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

- Remember the resources page on the course wiki
  - And always look at the lectures page for readings and worksheets
- Piazza!!!
- Waitlist - done
- HW0
- Office hours are happening –wiki: “people/office hours”
- Social hours on gather.town –after class, and new Asian hours.
Nonparametric Learning
Lyle Ungar
Computer and information Science

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
  - $y^{est}(x) = \text{the majority label or average of the } y\text{'s of those points}$

http://videolectures.net/aaai07_bosch_knnnc/
Norms and Distances
Norms

For all \( a \in \mathbb{R} \) and all \( u, v \in V \),

- \( L_p(a v) = |a| L_p(v) \)
- \( L_p(u + v) \leq L_p(u) + L_p(v) \)
  - 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, \( \|x\|_p : (\sum_j |x_j|^p)^{1/p} \)
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

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

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

\[ d_p(x,y) = \|x-y\|_p \]
Distance function (metric)

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

https://en.wikipedia.org/wiki/Metric_(mathematics)
Lines of equal distance from (0,0)

$L_2\text{norm}$

$L_1\text{norm}$

$L_{\infty}\text{norm}$
Convexity

Is $\|x\|^{1/2}$ convex?

Concave  Convex

Image credit: https://writingexplained.org/concave-vs-convex-difference
Different norms give different decision boundaries

- $L_2$
- $L_1$
- $L_{\infty}$
Components of ML - K-NN

- **Representation: nonparametric**
  - \( \hat{y} = f(x; w) = w^T x \)

- **Loss function**
  - \( L(y, \hat{y}) = \|y - \hat{y}\|_2 \)

- **Optimization method: not required**
  - \( \text{argmin}_w L(y, \hat{y}(w)) \)
  - 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

What symptom tells you most about the disease?

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>2</td>
<td>n</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>3</td>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>4</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>5</td>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>

A) S1  
B) S2  
C) S3

Why?
What symptom tells you most about the disease?

<table>
<thead>
<tr>
<th>S1/D</th>
<th>S2/D</th>
<th>s3/D</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

A) S1  
B) S2  
C) S3

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

<table>
<thead>
<tr>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>n</td>
<td>n</td>
<td>y</td>
</tr>
<tr>
<td>n</td>
<td>y</td>
<td>y</td>
<td>y</td>
</tr>
<tr>
<td>n</td>
<td>y</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>y</td>
<td>y</td>
<td>n</td>
<td>y</td>
</tr>
</tbody>
</table>

A) S1
B) S2
C) S3

Why?
Resulting decision tree

```
S1
  y/  \nD   S3
  y/  \nD   ~ D
```

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

<table>
<thead>
<tr>
<th>$P(X=A)$</th>
<th>$P(X=B)$</th>
<th>$P(X=C)$</th>
<th>$P(X=D)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
<td>1/4</td>
</tr>
</tbody>
</table>

**So you might see:** BAACBADCDDDDDA...

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>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P(X=A) = 1/2</td>
<td>P(X=B) = 1/4</td>
<td>P(X=C) = 1/8</td>
<td>P(X=D) = 1/8</td>
</tr>
</tbody>
</table>

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)

Copyright © 2001, 2003, Andrew W. Moore
Fewer Bits

Suppose there are three equally likely values...

<table>
<thead>
<tr>
<th></th>
<th>P(X=A) = 1/3</th>
<th>P(X=B) = 1/3</th>
<th>P(X=C) = 1/3</th>
</tr>
</thead>
</table>

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

<table>
<thead>
<tr>
<th>A</th>
<th>00</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>01</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
</tr>
</tbody>
</table>

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... $V_1, V_2, \ldots V_m$

| $P(X=V_1) = p_1$ | $P(X=V_2) = p_2$ | \(\ldots\) | $P(X=V_m) = p_m$ |

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

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

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

$H(X) = \text{The entropy of } X$

- “High Entropy” means $X$ is from a uniform (boring) distribution
- “Low Entropy” means $X$ is from varied (peaks and valleys) distribution

Copyright © 2001, 2003, Andrew W. Moore
Suppose $X$ can have one of $m$ values... $V_1, V_2, ... V_m$

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

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

It’s

$$H(X) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$$

$H(X) = \text{The entropy of } X$

- “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?

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

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 \(\frac{1}{2}\)
C) \(\frac{1}{2}\)
D) between \(\frac{1}{2}\) 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”

Assume this reflects the true probabilities

e.g. From this data we estimate

- $P(\text{LikeG} = \text{Yes}) = 0.5$
- $P(\text{Major} = \text{Math} \& \text{LikeG} = \text{No}) = 0.25$
- $P(\text{Major} = \text{Math}) = 0.5$
- $P(\text{LikeG} = \text{Yes} \mid \text{Major} = \text{History}) = 0$

Note:

- $H(X) = 1.5$
- $H(Y) = 1$
Specific Conditional Entropy $H(Y|X=v)$

**Definition of Specific Conditional Entropy:**

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

**Example:**

- $H(Y|X=\text{Math}) = 1$
- $H(Y|X=\text{History}) = 0$
- $H(Y|X=\text{CS}) = 0$

$X =$ College Major

$Y =$ Likes “Gladiator”

<table>
<thead>
<tr>
<th>$X$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Conditional Entropy $H(Y|X)$

$X = \text{College Major}$

$Y = \text{Likes “Gladiator”}$

**Definition of Conditional Entropy:**

$H(Y|X) = \text{The average specific conditional entropy of } Y$

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

$= \text{Expected number of bits to transmit } Y \text{ if both sides will know the value of } X$

$= \sum_j \text{Prob}(X=v_j) \ H(Y \mid X = v_j)$
**Conditional Entropy**

- **X** = College Major
- **Y** = Likes “Gladiator”

**Definition of Conditional Entropy:**

\[ H(Y|X) = \text{The average conditional entropy of } Y \]

\[ = \sum_j \text{Prob}(X=v_j) \cdot H(Y | X = v_j) \]

**Example:**

| \( v_j \) | Prob(\( X=v_j \)) | \( H(Y | X = v_j) \) |
|-----------|--------------------|---------------------|
| Math      | 0.5                | 1                   |
| History   | 0.25               | 0                   |
| CS        | 0.25               | 0                   |
| Math      | No                 |                     |
| History   | Yes                |                     |
| CS        | Yes                |                     |
| History   | No                 |                     |
| Math      | Yes                |                     |

\[ H(Y|X) = 0.5 \cdot 1 + 0.25 \cdot 0 + 0.25 \cdot 0 = 0.5 \]
Information Gain

**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 \[ IG(Y|X) = 1 - 0.5 = 0.5 \]

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>No</td>
</tr>
<tr>
<td>CS</td>
<td>Yes</td>
</tr>
<tr>
<td>History</td>
<td>No</td>
</tr>
<tr>
<td>Math</td>
<td>Yes</td>
</tr>
</tbody>
</table>
# Information Gain Example

<table>
<thead>
<tr>
<th>wealth values:</th>
<th>poor</th>
<th>rich</th>
</tr>
</thead>
</table>

| gender | Female | 14423 | 1769 | \( H(\text{wealth} | \text{gender} = \text{Female} ) \) = 0.497654 |
|--------|--------|-------|------|------------------------------------------------|
|        | Male   | 22732 | 9918 | \( H(\text{wealth} | \text{gender} = \text{Male} ) \) = 0.885847    |

\[
H(\text{wealth}) = 0.793844 \quad H(\text{wealth}|\text{gender}) = 0.757154
\]

\[
IG(\text{wealth}|\text{gender}) = 0.0366896
\]
Another example

<table>
<thead>
<tr>
<th>agegroup</th>
<th>10s</th>
<th>20s</th>
<th>30s</th>
<th>40s</th>
<th>50s</th>
<th>60s</th>
<th>70s</th>
<th>80s</th>
<th>90s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2507</td>
<td>11262</td>
<td>9468</td>
<td>6738</td>
<td>4110</td>
<td>2245</td>
<td>668</td>
<td>115</td>
<td>42</td>
</tr>
<tr>
<td>wealth values:</td>
<td>poor</td>
<td>rich</td>
<td>H(wealth</td>
<td>agegroup = 10s) = 0.0133271</td>
<td>H(wealth</td>
<td>agegroup = 20s) = 0.334906</td>
<td>H(wealth</td>
<td>agegroup = 30s) = 0.838134</td>
<td>H(wealth</td>
</tr>
</tbody>
</table>

H(wealth) = 0.793844  H(wealth|agegroup) = 0.709463  IG(wealth|agegroup) = 0.0843813
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 based to maximize information gain
What questions do you have on today's class?
Next up

◆ Office hours
  ● 12:00-1:00 Lyle Ungar
    https://upenn.zoom.us/j/92527713740
  ● And lots more

◆ Meet some other students
  ● https://gather.town/aQMGI0l1R8DP0Ovv/penn-cis
  ● Firefox or Chrome; not mobile