Errors, Naïve Bayes, and Decision Trees

David Quigley
CSCI 5622
2021 Fall

Week 2 -
this is Real!

Course Checkup

- HW 1 is out! It will run through an autograder, so…
  - keep all predefined functions with the same input / output options
    - You can add arguments *with defaults*
    - If a function *doesn't* have given return values, you can add them (usually in report functions)
  - Generalize your code
    - We should *never* have to, e.g., comment / uncomment code to get it to run
- Project Updates
  - Posting welcome on Piazza (due *Now*) - Counting in weekly participation
  - Group Formation Due Sept. 10 - All group members upload identical document
- Course Zoom is not "drop-in"
  - Feel free to use it – but I won't show up unless scheduled

Project Groups

- Forming Groups – up to 3 people maximum
    - Recommended 2 – 3 people
    - If you want to be solo, you need to seek permission at this group formation
- Starting to Form Problem Spaces / Datasets
    - This is not "set in stone" – things can change!
    - These are not "research" – we are finding public, accessible datasets or working with data you already have access to!*
- You can "double dip" your course project (with other courses, lab research, work, etc)
    - You will have to connect me with the other people and delineate what expectations and deliverables are for which purposes

*If you want to establish a new dataset via this class, we can talk!

Project Group Formation

- I (or the ISS) tried to comment on everyone's welcome (as of this afternoon)
    - If you gave an interest, background experience, or related idea, I tried to tie it into project inspiration!
    - This does *not* mean you have to go down that path at all with your projects.
        - Lots of folks probably shouldn't...
- There are a lot of connections between folks!
    - A lot of outdoors enthusiasts, a lot of sports fans, a lot of video game players, a lot of readers...
    - A lot of folks interested in NLP, a lot interested in CV, several interested in Space as a... problem space...
    - Any of these can inspire a project or more!

Prediction – College Admission

What are you going to predict for this NEW case with K = 1, using Manhattan distance?

Student	X_1	X_2	Y
A	1200	26	1
B	1450	28	1
C	1000	20	1
D	730	15	-1
NEW	720	16	???

Prediction – College Admission

What are you going to predict for this NEW case with K = 1, using Manhattan distance?

Student	X_1	X_2	Y
A	1200	26	1
B	1450	28	1
C	1180	20	1
D 230	730 10	15 240	-1
NEW	950 220	30 15 235	??? -1

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	
4	82	9421	.008	Positive	
3	17	9321	.04	Negative	
4	90	9128	.001	Negative	

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	
4	82	9421	.008	Positive	
3	17	9321	.04	Negative	
4	90	9128	.001	Negative	
3	16	9830	.04	???	

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	126.001
4	82	9421	.008	Positive	171638.001
3	17	9321	.04	Negative	259082
4	90	9128	.001	Negative	498281.0015
3	16	9830	.04	???	

Euclidian Distance – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
2	5	9832	.005	Positive	126.001
4	82	9421	.008	Positive	171638.001
3	17	9321	.04	Negative	259082
4	90	9128	.001	Negative	498281.0015
3	16	9830	.04	Positive?	

Normalization / Scaling

Transform X (Data) to X' (Scaled Data)

For (x_i) in X

  scale = $\max(x_i) - \min(x_i)$

  For $(x_{i,j})$ in (x_i)

    $x'_{i,j} = (x_{i,j} - \min(x_i)) / \text{scale}$

Scaling – N-Dimensional Vector

Euclidian Distance: $||x_i - x||^2$

X_1	X_2	X_3	X_4	Y	Distance
0	0	1	0.103	Positive	1.07
1	0.906	0.416	0.179	Positive	1.87
.5	0.141	0.274	1	Negative	0.52
1	1	0	0	Negative	3.00
.5	0.129	0.997	1	Negative	

Scaling – College Prediction

Student	SAT	ACT	GPA	Graduated?
A	1200 / 1600	26 / 36	3.2 / 4.0	Yes
B	1450 / 1600	28 / 36	3.5 / 4.0	Yes
C	1000 / 1600	20 / 36	3.0 / 4.0	Yes
D	730 / 1600	15 / 36	2.0 / 4.0	No
NEW	720 / 1600	16 / 36	2.2 / 4.0	???

Scaling – Housing Market?

Euclidian Distance: $||x_i - x||^2$

# Bedrms	Acres	Sq. Ft.	Radon	New Build?	Distance
2	5	9832	.005	Positive	
4	82	9421	.008	Positive	
3	17	9321	.04	Negative	
4	90	9128	.001	Negative	

Normalization / Scaling

Normalization is important for us to consider
 - It will allow us to consider variables on equal footing

Normalization and Scaling are important for us to consider *on a case by case basis*
 - Sometimes a "default" transformation won't make sense

Working with the built-in KNN library

Open Week 1 in-class Jupyter Notebook (the .ipynb file from Canvas)

K-Nearest Neighbors Algorithmic Complexity

One Query, m training examples, each with D features

$O(m * D)$

For the Naïve Case

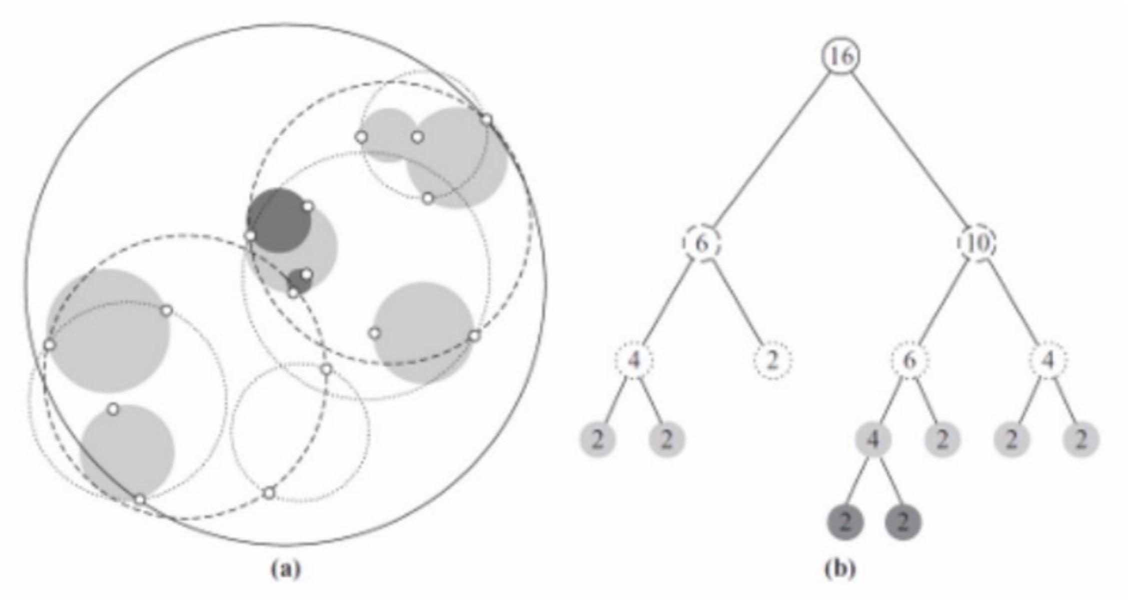
Scikit-learn has additional details in their implementation: https://scikit-learn.org/stable/modules/neighbors.html

K-Nearest Neighbors Algorithmic Complexity

How many different operations must I complete to find the 1 nearest neighbor?

How many different operations must I complete to find the 5 nearest neighbors?

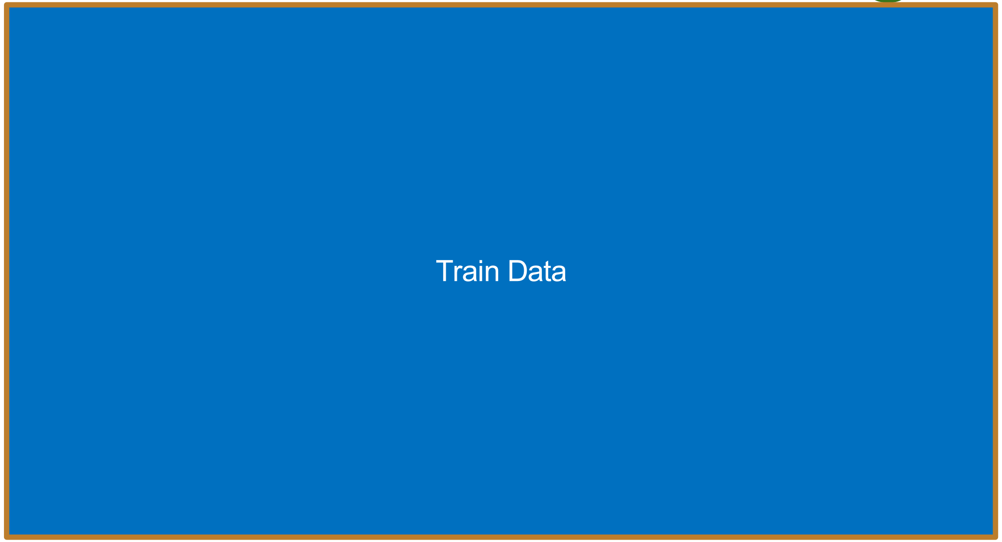
K-Nearest Neighbors – Tree Structure



Evaluating Models

How do we know if it works?

Homework 1 – Training & Test Sets



I pull out a random 20% of my data

Now I have something (probably) representative, AND

I'm not just testing inherent bias of my model or dataset

What is an Error?

We’ve looked at trying it out on a test set and getting an “accuracy”

Accuracy = # correct / # total

1) What is our “test set”?
2) Are all mistakes created equal?

Types of Errors

Classified As *Ground Truth*	**C**	**~C**
C		
~C		

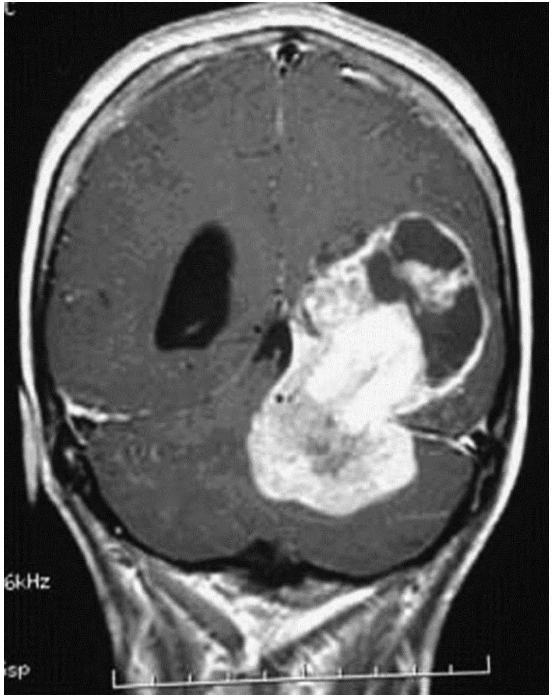
Types of Errors

Classified As / *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

incorrect

correct

What makes an error?



Types of Errors

Classified As *Ground Truth*	**Cancer**	**Not Cancer**
Cancer	True Positive (Hit)	False Negative (Miss)
Not Cancer	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

Types of Errors

Classified As / *Ground Truth*	**A**	**B**	**C**
A			
B			
C			

Types of Errors

Classified As / *Ground Truth*	**A**	**B**	**C**
A	hit	miss	miss
B	miss	hit	miss
C	miss	miss	hit

Types of Errors

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Calculating Errors

Error = $1/n * \Sigma_{i=1 \rightarrow n}(I(\hat{y}_i \neq y_i))$

 # errors / # instances

Is this always helpful? Meaningful?

Errors

Skewed Classes

- Classifying a day of sun in Yuma, AZ *(308 sunny days / year)*

    my classifier - isSunny = true

Errors

Skewed Classes

- Classifying a day of sun in Yuma, AZ *(308 sunny days / year)*

    my classifier - isSunny = true

    84 % Accuracy

- What if it's a case with 98% in one class?

    Is 98.5% accuracy helpful?

Errors

Skewed Classes

- Classifying a day of sun in Yuma, AZ *(308 sunny days / year)*

  my classifier - isSunny = true

  84 % Accuracy

- What if it's a case with 98% in one class?

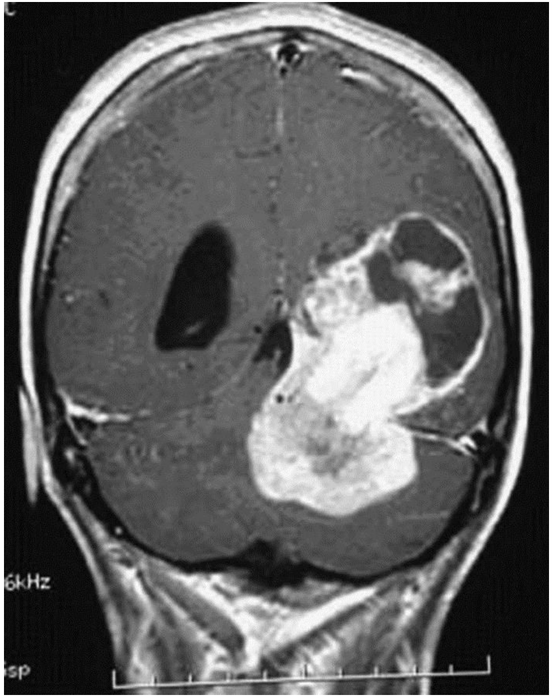
  Is 98.5% accuracy helpful?

Are all errors created equal?

Are all errors created equal?

Truth – No Cancer

Classified Cancer



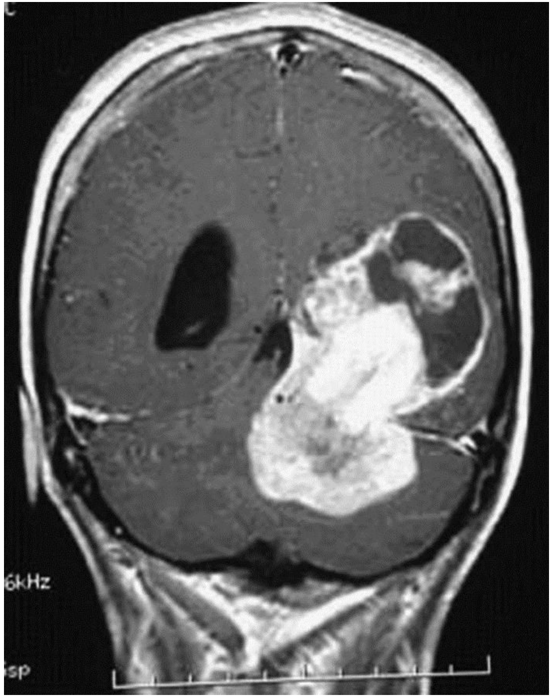
Truth – Cancer

Classified No Cancer

Are all errors created equal?

Truth – No Cancer
Classified Cancer

Goes to get a new test – more expensive but more accurate. Discovers true status.



Truth – Cancer
Classified No Cancer

Moves on with their day. Does not pursue treatments. Cancer worsens.

Types of Errors

Predicted Positive Rate (Precision) = Hits / (Hits + False Alarm)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

Predicted Negative Rate (Negative Predictive Value) = Corr. Rej / (Miss + Corr. Rej.)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

True Positive Rate (Sensitivity, Recall) = Hits / (Hits + Miss)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Types of Errors

False Positive Rate (Specificity) = Corr. Rej. / (False Alarm + Corr. Rej.)

Classified As *Ground Truth*	**C**	**~C**
C	True Positive (Hit)	False Negative (Miss)
~C	False Positive (False Alarm)	True Negative (Correct Rejection)

Confusion Matrix

Classified As *Ground Truth*	**C**	**~C**
C	Count of True Positives (Hit)	Count of False Negatives (Miss)
~C	Count of False Positives (False Alarm)	Count of True Negatives (Correct Rej.)

Confusion Matrix – Beyond Binary Decisions

https://en.wikipedia.org/wiki/Iris_flower_data_set

Confusion Matrix

Classified As *Ground Truth*	**C**	**~C**
C	Count of True Positives (Hit)	Count of False Negatives (Miss)
~C	Count of False Positives (False Alarm)	Count of True Negatives (Correct Rej.)

Confusion Matrix

Classified As / *Ground Truth*	**A**	**B**	**C**
A	Count of A / A Hits	Count of B / A Misses	Count of C / A Misses
B	Count of A / B Misses	Count of B / B Hits	Count of C / B Misses
C	Count of A / C Misses	Count of B / C Misses	Count of C / C Hits

Problem Set 1

You now have everything you need to work through Problem Set 1!

Problem Space – College Admissions (Week 1)

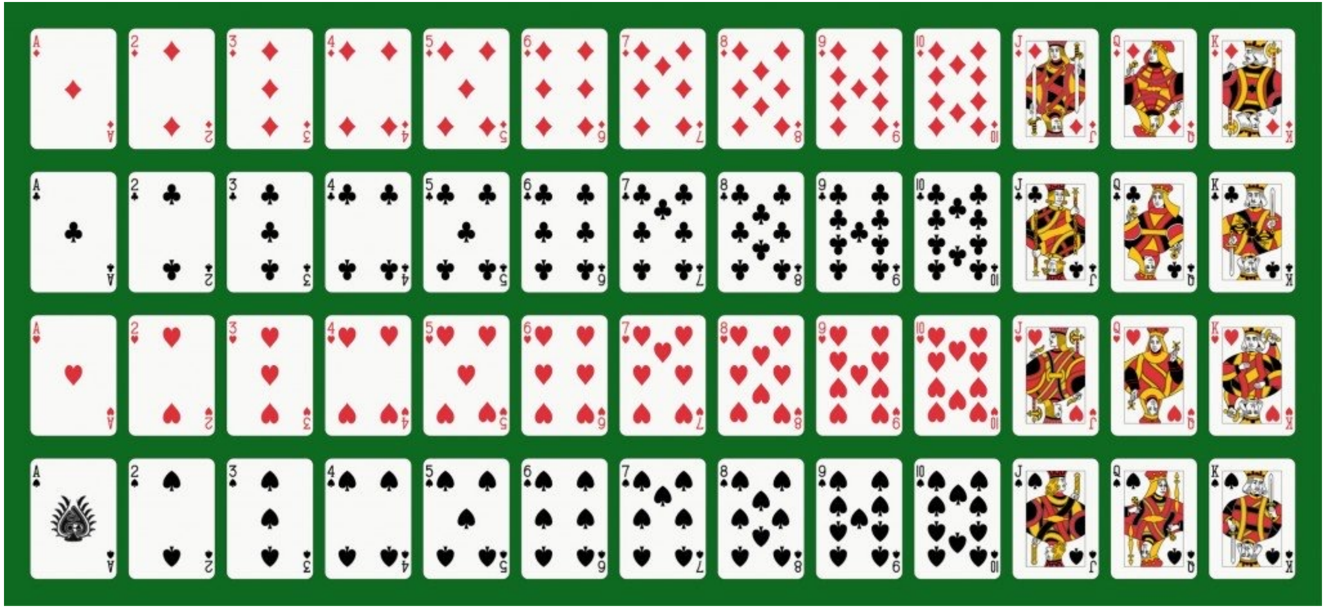
The following scenario isn't fully true, but it's close to what we do in college admissions…

I am trying to decide if a student should be admitted to my university. I have their SAT and ACT scores and their HS GPA. I also have the history of students who have attended in the past, their SAT / ACT / HS GPA as well as whether or not they graduated. I only want to admit new students if they will graduate.

Problems We Faced

Naïve Bayes

Probability Refresher (see Add'l Reading)



52 Cards

2 Colors {Red, Black}

4 Suits {D,C,H,S}

13 Values {A,2,3,...K}

Probability Refresher

V – Random variable of *Value*

v = "ace", v = 7, etc.

p(v) = probability that card has value "v"

C – Random variable of *Color*

C = Red or C = Black

p(c) = probability that card has color "c"

Joint probability – probability of *multiple events simultaneously*

p(v,c) = probability that card has value "v" *and* color "c"

probability of a Red 7 can be written as p(v=7,c=red) or p(7,red)

1/26

Probability Refresher – Product Rule

Product rule: p(A,B) = p(A|B)*p(B) = p(B|A)*p(A)

The *Joint* probability is equal to to the *conditional* probability multiplied by the probability of the condition (the *marginal* probability).

red 7 red 7 7

p(C,V) = p(C|V)*p(V) = p(V|C)*p(C)

1/26 1/2 1/13

P(Red,7) =

p(Red) =

P(7) = 1/13

p(Red,7) =

Probability Refresher – Chain Rule

For three variables, A,B,C, we can use the Product Rule repeatedly to show:

$$p(A,B,C) = p(a) * p(b|a) * p(c|b,a)$$

For *i* random variables $X_{1:I}$, we have:

$$p(X_{1:i}) = p(X_1) * p(X_2|X_1) * \ldots * p(X_i|X_{1:i-1})$$

Proving the Chain Rule is left as an exercise for the interested student

Probability Refresher – Sum Rule

Conversely, if we know the *joint* probabilities, we can compute the *marginal.*

$p(A) = \Sigma_b p(A,B=b) = \Sigma_b p(A|B=b) * p(B=b)$

Given p(C=Red,V=7) and p(C=Black,V=7), calculate p(V=7)

Probability – Bayes Rule

p(A,B) = p(A|B)*p(B) = p(B|A)*p(A)

p(X,Y) = p(Y|X)*p(X) → p(Y|X) = p(X,Y) / p(X) $→ p(Y|X) = (p(X|Y)*p(Y)) / p(X)

$$p(Y|X) = \frac{p(X|Y)*p(Y)}{p(X)}$$

$ = assuming p(X) > 0 (i.e. our data is possible)

Probability – Bayes Rule + Sum Rule

If we evaluate the

$$p(Y = y|X = x) = \frac{p(X = x \mid Y = y)*p(Y = y)}{p(X = x)}$$

And we use the *sum rule*...

$$p(Y = y|X = x) = \frac{p(X = x \mid Y = y)*p(Y = y)}{\Sigma_{y'} p(X=x|Y=y') * p(Y=y')}$$

We can compute p(y|x) from nothing but p(x|y) and p(y)

Bayes Rule - Cancer

We have a Cancer Test = {pos,neg} and Cancer {C,~C}

1% of people have cancer $p(\overset{Y}{C=\text{true}}) = .01$

90% of people *with cancer* will test *positive* $p(\overset{X}{\text{test}=\text{pos}} \mid \overset{Y}{C=\text{true}}) = .9$

8% of people *without cancer* will test *positive* $p(\overset{X}{\cancel{t}}=\text{pos} \mid \overset{Y}{C=\text{false}}) = .08$

Bayes Rule - Cancer

END

Say you have a *positive* test. What is the probability that you have cancer?

$$p(C \mid pos)$$

Thursday

Course Check-In

- Problem Set 1
  - You will turn in your Jupyter Notebook (.ipynb file)
    - You'll submit "after a completed run" – so we'll see the graphs you just generated, etc.
    - Your text descriptions should be generalized, but can reference those specific plots
- Project Group Formation
  - Post available on Piazza – "Search for Teammates!" is a great place to start

Bayes Rule - Cancer

We have a Cancer Test = {pos,neg} and Cancer {C,~C}

1% of people have cancer $\quad p(\overset{Y}{C=\text{true}}) = .01$

90% of people *with cancer* will test *positive* $\quad p(\overset{X}{\text{test}=\text{pos}} \mid \overset{Y}{C=\text{true}}) = .9$

8% of people *without cancer* will test *positive* $\quad p(\overset{X}{}=\text{pos} \mid \overset{Y}{C=\text{false}}) = .08$

Bayes Rule - Cancer

Say you have a *positive* test. What is the probability that you have cancer?

$$p(C \mid pos)$$

Bayes Rule - Cancer

Say you have a *positive* test. What is the probability that you have cancer?

$$p(C \mid pos)$$

Can't ignore the fact that there's a *small percentage of people with cancer*

$$p(C) = 0.01$$

Bayes Rule - Cancer

$$p(C|pos) = \frac{p(pos|C)*p(C)}{p(pos|C)*p(C) + p(pos|\sim C)*p(\sim C)}$$ as our formulation

Bayes Rule - Cancer

$$p(C|pos) = \frac{.90 * .01}{.90 * .01 + .08 * .99} = .10$$

Naïve Bayes Classification

We want to model the joint probability: p(x,y)

What we *actually care* about is the probability of our *answer* given the *data*: p(y|x)

$$p(Y|X) = \frac{p(X|Y) * p(Y)}{p(X)}$$

Or…

$$\text{Posterior probability} = \frac{\text{class-conditional} * \text{prior}}{\text{evidence}}$$

Naïve Bayes Classification

“What is the possibility that an example is of class *c* given its observed features?”

- What is the probability it is a grad (pos) given its HS outcomes?

Given a HS *x*:

If p(pos | x) >= p(neg | x)

classify as pos

Else

classify as neg

Class-conditional probability p(X | Y)

i.e. "Likelihood"
"Given (or *assuming*) y = c, what is the probability x is observed?"

"Given my assumptions about failing students, what's the probability that I see this particular student?"

p(x = [GPA=1.2,SAT=550] | y = neg)

NOTE: Now we're assuming joint probability of *features*

Naïve Bayes – Assumption of Independence

Features of **x** are *conditionally independent* for a given class **y**.

Naïve Bayes – Assumption of Independence

Features of **x** are *conditionally independent* for a given class **y**.

Take two weighted coins, C_1 and C_2.

Naïve Bayes – Assumption of Independence

Features of **x** are *conditionally independent* for a given class **y**.

Take two weighted coins, C_1 and C_2.

Pick a coin and flip it 3 times. Which coin did we flip?

$p(x=[HHT] \mid C_1) = p(H|C_1) * p(H|C_1) * p(T|C_1)$

Vs

$p(x=[HHT] \mid C_2) = p(H|C_2) * p(H|C_2) * p(T|C_2)$

Naïve Bayes – Assumption of Independence

Features of **x** are *conditionally independent* for a given class **y**.

Take two weighted coins, C_1 and C_2.
Pick a coin and flip it 3 times. Which coin did we flip?
$p(x=[HHT] \mid C_1) = p(H|C_1) * p(H|C_1) * p(T|C_1)$
Vs
$p(x=[HHT] \mid C_2) = p(H|C_2) * p(H|C_2) * p(T|C_2)$

Draw 3 cards from a deck of cards, one at a time (no replacement). Does this still hold?

Naïve Bayes – Assumption of Independence

Features of **x** are *conditionally independent* for a given class **y**.

p(x=[GPA=1.2,SAT=550,ACT=21] | neg) = p(GPA=1.2 | neg) * p(SAT=550| neg) * p(ACT=21 | neg)

Is this valid?

Probably not, but a) we'll make the assumptions for both pos and neg, and b) it makes feature conditionals easy to estimate

Naïve Bayes – Class-Conditional Probability

Features of **x** are *conditionally independent* for a given class **y**.

p(x=[GPA=1.2,SAT=550,ACT=12] | neg) = p(GPA=1.2| neg) * p(SAT=550| neg) * p(ACT=21 | neg)

How do we calculate the class-conditional probability?

Estimate from Data!

$$P(GPA=1.2|\ neg) = \frac{\#\ \text{failing students with a HS GPA of 1.2}}{\#\ \text{failing students}}$$

Naïve Bayes Classification

We want to model the joint probability: $p(x,y)$

What we actually care about is the probability of our *answer* given the *data*: $p(y|x)$

$$p(Y|X) = \frac{p(X|Y)*p(Y)}{p(X)}$$

Or…

$$\text{Posterior probability} = \frac{\cancel{\text{class-conditional}} * \text{prior}}{\text{evidence}}$$

Prior Probability p(y)

"What is the probability of encountering class *c*?"

p(neg) = "What is the probability that any given student will fail?"

How do we get this information?

from data

Prior Probability p(y)

“What is the probability of encountering class *c*?”

p(neg) = “What is the probability that any given student will fail?”

How do we get this information?
Ask Someone Who Knows!

Assume 80% of students fail

this class assumption does not necessarily match the real world!

Prior Probability p(y)

"What is the probability of encountering class *c*?"

p(neg) = "What is the probability that any given student will fail?"

How do we get this information?

Estimate it from Data!

$$P(\text{neg}) = \frac{\text{\# failing students}}{\text{\# students}}$$

Naïve Bayes Classification

We want to model the joint probability: p(x,y)

What we actually care about is the probability of our *answer* given the *data*: p(y|x)

$$p(Y|X) = \frac{p(X|Y)*p(Y)}{p(X)}$$

Or…

$$\text{Posterior probability} = \frac{\cancel{\text{class conditional}} * \cancel{\text{prior}}}{\text{evidence}}$$

Evidence p(x)

"The probability of encountering x independent of class"

"the probability of getting that exact HS outcome"

We *could* estimate this using the sum rule…

Naïve Bayes Classification

"What is the possibility that an example is of class *c* given its observed features?"

- What is the probability it is a positive grad given its HS features ?

Given HS features *x*:

If p(pos | x) >= p(neg | x)

classify as pos

Else

classify as neg

Evidence p(x)

“The probability of encountering x independent of class”

“the probability of getting that exact message”

We *could* estimate this using the sum rule…

But it doesn’t actually matter in our decision making process

$$\cancel{p(x)} \cdot \frac{p(x \mid pos) * p(pos)}{\cancel{p(x)}} \;\; >=<? \;\; \frac{p(x \mid neg) * p(neg)}{\cancel{p(x)}} \cdot \cancel{p(x)}$$

Evidence p(x)

“The probability of encountering x independent of class”

“the probability of getting that exact message”

We *could* estimate this using the sum rule…

But it doesn’t actually matter in our decision making process

p(x | pos) * p(pos) >=<? p(x | neg) * p(neg)

It’s not really *probability* anymore, but we have *scores* for each condition…

Naïve Bayes Classification - Example

	Pos	**Neg**	**Neg**	**Neg**	**Pos**
GPA	2.5	1.9	1.2	2.1	4.0
SAT	1100	990	750	1100	1600
ACT	21	16	16	21	36

Naïve Bayes Classification - Example

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

	Pos	**Neg**	**Neg**	**Neg**	**Pos**
GPA	2.5	1.9	1.2	2.1	4.0
SAT	1100	990	750	1100	1600
ACT	21	12	16	21	36

Naïve Bayes Classification - Example

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Naïve Bayes Classification - Example

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

$p(pos \mid 1.2, 550, 21) >=<? \; p(neg \mid 1.2,550,21)$

	Pos	Neg	Neg	Neg	Pos	???
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Naïve Bayes Classification - Example

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

p(pos | 1.2, 550, 21) = ???

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Naïve Bayes Classification - Example

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

PRODUCT RULE

p(pos | 1.2, 550, 21) = p(1.2, 550, 21 | pos) * p(pos)

	Pos	Neg	Neg	Neg	Pos	???
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Naïve Bayes Classification - Example

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

ASSUME INDEPENDENCE

0/2 1/2 2/5

p(pos | 1.2, 550, 21) = p(1.2 | pos) * p(550 | pos) * p(21 | pos) * p(pos)

	Pos	Neg	Neg	Neg	Pos	???
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Overcoming Data Problems

We've never seen a student with an SAT of 550 before! How might you address this?

Threshold

add a correction factor

discard feature

find the NN

Fit a Distribution

	Pos	Neg	Neg	Neg	Pos	???
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Overfitting Solutions - Binning

Binning – Assign values to bins

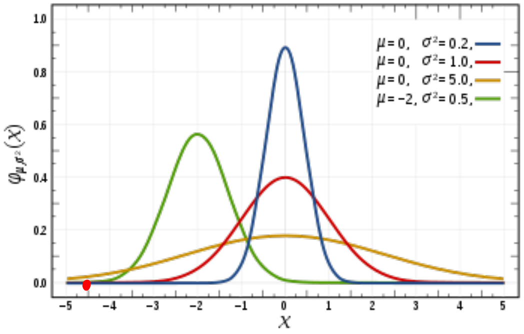
Bins can vary in size, number, shape

- Equity concern?

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	C	D	D	C	A	D
SAT	1001-1400	751-1000	0-750	1001-1400	1401-1600	0-750
ACT	21-25	11-15	16-20	21-25	31-36	21-25

Overfitting Solutions – Increasing Sample Size

- *"If I just keep taking samples, surely I'll eventually cover my whole sample space!"*



Overfitting Solutions – Targeted Sampling

- *"If I know to expect something, I can go seek out examples for training!"*

	Pos	Neg	Neg	Neg	Pos	???
GPA	C	D	D	C	A	D
SAT	1001-1400	751-1000	0-750	1001-1400	1401-1600	0-750
ACT	21-25	11-15	16-20	21-25	31-36	21-25

Overfitting Solutions – Trust the Experts

Sometimes there could be databases or lists of interest to your problem, e.g. a "cutoff score" for SAT scores, or a "valence score" for NLP word connotations.

- Does it actually fit my problem space?
- Do I trust it?

Overfitting Solutions – Smoothing

Builds on the assumption that “nothing is impossible”

Overfitting Solutions – Smoothing

Step 1) Start with a *completely* naïve, untrained understanding

- “a) All possible examples are equally likely to exist in my problem space and b) All classes are equally likely”
- All HS features are equally likely to appear in a student and all students are equally likely to be positive or negative

$p(x) = \frac{1}{|V|}$ $p(y) = \frac{1}{|Y|}$

Overfitting Solutions – Smoothing

Step 2) Add our training cases to the evidence set

BEFORE

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{\text{\# positive grads with “x” in them}}{\text{\# positive grads}}$$

AFTER

$$p(x|y) = \frac{p(x,y) + 1}{p(y) + |V|} = \frac{\text{\# positive grads with “x” (found in training) + 1}}{\text{\# positive grads (in training) + tot. student options}}$$

Add-1 or Laplace Smoothing

Overfitting Solutions – Smoothing

Step 2) Add our training cases to the evidence set

BEFORE

$$p(y) = \frac{\text{\# positive students}}{\text{\# students}}$$

AFTER

$$p(y) = \frac{\text{\# positive students (found in training)} + 1}{\text{\# students (found in training)} + \text{total classes}}$$

Add-1 or Laplace Smoothing

Overcoming Data Problems

Binning + Smoothing

GPA: 5 Bins - A, B, C, D, F

SAT: 2 Bins - >600

ACT: 2 Bins - >25

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	2.5	1.9	1.2	2.1	4.0	1.2
SAT	1100	990	750	1100	1600	550
ACT	21	12	16	21	36	21

Naïve Bayes Classification

If I get a new student with [GPA = 1.2, SAT = 550, ACT = 21], do we expect them to graduate (pos) or not (neg)?

p(pos | 1.2, 550, 21) = p(D | pos) * p(- | pos) * p(Low | pos) * p(pos)

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	C	D	D	C	A	D
SAT	+	+	+	+	+	-
ACT	Low	Low	Low	Low	High	Low

Naïve Bayes Classification

p(pos | 1.2, 550, 21) = p(D | pos) * p(- | pos) * p(Low | pos) * p(pos)

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	C	D	D	C	A	D
SAT	+	+	+	+	+	-
ACT	Low	Low	Low	Low	High	Low

Naïve Bayes Classification

p(pos | 1.2, 550, 21) = p(D | pos) * p(- | pos) * p(Low | pos) * p(pos)

Add-1 Smoothing

$$p(x|y) = \frac{p(x,y) + 1}{p(y) + |V|} = \frac{\text{\# positive grads with “x” (found in training)} + 1}{\text{\# positive grads (in training)} + \text{tot. student options}}$$

$$p(y) = \frac{\text{\# positive students (found in training)} + 1}{\text{\# students (found in training)} + \text{total classes}}$$

	Pos	**Neg**	**Neg**	**Neg**	**Pos**	**???**
GPA	C	D	D	C	A	D
SAT	+	+	+	+	+	-
ACT	Low	Low	Low	Low	High	Low

Naïve Bayes Classification

p(pos | 1.2, 550, 21) = p(D | pos) * p(- | pos) * p(Low | pos) * p(pos)

*= 1/7 * 1/4 * 2/4 * 3/7 = 3/392 or 0.0077ish*

Add-1 Smoothing

$$p(x|y) = \frac{p(x,y) + 1}{p(y) + |V|} = \frac{\text{\# positive grads with “x” (found in training) + 1}}{\text{\# positive grads (in training) + tot. student options}}$$

$$p(y) = \frac{\text{\# positive students (found in training) + 1}}{\text{\# students (found in training) + total classes}}$$

	Pos	Neg	Neg	Neg	Pos	???
GPA	C	D	D	C	A	D
SAT	+	+	+	+	+	-
ACT	Low	Low	Low	Low	High	Low

Naïve Bayes Classification

p(neg | 1.2, 550, 21) = p(D | neg) * p(- | neg) * p(Low | neg) * p(neg)

Add-1 Smoothing

$$p(x|y) = \frac{p(x,y) + 1}{p(y) + |V|} = \frac{\text{\# negative grads with “x” (found in training)} + 1}{\text{\# negative grads (in training)} + \text{tot. student options}}$$

$$p(y) = \frac{\text{\# negative students (found in training)} + 1}{\text{\# students (found in training)} + \text{total classes}}$$

	Pos	Neg	Neg	Neg	Pos	???
GPA	C	D	D	C	A	D
SAT	+	+	+	+	+	-
ACT	Low	Low	Low	Low	High	Low

Naïve Bayes Classification

p(neg | 1.2, 550, 21) = p(D | neg) * p(- | neg) * p(Low | neg) * p(neg)

= *3/8* * *1/5* * *4/5* * *4/7* = *6/175 or 0.034ish*

Add-1 Smoothing

$$p(x|y) = \frac{p(x,y) + 1}{p(y) + |V|} = \frac{\text{\# negative grads with “x” (found in training)} + 1}{\text{\# negative grads (in training)} + \text{tot. student options}}$$

$$p(y) = \frac{\text{\# negative students (found in training)} + 1}{\text{\# students (found in training)} + \text{total classes}}$$

NEG

	Pos	Neg	Neg	Neg	Pos	???
GPA	C	D	D	C	A	D
SAT	+	+	+	+	+	-
ACT	Low	Low	Low	Low	High	Low

Naïve Bayes - Probabilities

The probability of tweet x of length D being of class y

$p(y|x) = p(y) * \Pi_{i=1 \rightarrow D}(p(x_i|y))$

We get some complicated, small numbers multiplication

- Theoretically valid
- We work with finite computing machines
- Leads to *underflow*

Let's take the logarithm instead!

Naïve Bayes – Logarithms (ln)

Logarithm math refresher:

- $\ln(ab) = \ln(a) + \ln(b)$
- $\ln(x)$ is monotonically increasing, so argmax() is still valid

$$y = \text{argmax}_c \ \ln(p(c)) + \Sigma_{i=1 \rightarrow D} \ln(p(x_i|c))$$

Decision Trees

Machine Decision Making – 101

Think back to your very first days of programming…

I ask you to have a program take in two pieces of information:

- sun
- wind

And output whether or not I am going to play tennis.

- Only if it's sunny and not too windy

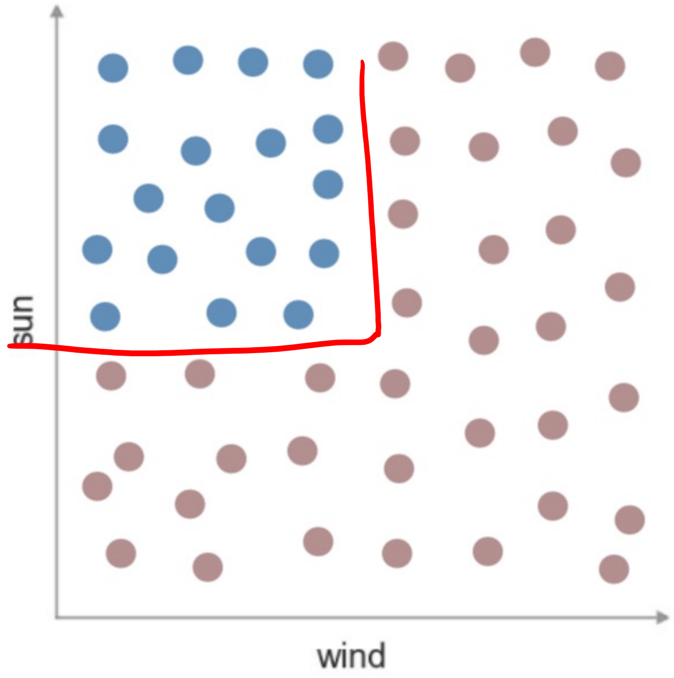
What does this program look like?

```
if(sunny && !windy):
    tennis = True
```

Machine Decision Making – Tennis

Consider this visualization of my data
And our model of a predictor

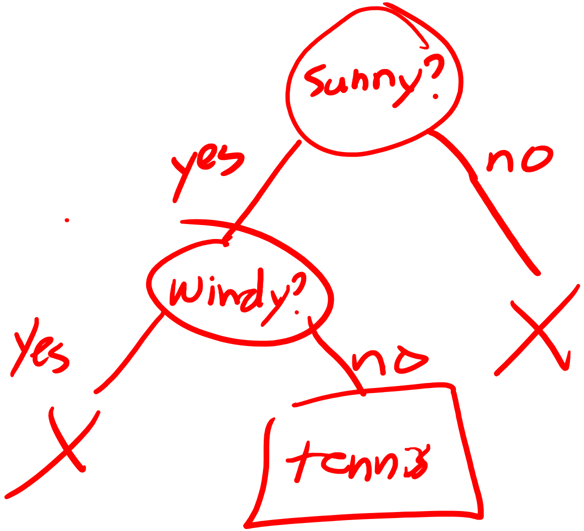
What would the decision boundary look like here?



Machine Decision Making - 201

Think back to your very second days of programming (i.e. data structures)...

How might we represent a series of nested if/else statements?



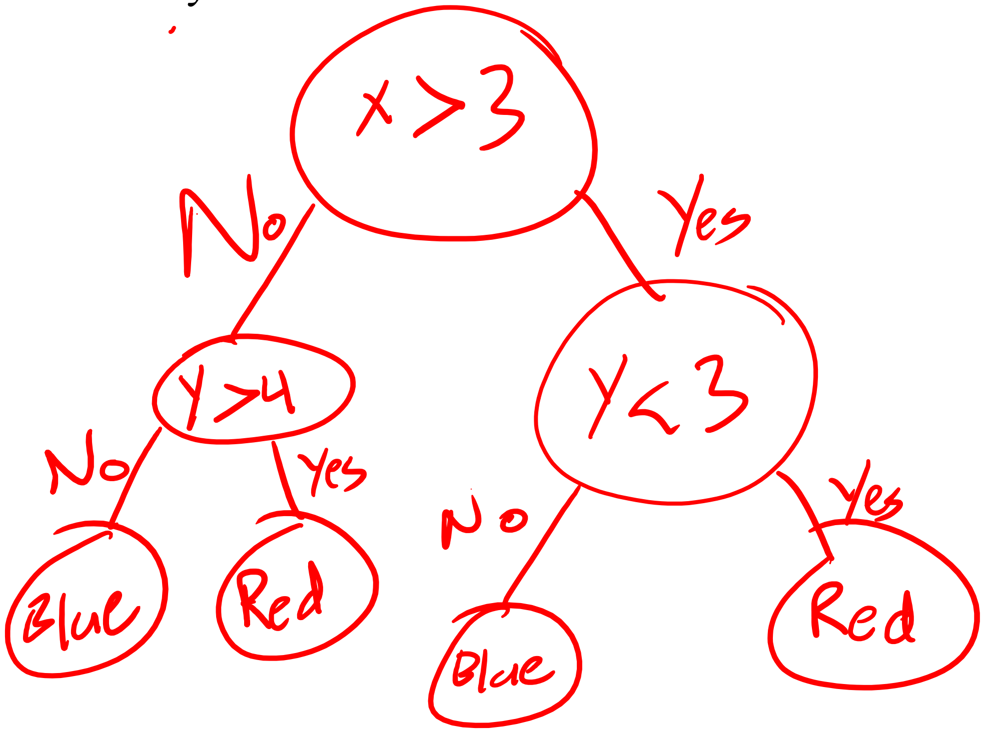
Decision Trees

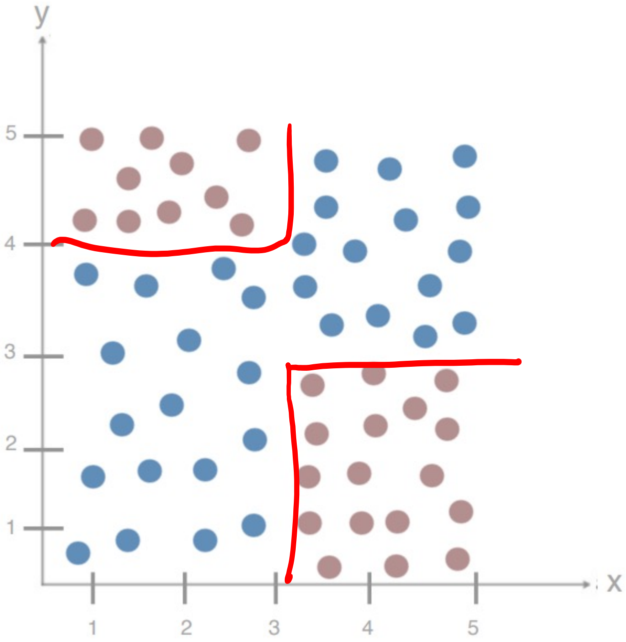
Creating nonlinear decision boundaries via the *union* of multiple linear decision boundaries

These are the basis of a lot of more complex algorithms – *Stay Tuned*

Decision Trees – Adding Complexity

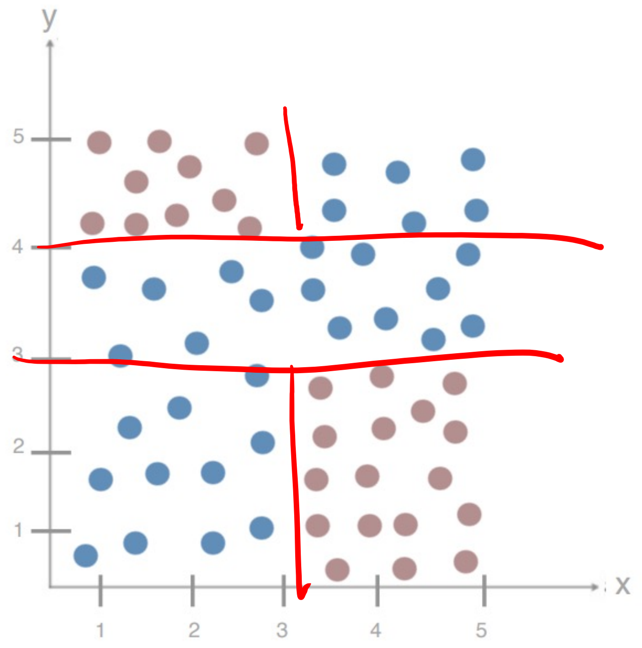
What is my Decision Tree for this dataset?





Decision Trees – Adding Complexity

What happens if I split on Y first?

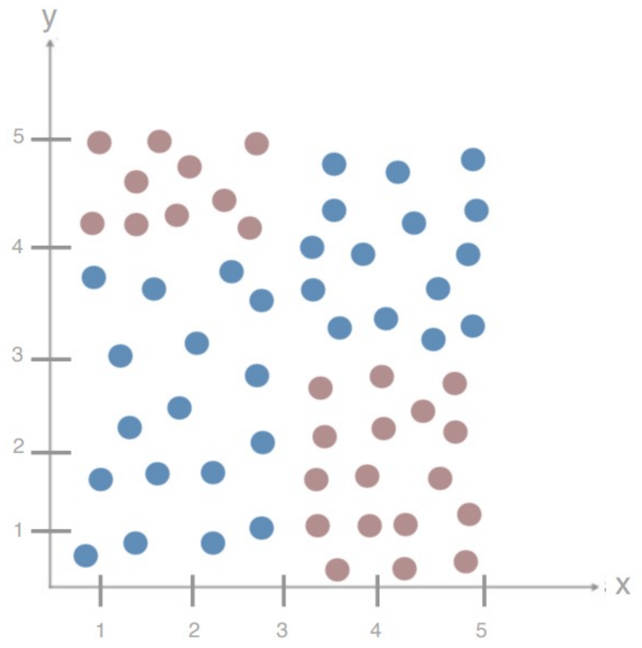


Decision Trees – Adding Complexity

How do we decide the order of our splits?

How do we decide the location of splits?
With continuous variables?

How do we know we're done?



Decision Trees – Choosing your Split

Simplifying our problem space – binary features

Test A	Test B	Test C	Cancer?
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes

Decision Trees – Choosing your Split

What feature should I split on first?

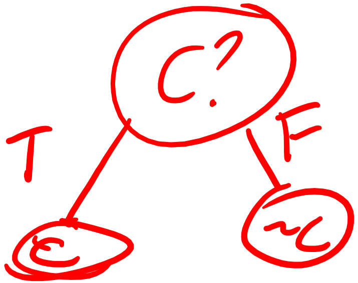
Test A	**Test B**	**Test C**	**Cancer?**
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes

Decision Trees – Choosing your Split

What feature should I split on first?

What is my entire tree?

Test A	Test B	Test C	Cancer?
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes



Decision Trees – Finding the Best Split

We need to find the split that gives us the best arrangement

We need to find the split that creates the most order from our chaos

Test A	Test B	Test C	Cancer?
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes

Decision Trees - Entropy

A measure of the impurity / messiness / chaos of a set of examples

Decision Trees - Entropy

A measure of the impurity / messiness / chaos of a set of examples

$\Sigma_{c=1...n} -p_c * \log_2(p_c)$

Decision Trees - Entropy

A measure of the impurity / messiness / chaos of a set of examples

$\Sigma_{c=1\ldots n}$ [$-p_c * \log_2(p_c)$]
C for each possible class

In the binary case...

$-p * \log_2(p) - (1-p) * \log_2(1-p)$

Decision Trees - Entropy

A measure of the impurity / messiness / chaos of a set of examples

Entropy = $\Sigma_{c=1...n}$ [$-p_c * \log_2(p_c)$]

C for each possible class

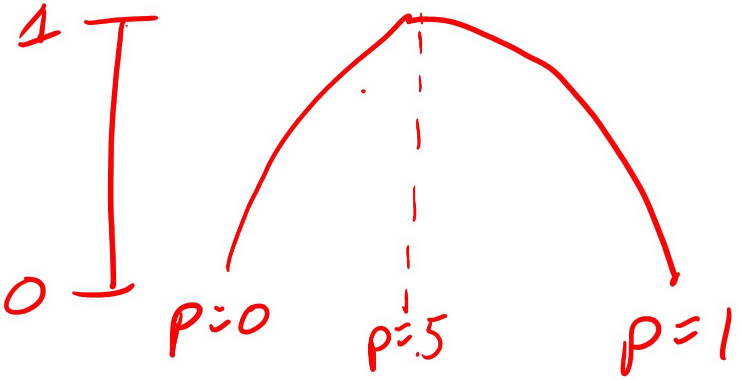
In the binary case...

Entropy = $-p * \log_2(p) - (1-p) * \log_2(1-p)$

$1 \times \log(1)$

What happens at p = 1? p = 0? p = 0.5?

0 0 1



Decision Trees – Entropy

A measure of the impurity / messiness / chaos of a set of examples

Entropy = $\Sigma_{c=1...n}$ [$-p_c * \log_2(p_c)$]

C for each possible class

In the binary case...

Entropy = $-p * \log_2(p) - (1-p) * \log_2(1-p)$

What happens at p = 1? p = 0? p = 0.5?

Minimum – perfect isolation of one class, Entropy = 0

Maximum – perfect split of data, Entropy = 1

Decision Trees - Entropy

A measure of the impurity / messiness / chaos of a set of examples

Entropy = $\Sigma_{c=1...n} [-p_c * \log_2(p_c)]$

C for each possible class

In the general case... (3 class case)

$P_c = .33$ for all c

Entropy is maximized (~ 1.56)

$P_1 = .4$ $P_2 = .4$ $P_3 = .2$

Entropy is reduced (~1.52)

P = {1,0,0}

Entropy is 0

Decision Trees – Entropy of Cancer

Entropy of our root node?

Test A	Test B	Test C	Cancer?
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes

Decision Trees – Entropy of Cancer

Entropy of our root node?

Test A	Test B	Test C	Cancer?
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes

$-.5 * \log_2(.5) - .5 * \log_2(.5) = 1$

Maximum Entropy

Decision Trees – Entropy of Cancer

How do we pick our split?

Test A	Test B	Test C	Cancer?
Pos	Neg	Neg	No
Pos	Pos	Neg	No
Neg	Pos	Pos	Yes
Pos	Neg	Pos	Yes

Choose the feature that reduces our entropy the most!

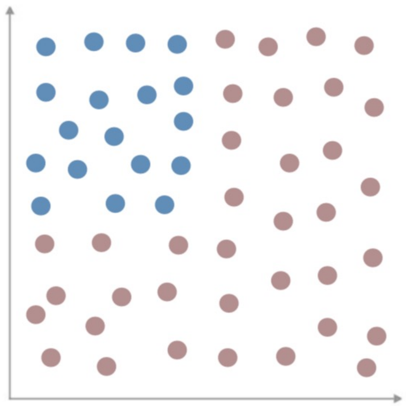
Decision Trees – Minimizing Entropy

We are creating a split to minimize the entropy of our set...

Decision Trees – Minimizing Entropy

We are creating a split to minimize the entropy of our set...

But really, we're creating *two* sets, each with their own entropy, but a smaller number of samples in each.

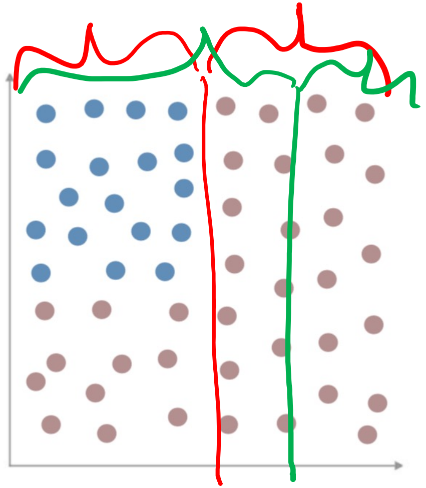


Decision Trees – Minimizing Entropy

We are creating a split to minimize the entropy of our set...

But really, we're creating *two* sets, each with their own entropy, but a smaller number of samples in each.

Is it better to split evenly, or finely?

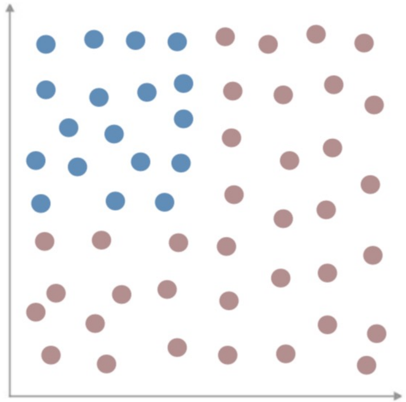


Decision Trees – Minimizing Entropy

We are creating a split to minimize the entropy of our set…

But really, we're creating *two* sets, each with their own entropy, but a smaller number of samples in each.

What makes for a good split?



Decision Trees – Minimizing Entropy END

We are creating a split to minimize the entropy of our set...

But really, we're creating *two* sets, each with their own entropy, but a smaller number of samples in each.

D_{par} = Data found in Parent Node
D_{left} = Data found in Left Node
D_{right} = Data found in Right Node
I() = Impurity function (entropy)
x_i = feature for split