Classification & Information Theory

Lecture #8

Introduction to Natural Language Processing

CMPSCI 585, Fall 2007

University of Massachusetts Amherst

Andrew McCallum

Today's Main Points

- Automatically categorizing text
  - Parameter estimation and smoothing
  - a general recipe for a statistical CompLing model
  - Building a Spam Filter

- Information Theory
  - What is information? How can you measure it?
  - Entropy, Cross Entropy, Information gain

Maximum Likelihood Parameter Estimation Example: Binomial

- Toss a coin 100 times, observe *r* heads
- Assume a binomial distribution
  - Order doesn't matter, successive flips are independent
  - One parameter is *q* (probability of flipping a head)
  - Binomial gives $p(r|n,q)$. We know r and n.
  - Find $\arg\max_q p(r|n, q)$

Maximum Likelihood Parameter Estimation Example: Binomial

- Toss a coin 100 times, observe *r* heads
- Assume a binomial distribution
  - Order doesn't matter, successive flips are independent
  - One parameter is *q* (probability of flipping a head)
  - Binomial gives p(r|n,q). We know r and n.
  - Find arg $\max_q$ p(r|n, q)

(Notes for board)

$$\text{likelihood} = p(R = r \mid n,q) = \binom{n}{r} q^r (1-q)^{n-r}$$

$$\text{log-likelihood} = L = \log(p(r \mid n,q)) \propto \log(q^r (1-q)^{n-r}) = r\log(q) + (n-r)\log(1-q)$$

$$\frac{\partial L}{\partial q} = \frac{r}{q} - \frac{n-r}{1-q} \Rightarrow r(1-q) = (n-r)q \Rightarrow q = \frac{r}{n}$$

Our familiar ratio-of-counts
is the maximum likelihood estimate!

Binomial Parameter Estimation Examples

- Make 1000 coin flips, observe 300 Heads
  - P(Heads) = 300/1000
- Make 3 coin flips, observe 2 Heads
  - P(Heads) = 2/3 ??
- Make 1 coin flips, observe 1 Tail
  - P(Heads) = 0 ???
- Make 0 coin flips
  - P(Heads) = ???

- We have some "*prior*" belief about P(Heads) before we see any data.
- After seeing some data, we have a "*posterior*" belief.

Maximum A Posteriori Parameter Estimation

- We've been finding the parameters that maximize
  – p(data|parameters),

  not the parameters that maximize
  – p(parameters|data) **(parameters are random variables!)**
- $p(q|n,r) = \dfrac{p(r|n,q)\ p(q|n)}{p(r|n)} = \dfrac{p(r|n,q)\ p(q)}{\text{constant}}$
- And let $p(q) = 2\ q(1-q)$

Maximum A Posteriori Parameter Estimation Example: Binomial

$$\text{posterior} = p(r \mid n,q)\,p(q) = \binom{n}{r} q^{r}(1-q)^{n-r}(2q(1-q))$$

$$\log-\text{posterior} = L \propto \log(q^{r+1}(1-q)^{n-r+1}) = (r+1)\log(q) + (n-r+1)\log(1-q)$$

$$\frac{\partial L}{\partial q} = \frac{(r+1)}{q} - \frac{(n-r+1)}{1-q} \Rightarrow (r+1)(1-q) = (n-r+1)q \Rightarrow q = \frac{r+1}{n+2}$$

Bayesian Decision Theory

- We can use such techniques for choosing among models:
  - Which among several models best explains the data?
- Likelihood Ratio

$$\frac{P(\text{model1} \mid \text{data})}{P(\text{model2} \mid \text{data})} = \frac{P(\text{data}|\text{model1})\, P(\text{model1})}{P(\text{data}|\text{model2})\, P(\text{model2})}$$

...back to our example: French vs English

- p(French | glacier, melange) versus
  p(English | glacier, melange) ?

- We have real data for
  - Jane Austin
  - William Shakespeare

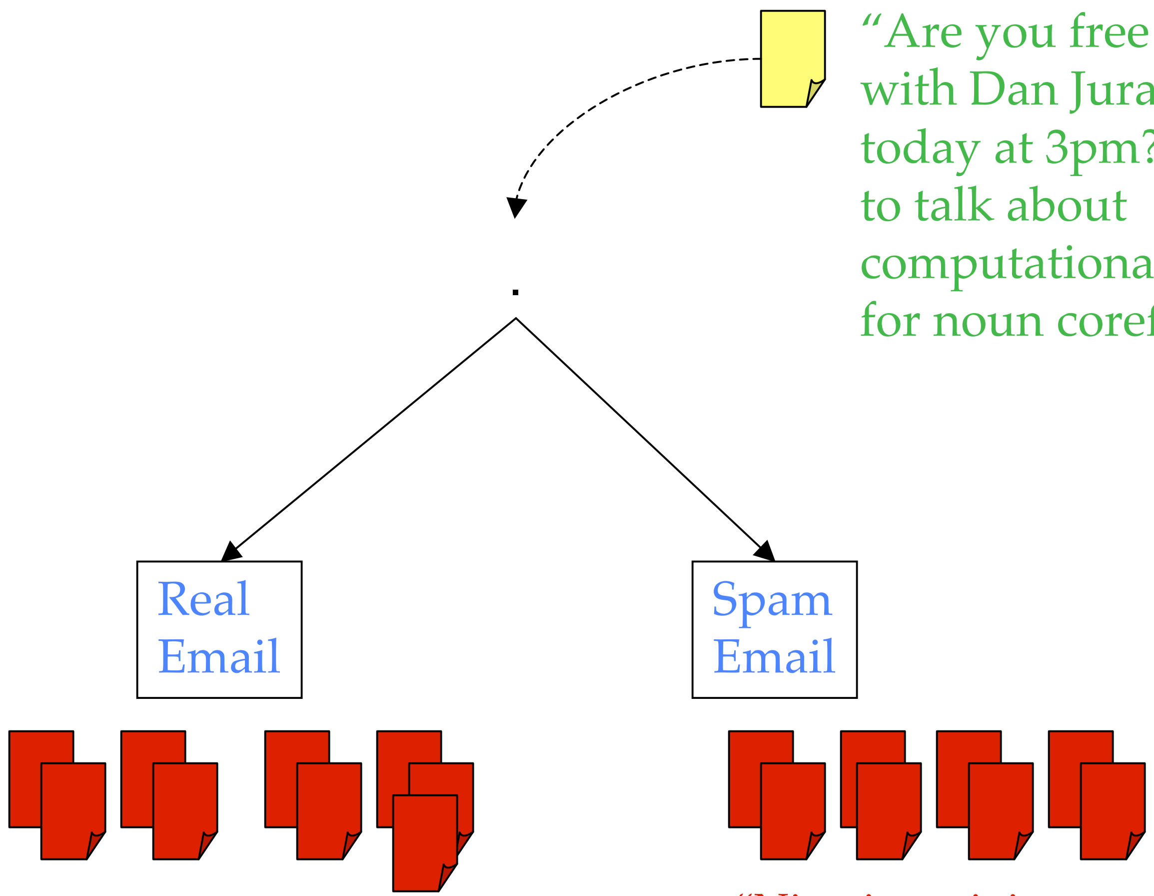
- p(Austin | “stars”, “thou”)
  p(Shakespeare | “stars”, “thou”)

Statistical Spam Filtering

Testing Document:

“Are you free to meet with Dan Jurafsky today at 3pm? He wants to talk about computational methods for noun coreference.”

Categories:

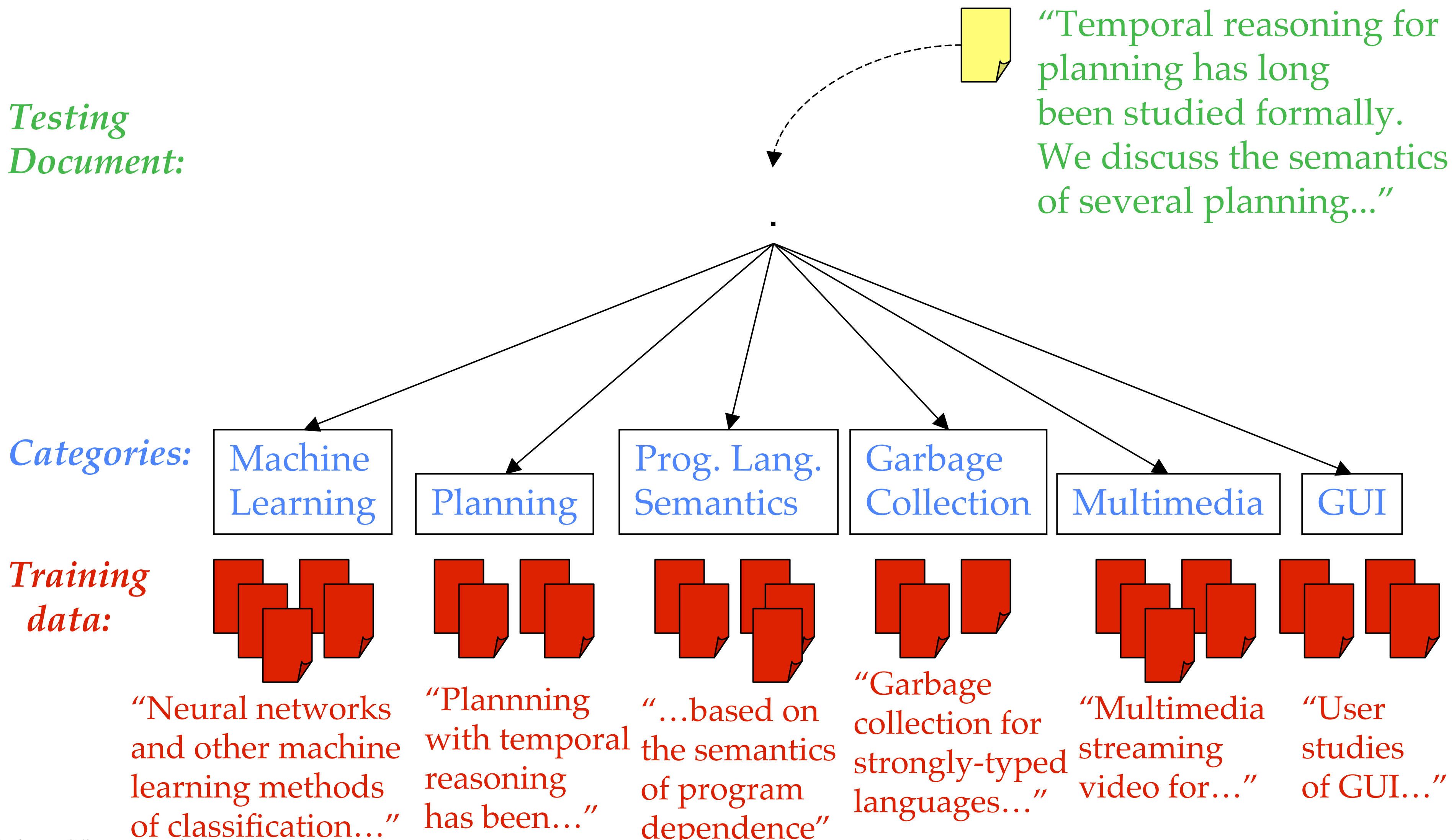


Training data:

“Speaking at awards ceremony...”
“Coming home for dinner...”
“Free for a research meeting at 6pm...”
“Computational Linguistics office hours...”

“Nigerian minister awards…”
“Earn money at home today!...”
“FREE CASH”
“Just hours per day...”

Document Classification by Machine Learning



Work out Naïve Bayes formulation interactively on the board

Recipe for Solving a NLP Task Statistically

1) **Data**: Notation, representation
2) **Problem**: Write down the problem in notation
3) **Model**: Make some assumptions, define a parametric model
4) **Inference**: How to search through possible answers to find the best one
5) **Learning**: How to estimate parameters
6) **Implementation**: Engineering considerations for an efficient implementation

(Engineering) Components of a Naïve Bayes Document Classifier

- Split documents into training and testing
- Cycle through all documents in each class
- Tokenize the character stream into words
- Count occurrences of each word in each class
- Estimate $P(w|c)$ by a ratio of counts (+1 prior)
- For each test document, calculate $P(c|d)$ for each class
- Record predicted (and true) class, and keep accuracy statistics

A Probabilistic Approach to Classification: "Naïve Bayes"

Pick the most probable class, given the evidence:

$$c^* = \arg\max_{c_j} \Pr(c_j \mid d)$$

c_j - a class (like "Planning")

d - a document (like "language intelligence proof...")

Bayes Rule:

"Naïve Bayes":

$$\Pr(c_j \mid d) = \frac{\Pr(c_j)\Pr(d \mid c_j)}{\Pr(d)} \approx \frac{\Pr(c_j)\prod_{i=1}^{|d|}\Pr(w_{d_i} \mid c_j)}{\sum_{c_k}\Pr(c_k)\prod_{i=1}^{|d|}\Pr(w_{d_i} \mid c_k)}$$

w_{d_i} - the i th word in d (like "proof")

Parameter Estimation in Naïve Bayes

Estimate of P(c)

$$P(c_j) = \frac{1 + \mathrm{Count}(d \in c_j)}{|C| + \sum_k \mathrm{Count}(d \in c_k)}$$

Estimate of P(w|c)

$$P(w_i | c_j) = \frac{1 + \sum_{d_k \in c_j} \mathrm{Count}(w_i, d_k)}{|V| + \sum_{t=1}^{|V|} \sum_{d_k \in c_j} \mathrm{Count}(w_t, d_k)}$$

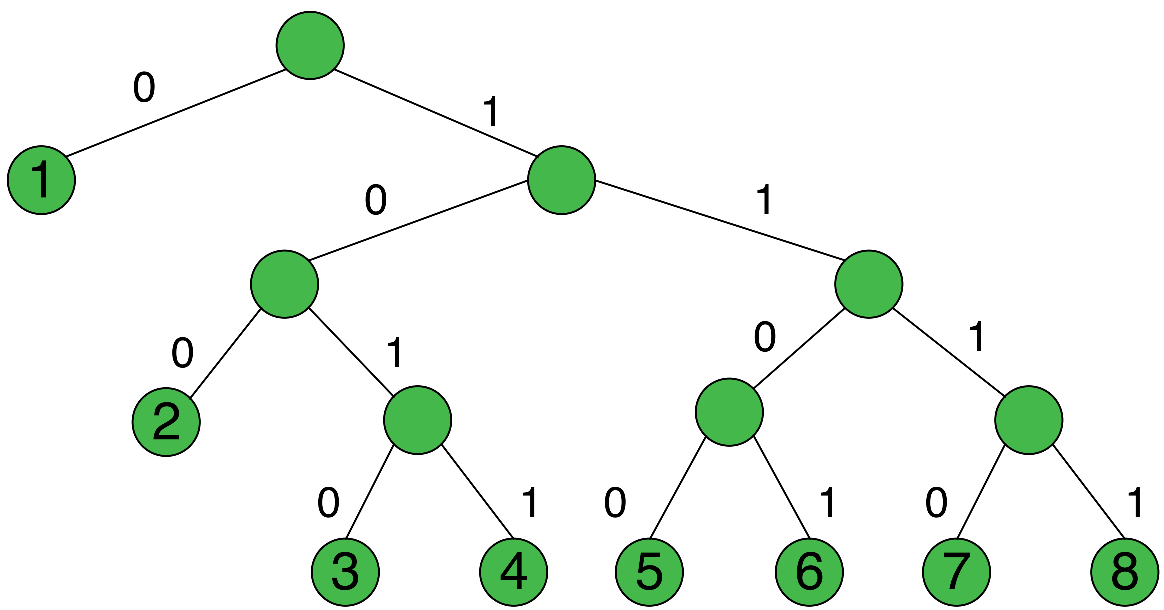
Information Theory

What is Information?

- “The sun will come up tomorrow.”
- “Condi Rice was shot and killed this morning.”

Efficient Encoding

- I have a 8-sided die.
  How many bits do I need to tell you what face I just rolled?
- My 8-sided die is unfair
  - P(1)=1/2, P(2)=1/8, P(3)=…=P(8)=1/16



Entropy (of a Random Variable)

- Average length of message needed to transmit the outcome of the random variable.

- First used in:
  - Data compression
  - Transmission rates over noisy channel

"Coding" Interpretation of Entropy

- Given some distribution over events P(X)…
- What is the average number of bits needed to encode a message (a event, string, sequence)
- = Entropy of P(X):

$$H(p(X)) = -\sum_{x \in X} p(x) \log_2(p(x))$$

- Notation: $H(X) = H_p(X) = H(p) = H_X(p) = H(p_X)$

What is the entropy of a fair coin? A fair 32-sided die?
What is the entropy of an unfair coin that always comes up heads?
What is the entropy of an unfair 6-sided die that always {1,2}
Upper and lower bound? (Prove lower bound?)

Entropy and Expectation

- Recall

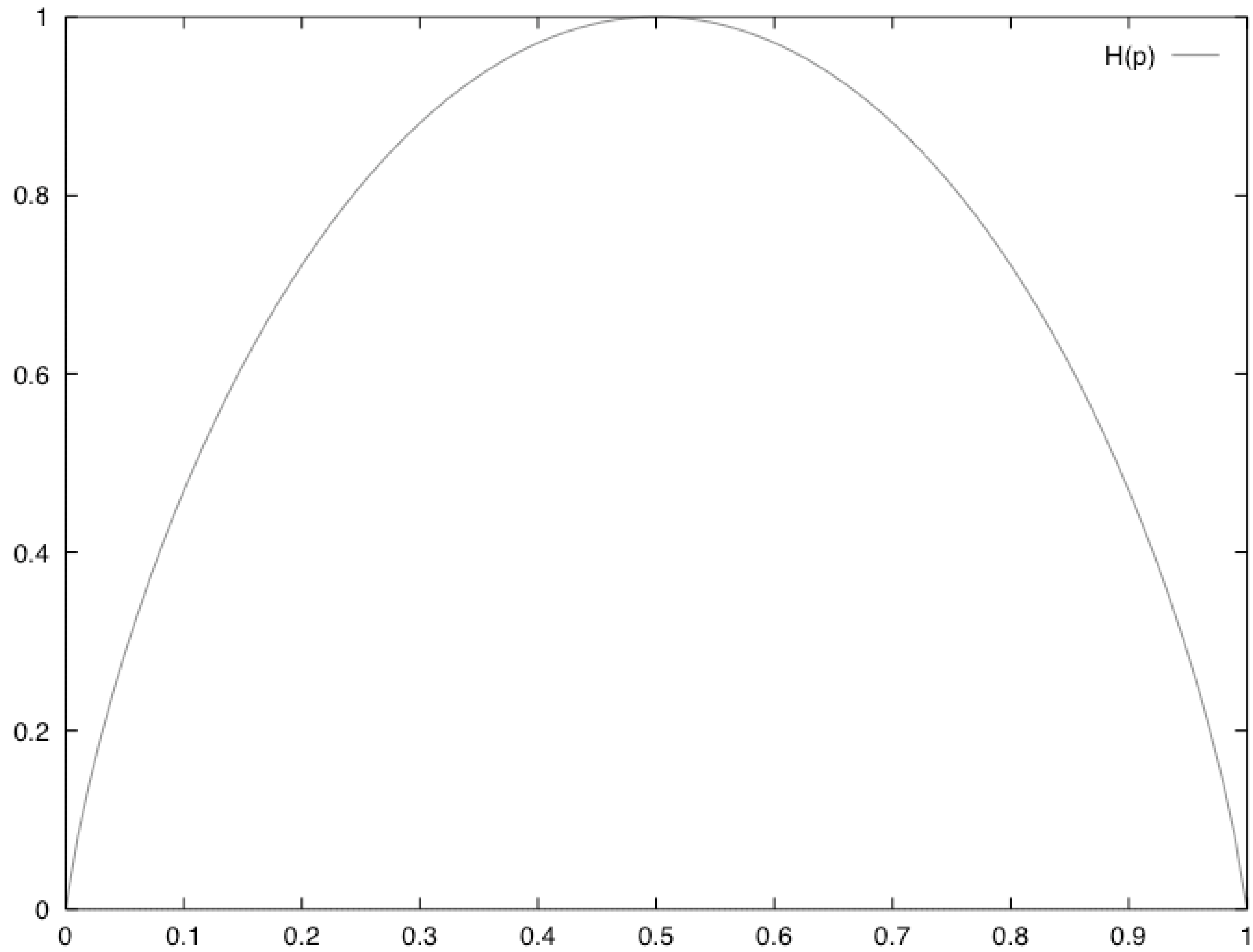
  $E[X] = \Sigma_{x \in X(\Omega)} x \cdot p(x)$

- Then

  $E[-\log_2(p(x))] = \Sigma_{x \in X(\Omega)} -\log_2(p(x)) \cdot p(x)$

  $= H(X)$

Entropy of a coin



Entropy, intuitively

- High entropy ~ "chaos", fuzziness, opposite of order
- Comes from physics:
  - Entropy does not go down unless energy is used

- Measure of uncertainty
  - High entropy: a lot of uncertainty about the outcome, uniform distribution over outcomes
  - Low entropy: high certainty about the outcome

Claude Shannon

1950

- Claude Shannon
  1916 - 2001
  Creator of Information Theory

- Lays the foundation for implementing logic in digital circuits as part of his <u>Masters</u> Thesis! (1939)
- "A Mathematical Theory of Communication" (1948)

Joint Entropy and Conditional Entropy

- Two random variables: X (space Ω), Y (Ψ)
- Joint entropy
  - no big deal: (X,Y) considered a single event:

$$H(X,Y) = - \Sigma_{x\in\Omega} \Sigma_{y\in\Psi} p(x,y) \log_2 p(x,y)$$

- Conditional entropy:

$$H(X|Y) = - \Sigma_{x\in\Omega} \Sigma_{y\in\Psi} p(x,y) \log_2 p(x|y)$$

  - recall that $H(X) = E[-\log_2(p(x))]$ (weighted average, and <u>weights are not conditional</u>)
  - How much extra information you need to supply to transmit X *given that the other person knows Y.*

Conditional Entropy (another way)

$$H(Y \mid X) = \sum_x p(x) H(Y \mid X = x)$$

$$= \sum_x p(x) (- \sum_y p(y \mid x) \log_2 (p(y \mid x))$$

$$= - \sum_x \sum_y p(x) p(y \mid x) \log_2 (p(y \mid x))$$

$$= - \sum_x \sum_y p(x, y) \log_2 (p(y \mid x))$$

Chain Rule for Entropy

- Since, like random variables, entropy is based on an expectation..

$$H(X, Y) = H(Y|X) + H(X)$$

$$H(X, Y) = H(X|Y) + H(Y)$$

Cross Entropy

- What happens when you use a code that is sub-optimal for your event distribution?
  - I created my code to be efficient for a fair 8-sided die.
  - But the coin is unfair and always gives 1 or 2 uniformly.
  - How many bits on average for the optimal code?
    How many bits on average for the sub-optimal code?

$$H(p,q) = -\sum_{x \in X} p(x) \log_2(q(x))$$

KL Divergence

- What are the average number of bits that are wasted by encoding events from distribution *p* using distribution *q*?

$$D(p \,||\, q) = H(p,q) - H(p)$$

$$= -\sum_{x \in X} p(x)\log_2(q(x)) + \sum_{x \in X} p(x)\log_2(p(x))$$

$$= \sum_{x \in X} p(x)\log_2\left(\frac{p(x)}{q(x)}\right)$$

A sort of "distance" between distributions *p* and *q*, but
It is not symmetric!
It does not satisfy the triangle inequality!

Mutual Information

- Recall: H(X) = average # bits for me to tell you which event occurred from distribution P(X).
- Now, first I tell you event y ∈ Y, H(X|Y) = average # bits necessary to tell you which event occurred from distribution P(X)?
- By how many bits does knowledge of Y lower the entropy of X?

$$I(X;Y) = H(X) - H(X \mid Y)$$

$$= H(X) + H(Y) - H(X,Y)$$

$$= -\sum_x p(x)\log_2 \frac{1}{p(x)} - \sum_y p(y)\log_2 \frac{1}{p(y)} + \sum_{x,y} p(x,y)\log_2 p(x,y)$$

$$= \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}$$

Mutual Information

- Symmetric, non-negative.
- Measure of independence.
  - $I(X;Y) = 0$ when X and Y are independent
  - $I(X;Y)$ grows both with degree of dependence and entropy of the variables.
- Sometimes also called “information gain”

- Used often in NLP
  - clustering words
  - word sense disambiguation
  - feature selection…

Pointwise Mutual Information

- Previously measuring mutual information between two random variables.
- Could also measure mutual information between two events

$$I(x,y) = \log \frac{p(x,y)}{p(x)\,p(y)}$$