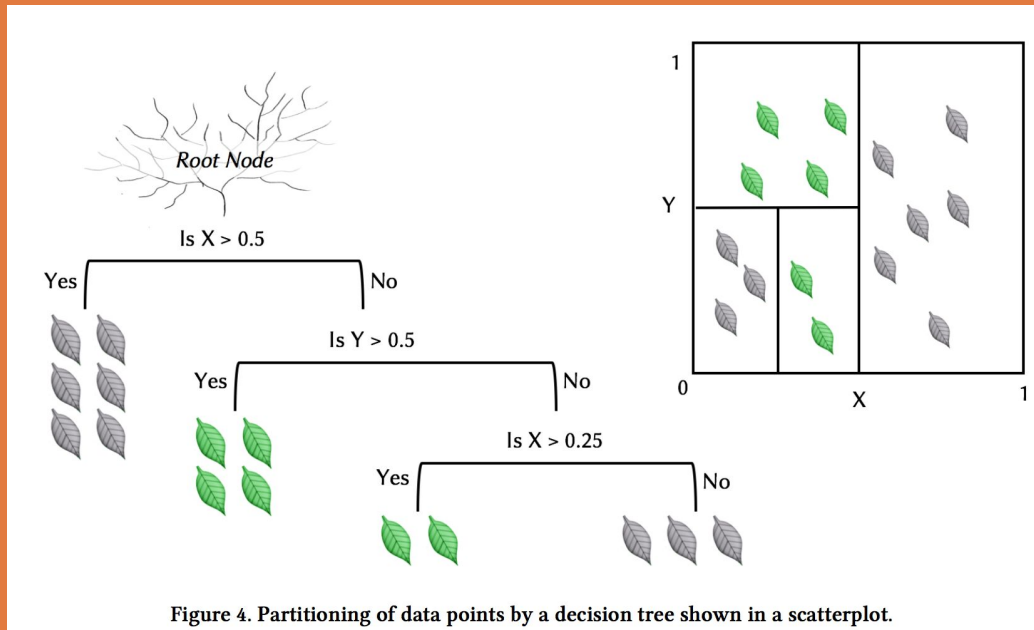


Source: "Numsense - Data Science for the Layman" by Ng & Soo.

- What are your chances of survival if you are:
  - A girl in 3rd class?
  - An adult male not in 3rd class?
- Who has the best survival rates?



# Decision Trees: Partitioning Data Points



Source: “Numsense - Data Science for the Layman” by Ng & Soo.

- Explain to your neighbor how the tree on the left maps to the partitioning on the right.
- Where are the decision boundaries?

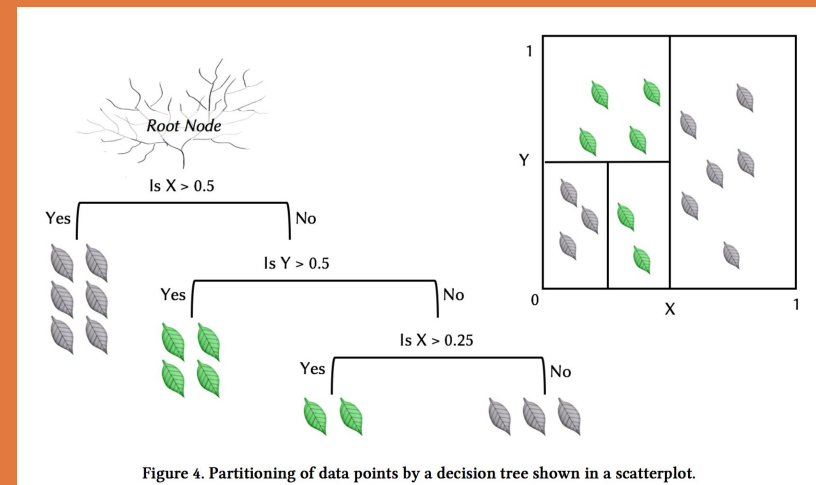
# Generating Decision Trees

Two Steps:

1. Identify a binary question based on one of the features that partitions the data into more homogeneous groups
2. Repeat Step 1 on each subtree until stopping criteria reached

Stopping Criteria (Examples):

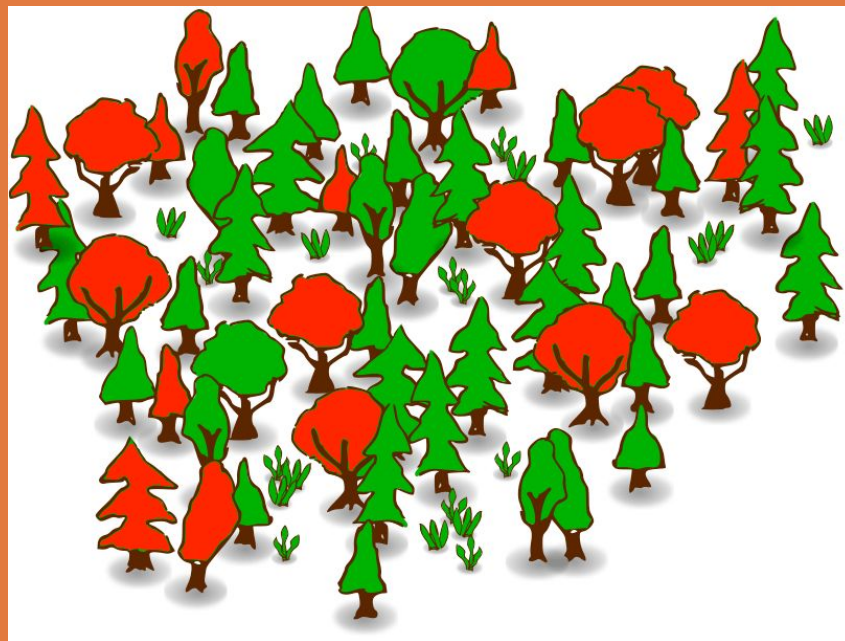
- All data points in leaf in same category
- Leaf contains fewer than 5 data points





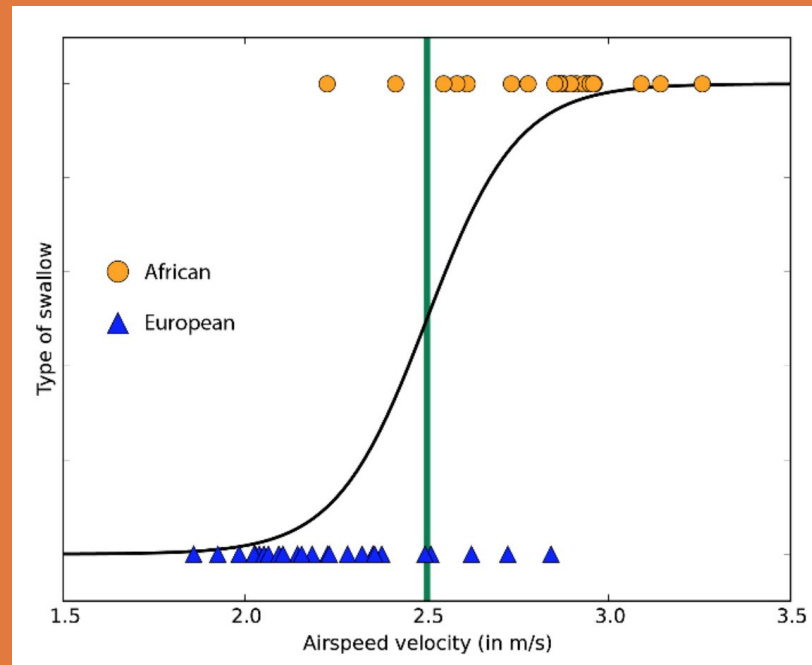
# Random Forests

- “Crowdsourcing”
- Ensemble of decision trees
- Take most popular vote as final prediction



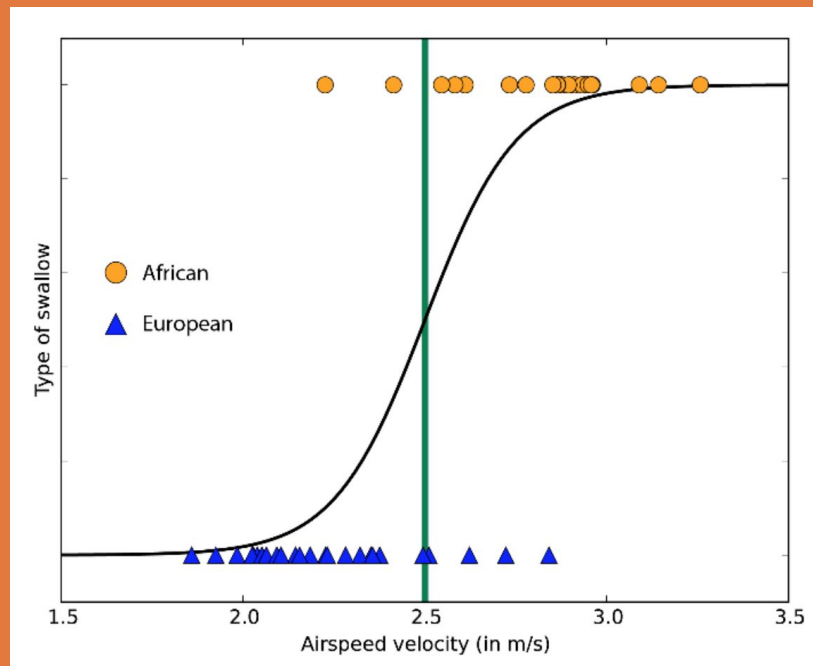
# *Review:* Logistic Regression

# Logistic Regression



- Fit logistic function to data points for each class
- Y - value corresponds to probability of that class

# Logistic Regression



- What type of swallow does the model predict if:
  - Velocity = 2.0
  - Velocity = 2.4
  - Velocity = 2.7?

# Let's code it up in Python!

Open

`lecture_10_classification_part_1_starter.ipynb`

# Announcements

- Midterm

- Midterm Problem 3

- Answer: Al, Ben, Cob, Dan, Fay
    - If you gave that answer but were deducted points, see me (Ethan) after class for regrade

- Stats

- Mean: 36.03

- Solutions are posted on course website

- Assignment 4

- Part A released, Part B released next week
  - Both parts A and B are due Tue May 23

# Up Next

- Classification Part 2
  - Naive Bayes
  - Support Vector Machines
- Evaluation metrics (if we have time)
  - Accuracy
  - False positives, false negatives
  - Precision
  - Recall
  - F1 Score

# Naive Bayes

- Define probability
- Define conditional probability
- Define Bayes Rule
- Define Conditional Independence
- Define Naive Bayes



# Thomas Bayes



# Probability

## Definition

Let event “Y” be if a person has (cancer or no cancer)

Let event “X” be the outcome of a test (positive or negative)

## Basic Probability

$P(Y = \text{cancer}) = 0.01$  means a person has 1% chance of having cancer

$P(Y = \text{no cancer}) = 0.99$  , 99% chance of no cancer

## Meaning

If you pick anyone on the street, there’s a 1% chance that person has cancer.

# Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

$P(Y = \text{cancer}) = 3/7$ ,  $P(Y = \text{no cancer}) = 4/7$

$P(X = \text{positive}) = 3/7$ ,  $P(X = \text{negative}) = 4/7$

# Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

$$P(Y = \text{cancer} \mid X = \text{positive})$$
$$= \frac{\text{\#cancer and positives}}{\text{\#positives}}$$

Probability of having cancer **given** that we know the test is positive.

# Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
D	Cancer	Positive
G	No Cancer	Positive

$P(Y = \text{cancer} \mid X = \text{positive})$

$= \frac{\text{\#cancer and positives}}{\text{\#positives}}$

$= \frac{\text{\#cancer and positives}}{3}$

We only want to look at rows where  $X = \text{positive}$  first

We can see that there are 3 rows that have positive tests

# Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
D	Cancer	Positive
G	No Cancer	Positive

$P(Y = \text{cancer} \mid X = \text{positive})$

$= \frac{\text{\#cancer and positives}}{\text{\#positives}}$

$= \frac{2}{3}$

$= \frac{2}{3}$

We can see that there are 2 rows where  $Y = \text{cancer}$  and  $X = \text{Positive}$ .

**We know know that  $P(Y = \text{cancer} \mid X = \text{positive}) = \frac{2}{3} = 0.66$**

# Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

What is  $P(Y = \text{Cancer} \mid X = \text{negative})$ ?

# Conditional Probability

Person	Y (cancer or not)	X (test positive or negative)
B	Cancer	Negative
C	No Cancer	Negative
E	No Cancer	Negative
F	No Cancer	Negative

What is  $P(Y = \text{Cancer} \mid X = \text{negative})$ ?  $\frac{1}{4}$ .



# Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

$$P(Y = \text{cancer}) = 3/7, P(Y = \text{no cancer}) = 4/7$$

$$P(X = \text{positive}) = 3/7, P(X = \text{negative}) = 4/7$$

$$P(Y = \text{cancer} \mid X = \text{positive}) = 2/3$$

$$P(Y = \text{cancer} \mid X = \text{negative}) = 1/4$$

# Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
C	No Cancer	Negative
D	Cancer	Positive
E	No Cancer	Negative
F	No Cancer	Negative
G	No Cancer	Positive

$$P(Y = \text{cancer}) = 3/7, P(Y = \text{no cancer}) = 4/7$$

How about  $P(X = \text{positive} \mid Y = \text{cancer})$ ?

$$P(X = \text{positive}) = 3/7, P(X = \text{negative}) = 4/7$$

$$P(Y = \text{cancer} \mid X = \text{positive}) = 2/3$$

$$P(Y = \text{cancer} \mid X = \text{negative}) = 1/4$$

# Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
D	Cancer	Positive

$P(Y = \text{cancer}) = 3/7$ ,  $P(Y = \text{no cancer}) = 4/7$

$P(X = \text{positive}) = 3/7$ ,  $P(X = \text{negative}) = 4/7$

$P(Y = \text{cancer} \mid X = \text{positive}) = 2/3$

$P(Y = \text{cancer} \mid X = \text{negative}) = 1/4$

How about  $P(X = \text{positive} \mid Y = \text{cancer})$ ?  
 $2/3$ .

# Probability

Person	Y (cancer or not)	X (test positive or negative)
A	Cancer	Positive
B	Cancer	Negative
D	Cancer	Positive

$P(Y = \text{cancer}) = 3/7$ ,  $P(Y = \text{no cancer}) = 4/7$

$P(X = \text{positive}) = 3/7$ ,  $P(X = \text{negative}) = 4/7$

$P(Y = \text{cancer} \mid X = \text{positive}) = 2/3$

$P(Y = \text{cancer} \mid X = \text{negative}) = 1/4$

How about  $P(X = \text{positive} \mid Y = \text{cancer})$ ?  
2/3.

We can also get this answer  $2/3$  without counting by using our previous results obtained through Bayes Rule.

# Bayes Rule

$$P(A|B) = P(B|A) * P(A) / P(B)$$

Not going to prove it in class.

Remember we wanted to find  $P(X = \text{positive} \mid Y = \text{Cancer})$ .

$$P(X=\text{positive} \mid Y=\text{Cancer})$$

$$= P(Y=\text{Cancer} \mid X=\text{positive}) * P(X = \text{positive}) / P(Y = \text{Cancer})$$

$$= (2/3) * (3/7) / (3/7)$$

$$= 2/3 \text{ (same as direct counting!)}$$

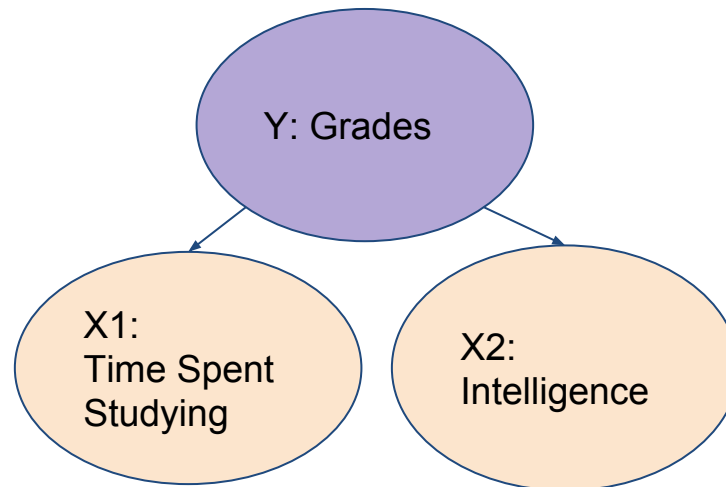
# Conditional Independence

**Definition:** Given we know the Label, the probability of feature X1 occurring is independent of feature X2.

**In Math:**  $P(X1, X2 \mid Y) = P(X1 \mid Y) * P(X2 \mid Y)$

Naive Bayes assumes all variables are conditionally independent, hence it is called “*naive*”.

This allows fast and efficient computation.



# Naive Bayes (Classification)

## Features

- X1: Age [young / old]
- X2: Tumor Size [none / small / large]

## Labels

- Y: [Cancer / No Cancer]

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

# Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if  $P(Y = \text{Cancer} \mid X1 = \text{old}, X2 = \text{small})$

or  $P(Y = \text{No Cancer} \mid X1 = \text{old}, X2 = \text{small})$  is greater

We can reform that equation using Bayes Rule:

$$P(Y = \text{Cancer} \mid X1, X2)$$

$$= P(X1, X2 \mid Y = \text{Cancer}) * P(Y = \text{Cancer}) / P(X1, X2)$$

$$P(Y = \text{No Cancer} \mid X1, X2)$$

$$= P(X1, X2 \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer}) / P(X1, X2)$$

Since we only care about which one is bigger, we can drop the  $P(X1, X2)$  term.

Determine if  $P(X1, X2 \mid Y = \text{Cancer}) * P(Y = \text{Cancer})$

or  $P(X1, X2 \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer})$  is greater.

Use the Conditional Independence Assumption

$$P(X1 \mid Y = \text{Cancer}) * P(X2 \mid Y = \text{Cancer}) * P(Y = \text{Cancer})$$

$$P(X1 \mid Y = \text{No Cancer}) * P(X2 \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer})$$



# Naive Bayes (Classification)

Suppose this person who is (X1) old and has a (X2) small tumor comes to you..

Determine if  $P(Y = \text{Cancer} \mid X1 = \text{old}, X2 = \text{small})$

or  $P(Y = \text{No Cancer} \mid X1 = \text{old}, X2 = \text{small})$  is greater

We can reform that equation using Bayes Rule:

$$P(Y = \text{Cancer} \mid X1, X2)$$

$$= P(X1, X2 \mid Y = \text{Cancer}) * P(Y = \text{Cancer}) / P(X1, X2)$$

$$P(Y = \text{No Cancer} \mid X1, X2)$$

$$= P(X1, X2 \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer}) / P(X1, X2)$$

Since we only care about which one is bigger, we can drop the  $P(X1, X2)$  term.

Determine if  $P(X1, X2 \mid Y = \text{Cancer}) * P(Y = \text{Cancer})$

or  $P(X1, X2 \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer})$  is greater.

Use the Conditional Independence Assumption

$$P(X1 \mid Y = \text{Cancer}) * P(X2 \mid Y = \text{Cancer}) * P(Y = \text{Cancer})$$

$$P(X1 \mid Y = \text{No Cancer}) * P(X2 \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer})$$

Depending on which term is larger, we predict if a person has cancer or not

# Going back to example data

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

$P(X1 = \text{old} \mid Y = \text{Cancer}) * P(X2 = \text{small} \mid Y = \text{Cancer}) * P(Y = \text{Cancer})$

=

$P(X1 = \text{old} \mid Y = \text{No Cancer}) * P(X2 = \text{small} \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer})$

=

# Going back to example data

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

$$P(X1 = \text{old} \mid Y = \text{Cancer}) * P(X2 = \text{small} \mid Y = \text{Cancer}) * P(Y = \text{Cancer}) \\ = \frac{2}{3} * \frac{1}{3} * \frac{3}{7} = 0.0952$$

$$P(X1 = \text{old} \mid Y = \text{No Cancer}) * P(X2 = \text{small} \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer}) \\ = \frac{2}{4} * \frac{1}{4} * \frac{4}{7} = 0.0714$$

# Going back to example data

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

$$P(X1 = \text{old} \mid Y = \text{Cancer}) * P(X2 = \text{small} \mid Y = \text{Cancer}) * P(Y = \text{Cancer}) \\ = \frac{2}{3} * \frac{1}{3} * \frac{3}{7} = 0.0952$$

$$P(X1 = \text{old} \mid Y = \text{No Cancer}) * P(X2 = \text{small} \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer}) \\ = \frac{2}{4} * \frac{1}{4} * \frac{4}{7} = 0.0714$$

Model Predicted this person has cancer!!! :o  
 $0.0952 > 0.0714$

# Naive Bayes Summary

Given training data that follows this format..

Feature X1	Feature X2	Feature X3	...	Feature X999	Y (Label) [A or B]
..	..	..	..	..	..
..	..	..	..	..	..

And you are given new data without labels that you want to classify

Feature X1	Feature X2	Feature X3	...	Feature X999	Y (Label) [A or B]
..	..	..	..	..	???
..	..	..	..	..	???

**Determine if**

$P(X1 | Y = A) * P(X2 | Y = A) * P(X3 | Y = A) * .. * P(X999 | Y = A) * P(Y = A)$

**OR**

$P(X1 | Y = B) * P(X2 | Y = B) * P(X3 | Y = B) * .. * P(X999 | Y = B) * P(Y = B)$

**Is greater**

# Now, your turn!

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

Given a new person who is X1 = young and X2 = large, what will the model predict?

**Remember! Determine if**

$P(X1 | Y = A) * P(X2 | Y = A) * P(X3 | Y = A) * .. * P(X999 | Y = A) * P(Y = A)$

**OR**

$P(X1 | Y = B) * P(X2 | Y = B) * P(X3 | Y = B) * .. * P(X999 | Y = B) * P(Y = B)$

**Is greater**

# Now, your turn!

Person	X1 Age [young / old]	X2 Tumor Size [none / S / L]	Y (cancer or not)
A	old	none	Cancer
B	old	small	Cancer
C	young	none	No Cancer
D	young	large	Cancer
E	old	none	No Cancer
F	old	none	No Cancer
G	young	small	No Cancer

**Given a new person who is X1 = young and X2 = large, what will the model predict?**

$$P(X1 = \text{young} \mid Y = \text{Cancer}) * P(X2 = \text{small} \mid Y = \text{Cancer}) * P(Y = \text{Cancer}) \\ = \frac{1}{3} * \frac{1}{3} * \frac{3}{7} = 0.0476$$

$$P(X1 = \text{young} \mid Y = \text{No Cancer}) * P(X2 = \text{small} \mid Y = \text{No Cancer}) * P(Y = \text{No Cancer}) \\ = \frac{2}{4} * \frac{1}{4} * \frac{4}{7} = 0.0714$$

0.0476 < 0.0714, NO CANCER!! :)

# Why Naive Bayes?

1. Simple and Easy to implement
2. Computationally fast
3. Works well on small datasets

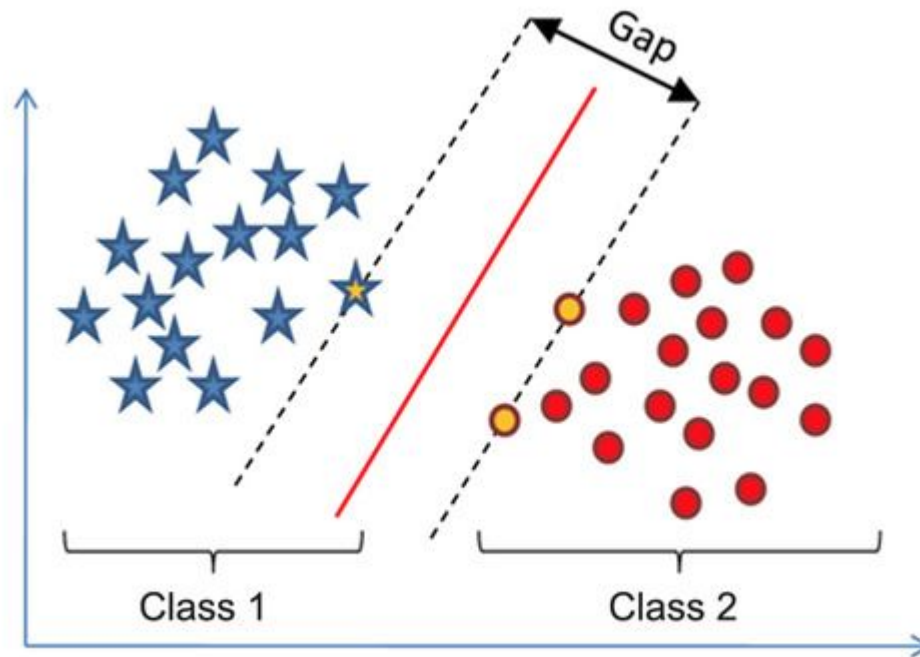
## Real World Examples

1. Classify an email as spam, or not spam
2. Classify a news article to its category



# Support Vector Machines

Finds a line that best separates 2 classes of points.



# Supervised Learning Recap:

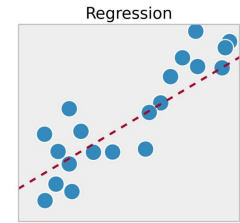
## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points

## Classification

- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take the most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points

# Supervised Learning Recap:



## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points

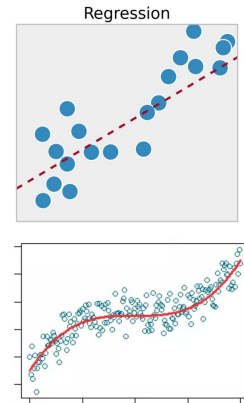
## Classification

- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take the most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points

# Supervised Learning Recap:

## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points



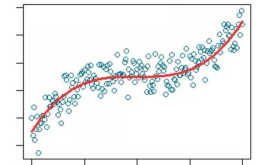
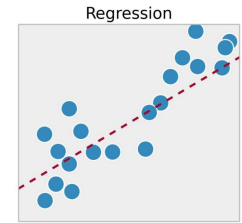
## Classification

- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take the most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points

# Supervised Learning Recap:

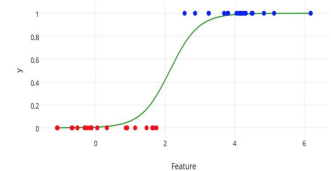
## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points



## Classification

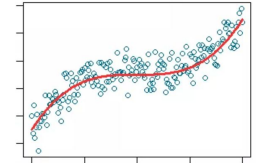
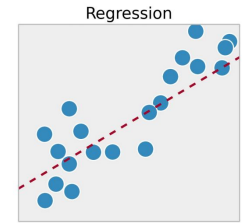
- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points



# Supervised Learning Recap:

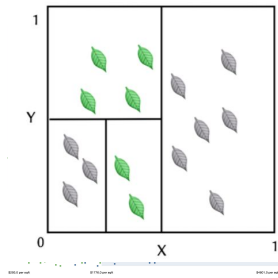
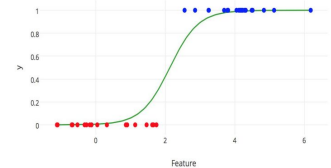
## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points



## Classification

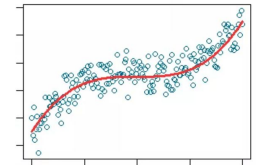
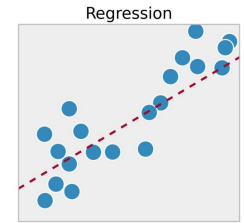
- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points



# Supervised Learning Recap:

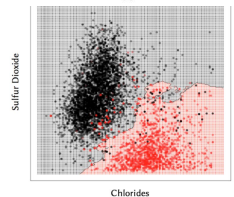
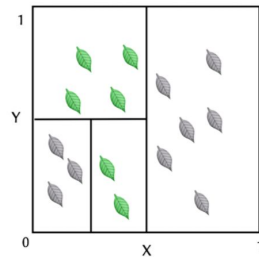
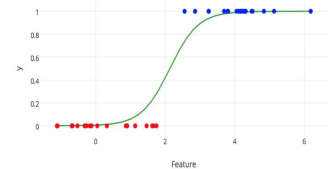
## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points



## Classification

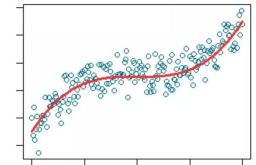
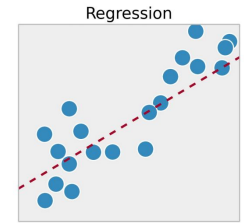
- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points



# Supervised Learning Recap:

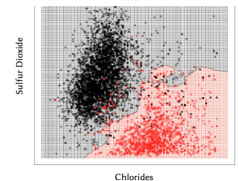
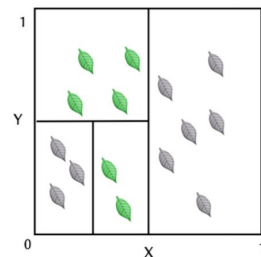
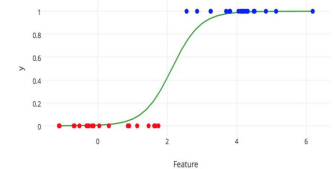
## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points



## Classification

- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points

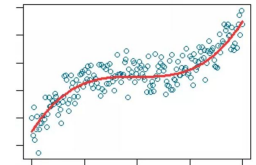
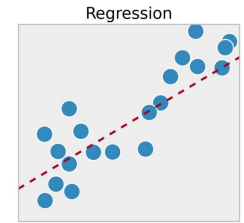




# Supervised Learning Recap:

## Regression

- Linear Regression
  - Draw a line of best fit through a set of points
- Polynomial Regression
  - Draw a curve of best fit through a set of points



## Classification

- Logistic Regression
  - Draw a line (logit curve) to separate  $\geq 2$  set of points.
- Decision Trees
  - Draw boundaries on axes to separate data points
- Random Forests
  - Group multiple decision trees and take most common vote
- K-Nearest Neighbors
  - Boundary determined by nearest points
- Naive Bayes!
  - Probability of features given that a label = Cancer or not
- Support Vector Machines
  - Draw a line/plane to separate 2 sets of points

