CIS520: Machine Learning

Lyle Ungar

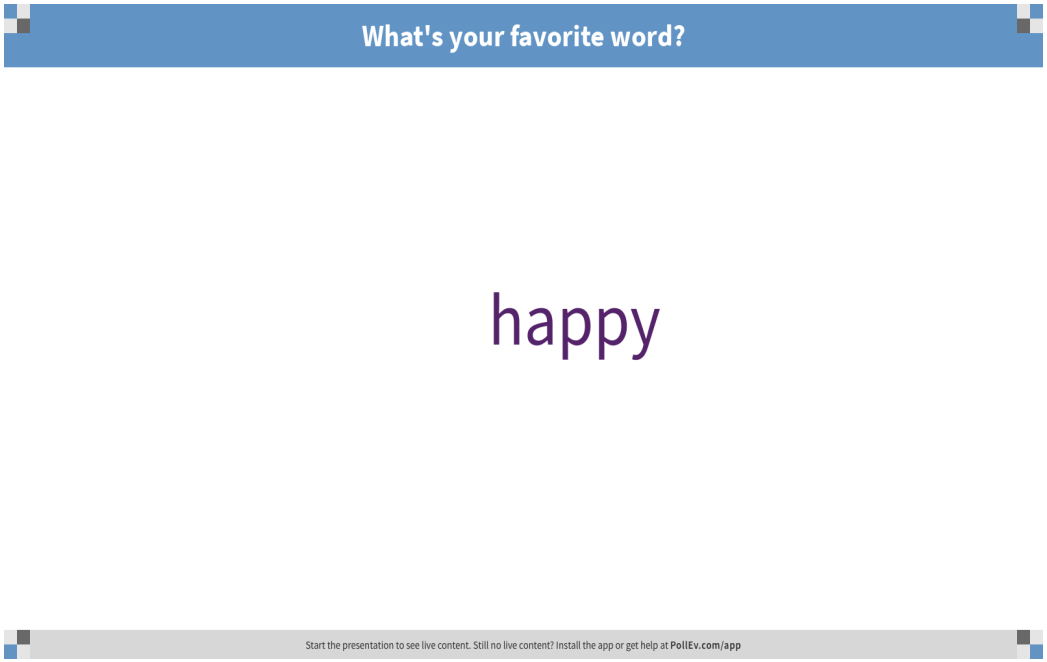
Poll Everywhere

Poll Everywhere, Inc. Communication

Everyone

This app is compatible with all of your devices.

Install *Poll Everywhere* from app store
or go to
https://pollev.com/lyleungar251



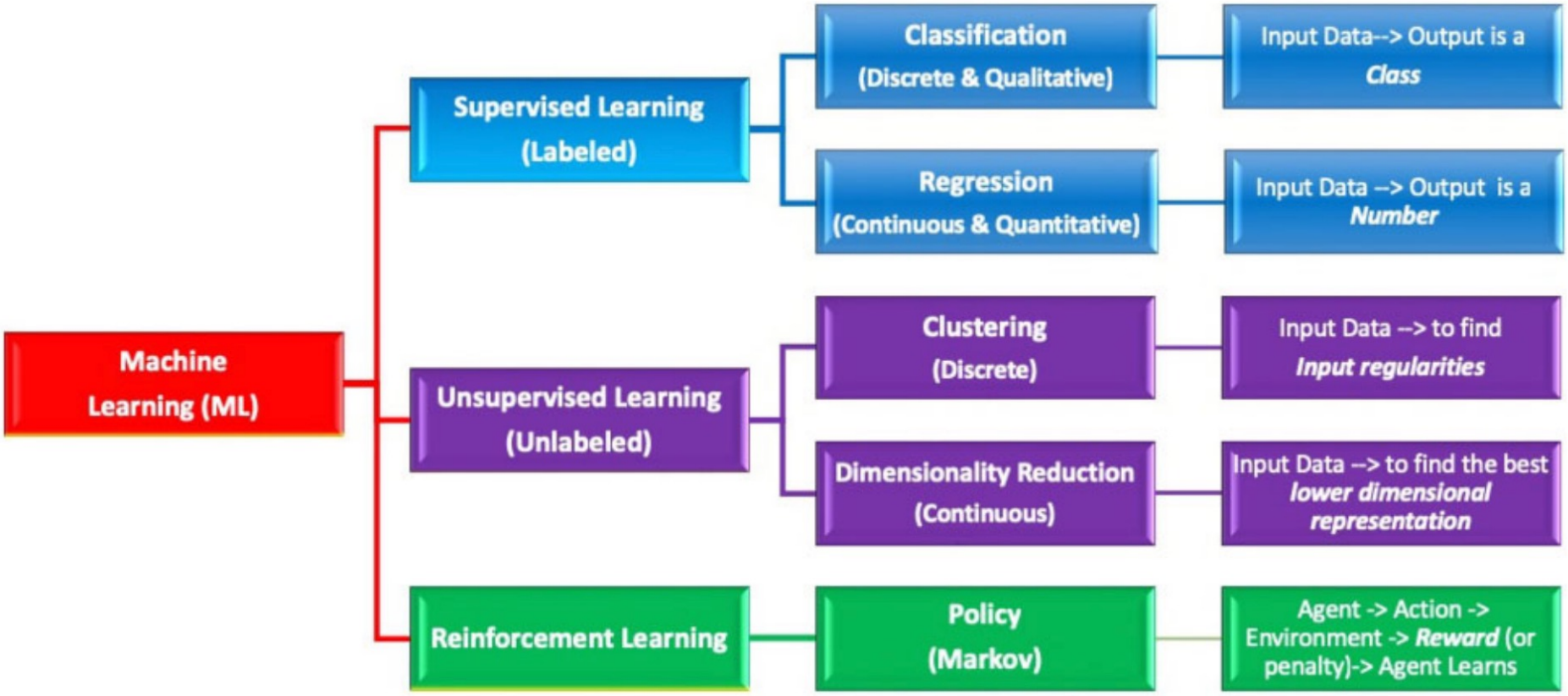
Administrivia

- **Always look at the wiki** lectures page
  - Readings, worksheets, quiz, survey
  - Remember the resources page on the course wiki
- **Ed** – you should be receiving announcements!!!
- **Waitlist - done**
- **HW0, HW1, worksheets…**
- **Office hours**: wiki: “people/office hours”
- **Pods** – sign up!!!

Course Cadence

- **MW:** new material in lectures
- **F:** review material DRLB A2
- **WTh Pods** – in person
  - Meet people; think about material
- **Th-T Worksheets, HW, quizzes**
  - Do everything by midnight Tuesday

Overview of ML



Rashidi, Betts & Green 2019

Nonparametric Learning

Lyle Ungar

K-NN
- Norms, Distance
- Overfitting and Model Complexity

Decision Trees
- Entropy, Information gain

k-Nearest Neighbors (kNN)

- **To predict *y* at a point *x***
  - Find the k nearest neighbors
  - $\hat{y}(\boldsymbol{x})$ = the majority label or
    the average of the *y*'s of those points

https://www.youtube.com/watch?v=PB4qATziTlQ

Norms and Distances

Norms

For all $a \in R$ and all $u, v \in V$,

- $L_p(av) = |a| L_p(v)$
- $L_p(u + v) \leq L_p(u) + L_p(v)$
  - *triangle inequality* or *subadditivity*
- If $L_p(v) = 0$ then v is the zero vector
  - implies $\|v\| = 0$ *iff* v is the zero vector

L_p norm, $\|x\|_p$: $(\Sigma_j |x_j|^p)^{1/p}$

What is

$\|(1,2,3)\|_1$?

A) 1

B) 3

C) sqrt(14)

D) sqrt(14/3)

E) none of the above

What is

$\|(1,2,3)\|_2$?

A) 1

B) 3

C) sqrt(14)

D) sqrt(14/3)

E) none of the above

What is

$\|(1,2,3)\|_{1/2}$?

A) 1

B) 3

C) sqrt(14)

D) sqrt(14/3)

E) none of the above

L_0 pseudo-norm

$||\mathbf{x}||_0$ = number of elements $x_j \neq 0$

How is this not a real norm?

What is

$||(1,2,3)||_0$?

A) 1

B) 3

C) sqrt(14)

D) sqrt(14/3)

E) none of the above

Distance

- Every norm generates a distance

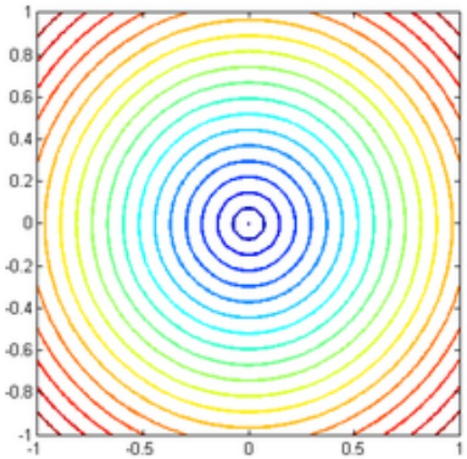
$$d_p(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||_p$$

Distance function (metric)

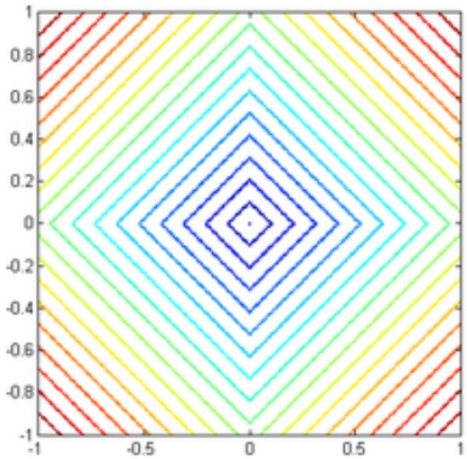
1. $d(x, y) \geq 0$ (*non-negativity*, or separation axiom)
2. $d(x, y) = 0$ if and only if $x = y$ (coincidence axiom)
3. $d(x, y) = d(y, x)$ (*symmetry*)
4. $d(x, z) \leq d(x, y) + d(y, z)$ (*subadditivity* / *triangle inequality*).

https://en.wikipedia.org/wiki/Metric_(mathematics)

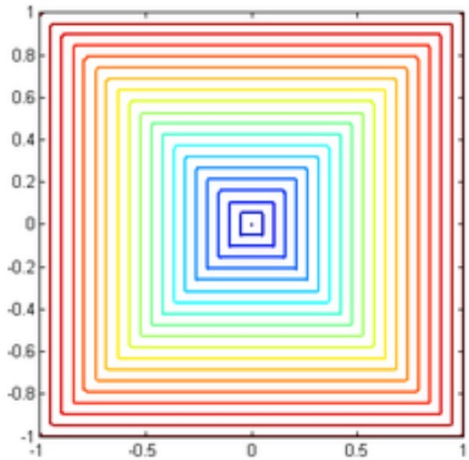
Lines of equal distance from (0,0)



$L_2\,norm$



$L_1\,norm$



$L_{\mathrm{inf}}\,norm$

Convexity

Is $||x||_{1/2}$ convex?

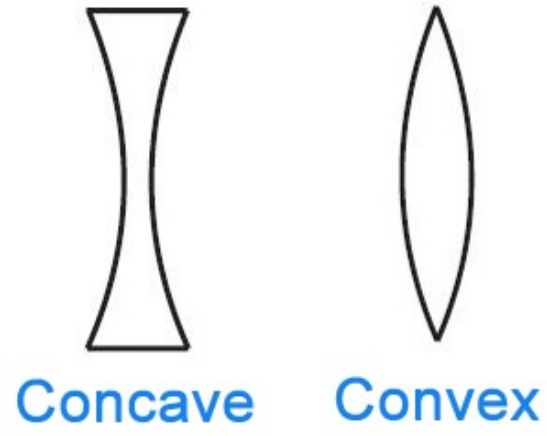
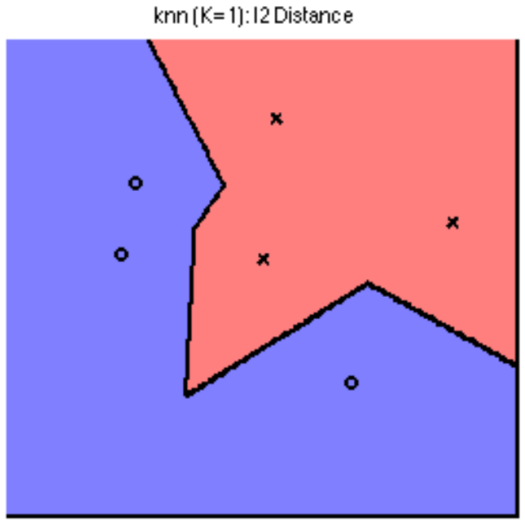


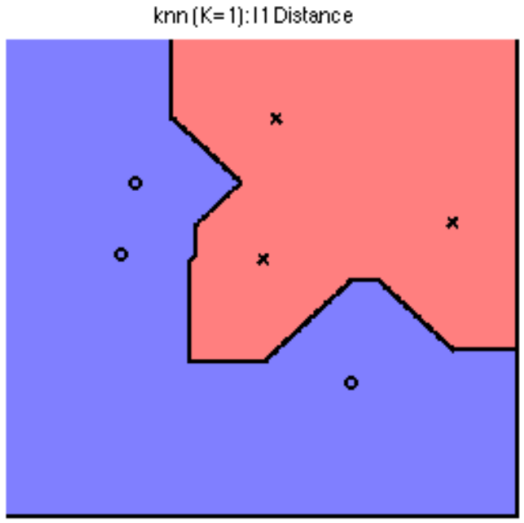
Image credit: https://writingexplained.org/concave-vs-convex-difference

A figure is convex if any line segment connecting two points on the surface of the figure lies entirely inside the figure; otherwise it is concave

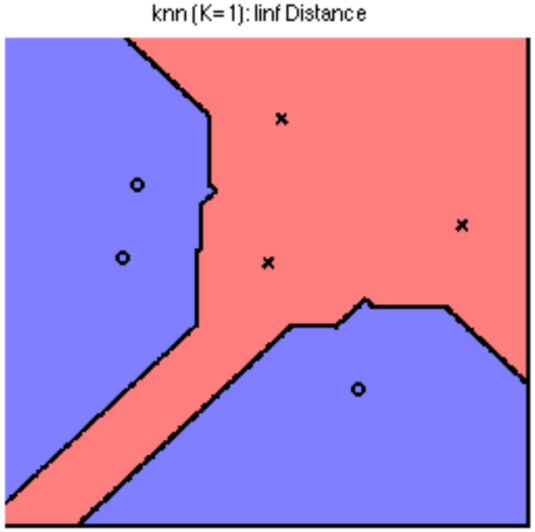
Different norms give different decision boundaries



L_2



L_1



L_{inf}

Components of ML - kNN

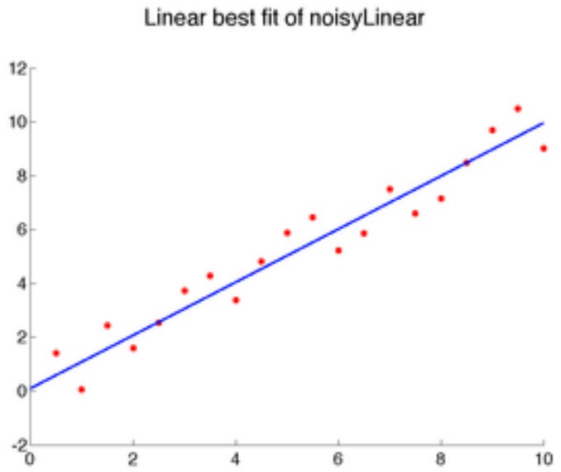
- ***Representation: nonparametric***
  - ~~$\hat{y} = f(\mathbf{x}; \mathbf{w}) = \mathbf{w}^T\mathbf{x}$~~
- ***Loss function (with L2 distance)***
  - $L(\mathbf{y}, \hat{\mathbf{y}}) = ||y - \hat{\mathbf{y}}||_2$
- ***Optimization method: not required***
  - $argmin_w\ L(y, \hat{y}(w))$
  - ~~gradient descent~~

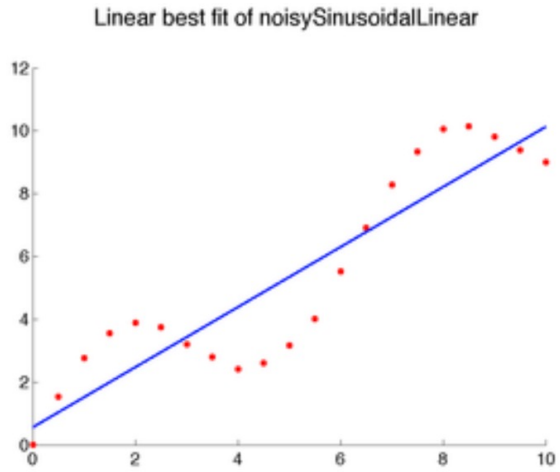
How to pick k?

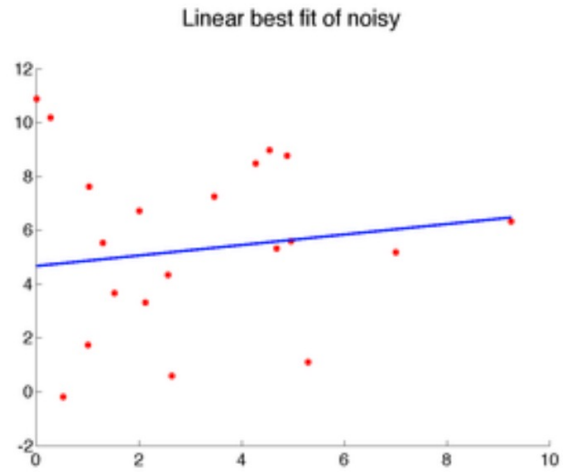
- **What loss function are we trying to minimize?**

  $||y - \hat{y}(x)||_p$

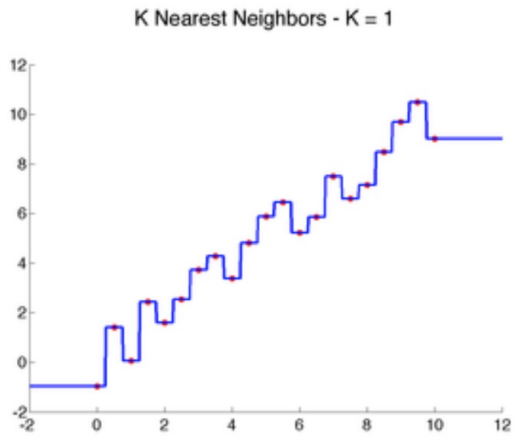
Linear regression on 3 data sets

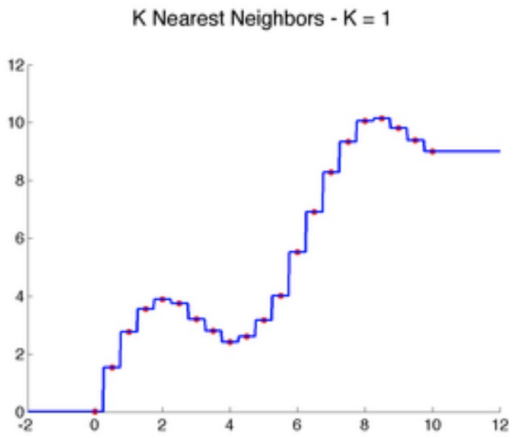


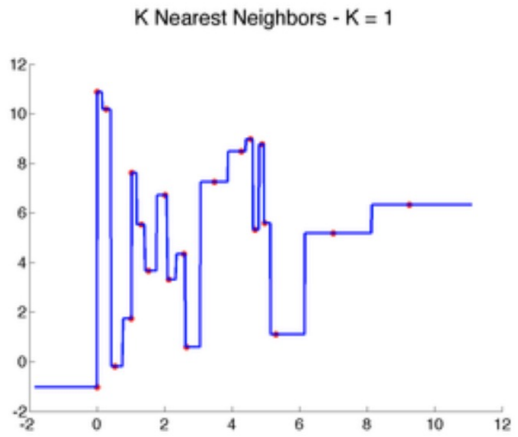




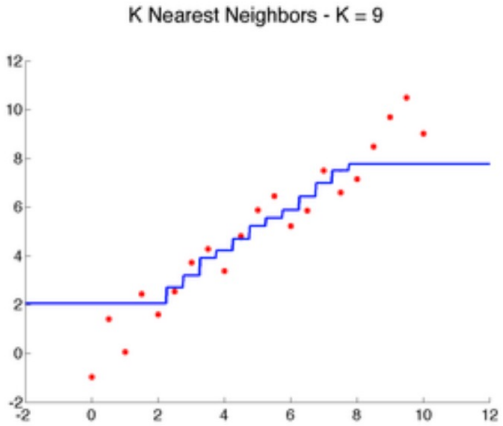
1-NN on 3 data sets (L_1)

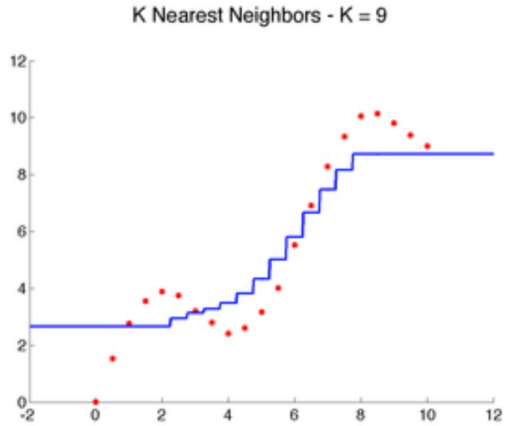


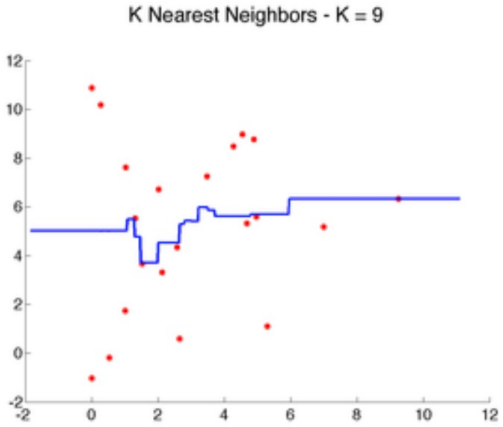




9-NN on 3 data sets (L_1)







In high dimensions most points are equally close to each other.

- **Consider a 100-dimensional cube.**
  - A vertex represented is a “one hot encoding” or “indicator” function, a vector with 99 zeros and one 1.
- **What is the distance between any two vertices?**
  - 0,1,2, more, it varies
- **Generate points at random with half 0’s and half 1’s.**
  - How far away (on average) are two such points?

    Half the coordinates will be the same, so sqrt(50)

Decision Trees
and Information Theory

Lyle Ungar
University of Pennsylvania

Decision Trees

Recursive partition trees, ID3, C4.5, CART, CHAID

- **Example**

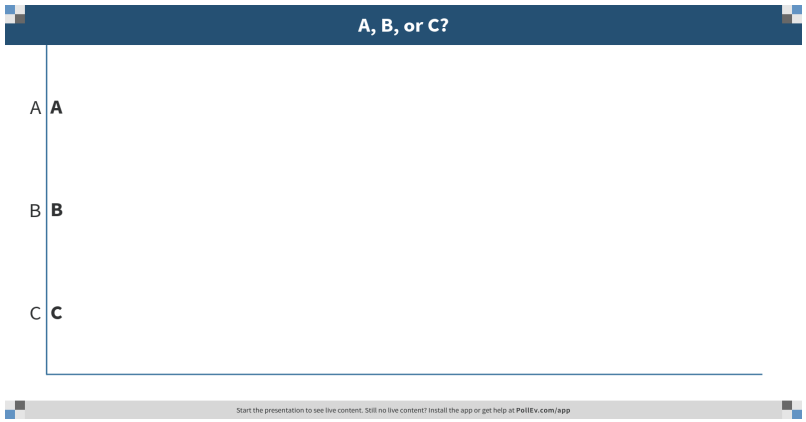
https://www.nytimes.com/interactive/2019/08/08/opinion/sunday/party-polarization-quiz.html

What symptom tells you most about the disease?

S1	S2	S3	D
y	n	n	y
n	y	y	y
n	y	n	n
n	n	n	n
y	y	n	y

A) S1
B) S2
C) S3

Why?



What symptom tells you most about the disease?

S1/D

	y	n
y	2	0
n	1	2

S2/D

	y	n
y	2	1
n	1	1

s3/D

	y	n
y	1	0
n	2	2

A) S1
B) S2
C) S3

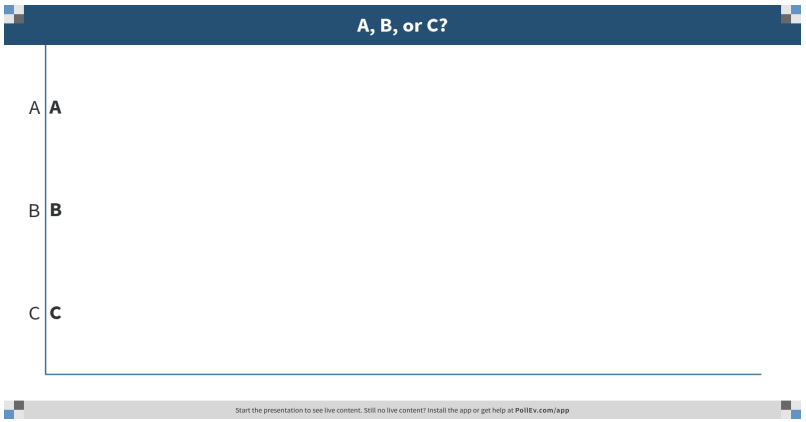
Why?

If you know S1=n, what symptom tells you most about the disease?

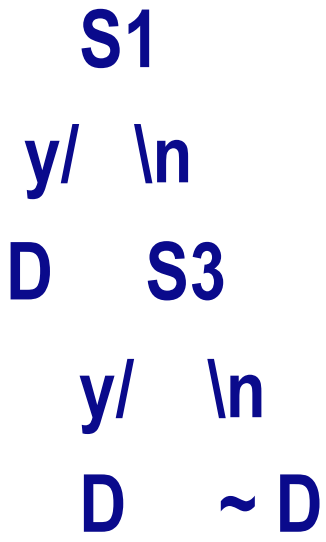
S1	S2	S3	D
y	n	n	y
n	y	y	y
n	y	n	n
n	n	n	n
y	y	n	y

A) S1
B) S2
C) S3

Why?



Resulting decision tree



The key question: what criterion to use do decide which question to ask?

Entropy and Information Gain

Andrew W. Moore

Carnegie Mellon University

www.cs.cmu.edu/~awm

awm@cs.cmu.edu

412-268-7599

modified by Lyle Ungar

Bits

You observe a set of independent random samples of X

You see that X has four possible values

P(X=A) = 1/4	P(X=B) = 1/4	P(X=C) = 1/4	P(X=D) = 1/4

So you might see: BAACBADCDADDDA…

You transmit data over a binary serial link. You can encode each reading with two bits (e.g. A = 00, B = 01, C = 10, D = 11)

0100001001001110110011111100…

Fewer Bits

Someone tells you that the probabilities are not equal

P(X=A) = 1/2	P(X=B) = 1/4	P(X=C) = 1/8	P(X=D) = 1/8

It is possible to invent a coding for your transmission that only uses 1.75 bits on average per symbol. How?

A	0
B	10
C	110
D	111

(This is just one of several ways)

Fewer Bits

Suppose there are three equally likely values…

P(X=A) = 1/3	P(X=B) = 1/3	P(X=C) = 1/3

Here's a naïve coding, costing 2 bits per symbol

A	00
B	01
C	10

Can you think of a coding that only needs 1.6 bits per symbol on average?

In theory, it can in fact be done with 1.58496 bits per symbol.

General Case: Entropy

Suppose X can have one of *m* values… $V_1, V_2, \ldots V_m$

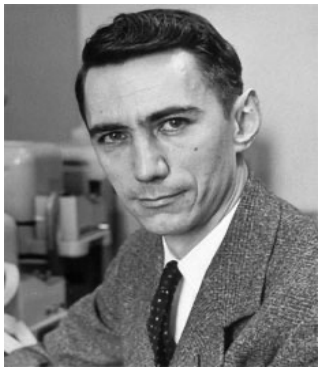
$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	….	$P(X=V_m) = p_m$

What's the smallest possible number of bits, on average, per symbol, needed to transmit a stream of symbols drawn from X's distribution?

It is

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$



H(X) = The entropy of X

- **"High Entropy" means X is from a uniform (boring) distribution**
- **"Low Entropy" means X is from varied (peaks and valleys) distribution**

General Case

Suppose X can have one of *m* values… $V_1, V_2, \ldots V_m$

$P(X=V_1) = p_1$	$P(X=V_2) = p_2$	….	$P(X=V_m) = p_m$

What's the smallest possible number of bits needed to transmit a stream of symbols drawn from X's distribution? It's

$$H(X) = -\sum_{j=1}^{m} p_j \log_2 p_j$$

H(X) = The entropy of X

- **"High Entropy" means X is from a uniform (boring) distribution**
- **"Low Entropy" means X is from varied (peaks and valleys) distribution**

A histogram of the frequency distribution of values of X would be flat

..and so the values sampled from it would be all over the place

A histogram of the frequency distribution of values of X would have many lows and one or two highs

..and so the values sampled from it would be more predictable

Entropy in a nut-shell

Low Entropy

High Entropy

Entropy in a nut-shell

Low Entropy

..the values (locations of soup) sampled entirely from within the soup bowl

High Entropy

..the values (locations of soup) unpredictable... almost uniformly sampled throughout our dining room

Why does entropy have this form?

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

Entropy is the expected value of the information content (surprise) of the message $log_2 p_j$

If an event is certain, the entropy is
A) 0
B) between 0 and ½
C) ½
D) between ½ and 1
E) 1

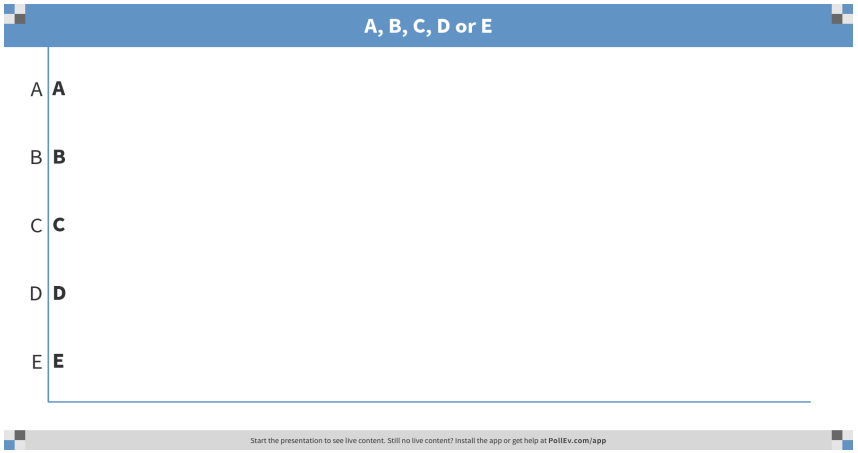
Why does entropy have this form?

$$H(X) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 - \ldots - p_m \log_2 p_m$$

$$= -\sum_{j=1}^{m} p_j \log_2 p_j$$

If two events are equally likely, the entropy is

A) 0
B) between 0 and ½
C) ½
D) between ½ and 1
E) 1



Specific Conditional Entropy H(Y|X=v)

Suppose I'm trying to predict output Y and I have input X

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Assume this reflects the true probabilities

e.g. From this data we estimate

- *P(LikeG = Yes) = 0.5*
- *P(Major = Math & LikeG = No) = 0.25*
- *P(Major = Math) = 0.5*
- *P(LikeG = Yes | Major = History) = 0*

Note:

- *H(X) = 1.5*
- *H(Y) = 1*

Specific Conditional Entropy H(Y|X=v)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Specific Conditional Entropy:

H(Y |X=v) = **The entropy of *Y* among only those records in which *X* has value *v***

Example:

- *H(Y|X=Math) = 1*
- *H(Y|X=History) = 0*
- *H(Y|X=CS) = 0*

Conditional Entropy H(Y|X)

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Conditional Entropy:

$H(Y|X)$ = The average specific conditional entropy of Y

If you choose a record at random what will be the conditional entropy of Y, conditioned on that row's value of X

= Expected number of bits to transmit Y if both sides will know the value of X

$= \Sigma_j Prob(X=v_j)\, H(Y \mid X = v_j)$

Conditional Entropy

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Conditional Entropy:

$H(Y|X)$ = The average conditional entropy of Y

$= \Sigma_j Prob(X=v_j)\ H(Y \mid X = v_j)$

Example:

v_j	$Prob(X=v_j)$	$H(Y \mid X = v_j)$
Math	0.5	1
History	0.25	0
CS	0.25	0

$H(Y|X) = 0.5 * 1 + 0.25 * 0 + 0.25 * 0 = 0.5$

Information Gain

X = College Major

Y = Likes "Gladiator"

X	Y
Math	Yes
History	No
CS	Yes
Math	No
Math	No
CS	Yes
History	No
Math	Yes

Definition of Information Gain:

IG(Y|X) = I must transmit *Y*. How many bits on average would it save me if both ends of the line knew *X*?

$IG(Y|X) = H(Y) - H(Y|X)$

Example:

- **H(Y) = 1**
- **H(Y|X) = 0.5**
- **Thus IG(Y|X) = 1 – 0.5 = 0.5**

Information Gain Example

wealth values: poor rich

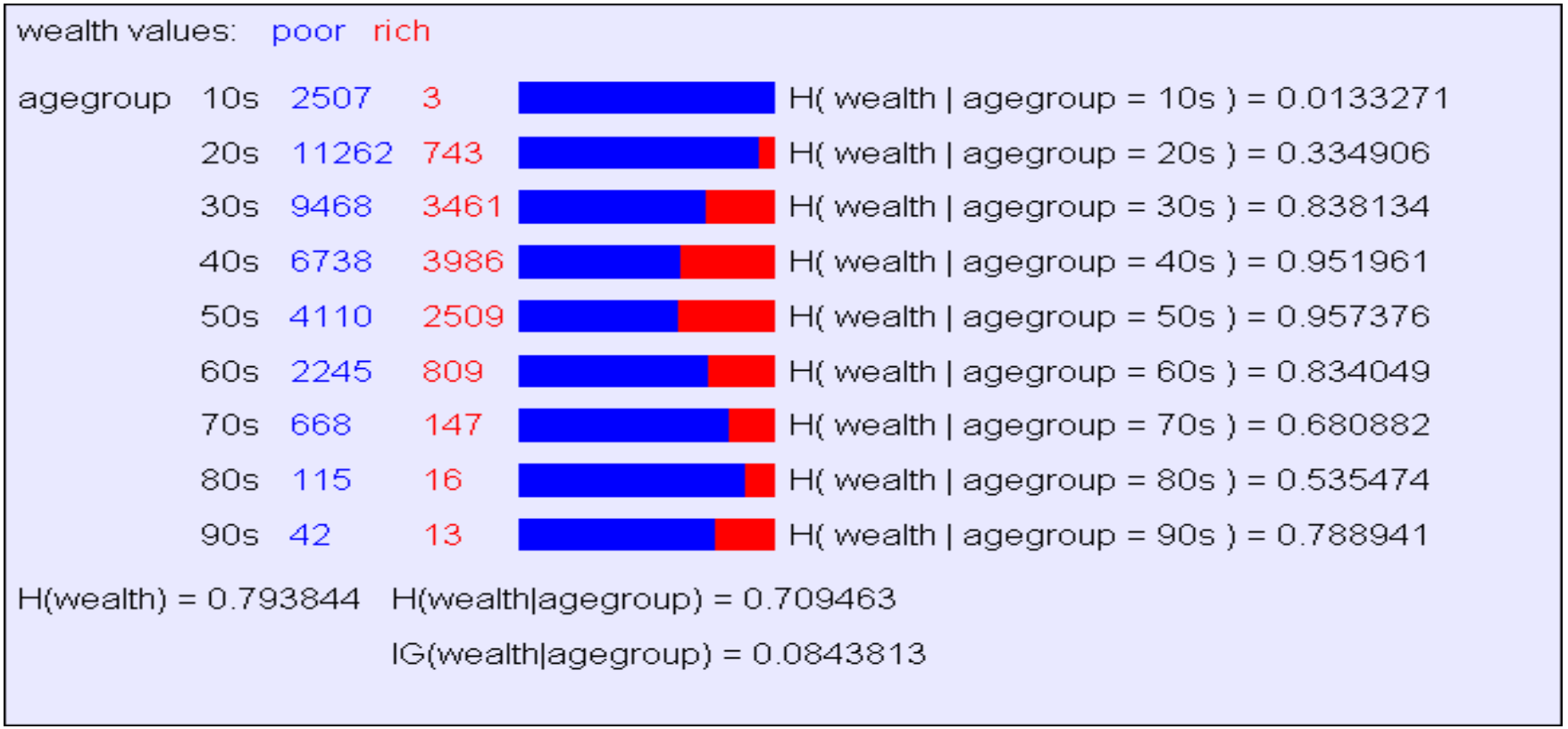
gender Female 14423 1769  H(wealth | gender = Female) = 0.497654

Male 22732 9918  H(wealth | gender = Male) = 0.885847

H(wealth) = 0.793844 H(wealth|gender) = 0.757154

IG(wealth|gender) = 0.0366896

Another example



What is Information Gain used for?

If you are going to collect information from someone (e.g. asking questions sequentially in a decision tree), the “best” question is the one with the highest information gain.

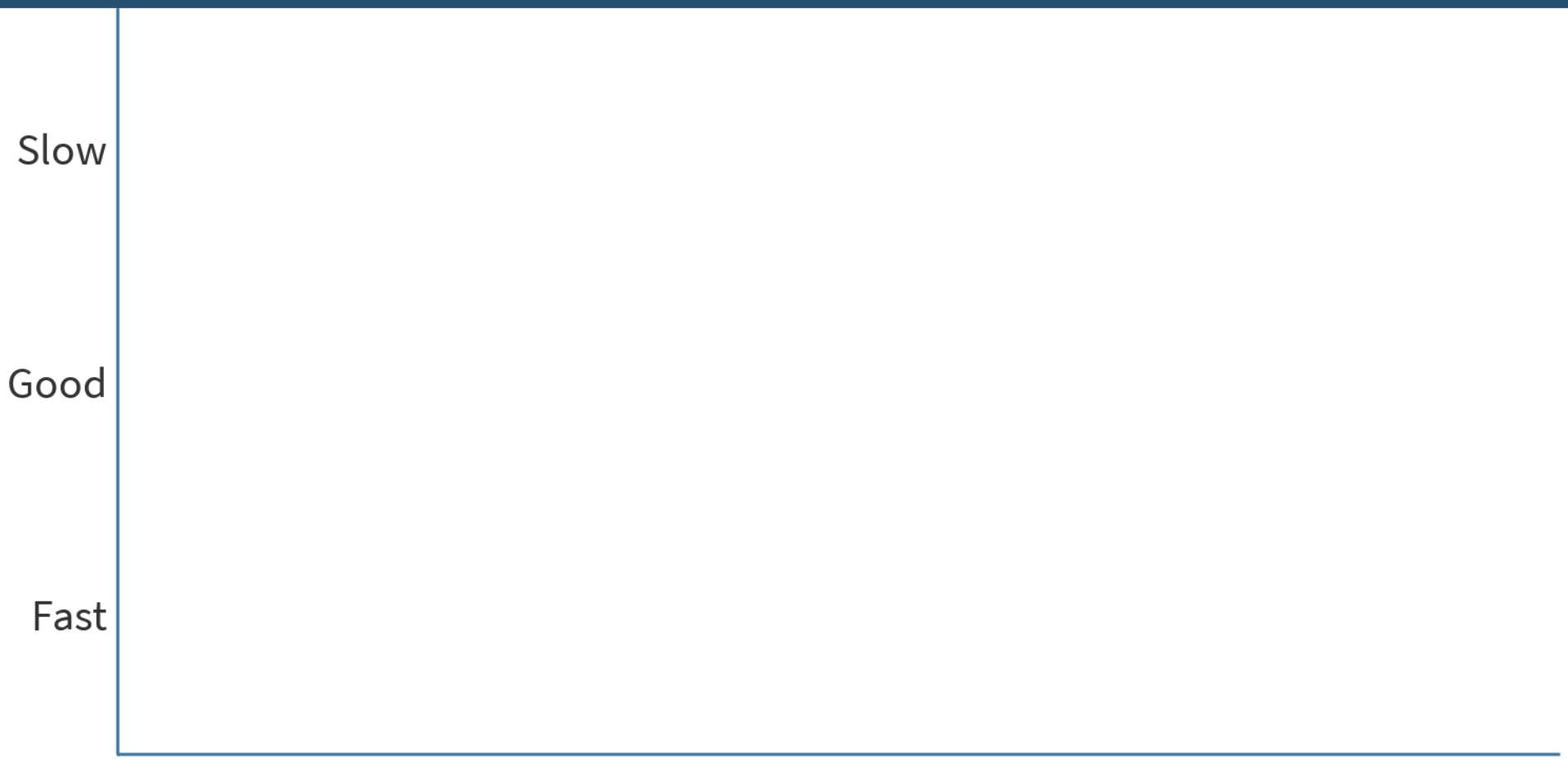
Information gain is useful for model selection

What question did we not ask (or answer) about decision trees?

What you should know

- **K-NN**
  - hyperparameter *k* controls model complexity
- **Norm, distance**
- **Convexity**
- **Entropy, information gain**
- **The standard decision tree algorithm**
  - Recursive partition to maximize information gain

How is my speed?



What questions do you have on today's class?

Top