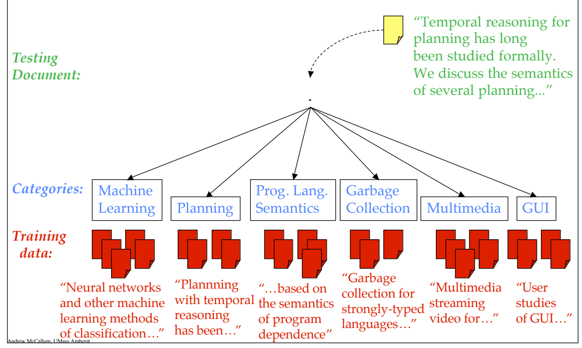
Classification & Information Theory
Lecture #3

Introduction to Natural Language Processing
CMPSCI 585, Spring 2004

University of Massachusetts Amherst

Andrew McCallum

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:

$$\Pr(c_j \mid d) = \frac{\Pr(c_j)\Pr(d \mid c_j)}{\Pr(d)}$$

"Naïve Bayes":

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

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

Parameter Estimation in Naïve Bayes

Estimate of P(c)

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

Estimate of P(w|c)

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

Common words in *Tom Sawyer*
(71,370 words)

Word	Freq	Use
the	3332	determiner (article)
and	2972	conjunction
a	1775	determiner
to	1725	preposition, verbal infinitive marker
of	1440	preposition
was	1161	auxiliary verb
it	1027	(personal/expletive) pronoun
in	906	preposition
that	877	complementizer, demonstrative
he	877	(personal) pronoun
I	783	(personal) pronoun
his	772	(possessive) pronoun
you	686	(personal) pronoun
Tom	679	proper noun
with	642	preposition

Frequencies of frequencies in *Tom Sawyer*

71,730 word tokens
8,018 word types

Word Frequency	Frequency of Frequency
1	3993
2	1292
3	664
4	410
5	243
6	199
7	172
8	131
9	82
10	91
11-50	540
51-100	99
>100	102

Ziph's law *Tom Sawyer*

Word	Freq. (f)	Rank (r)	f * r
the	3332	1	3332
and	2972	2	5944
a	1775	3	5235
he	877	10	8770
but	710	20	8400
be	294	30	8820
there	222	40	8880
one	172	50	8600
about	158	60	9480
more	138	60	9480
never	124	80	9920
Oh	116	90	10440
two	104	100	10400

Ziph's law *Tom Sawyer*

Word	Freq. (f)	Rank (r)	f * r
turned	51	200	10200
you'll	30	300	9000
name	21	400	8400
comes	16	500	8000
group	13	600	7800
lead	11	700	7700
friends	10	800	8000
begin	9	900	8100
family	8	1000	8000
brushed	4	2000	8000
sins	2	3000	6000
Could	2	4000	8000
Applausive	1	8000	8000

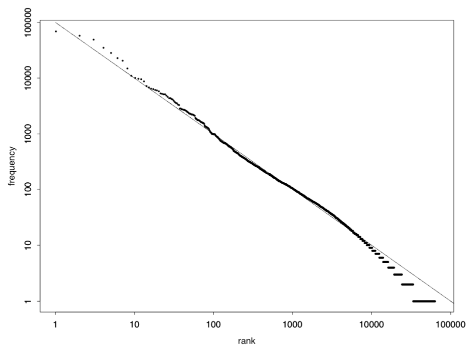
Zipf's law

$$f \propto \frac{1}{r}$$

In other words, there is a constant, k, such that

$$f \cdot r = k$$

Zipf's Law and the Brown Corpus



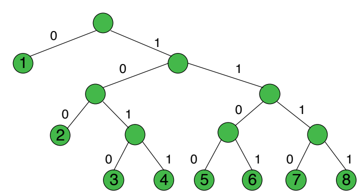
Information Theory

What is Information?

- "The sun will come up tomorrow."
- "Greenspan 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)=0.5, P(2)=0.125, P(3)=…=P(8)=0.0625



"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):

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

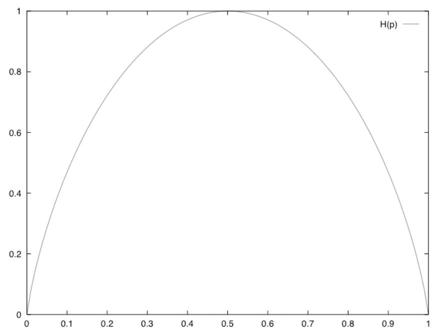
- Notation: $H(X) = H_p(X)=H(p)=H_X(p)=H(p_X)$

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 Masters 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 weights are not conditional)

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(X|Y) + H(X)

  H(X, Y) = H(Y|X) + 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 $y \in Y$, 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?

$$I(X;Y) = H(X) - H(X|Y)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= \sum_x p(x)\log_2\frac{1}{p(x)} + \sum_y p(y)\log_2\frac{1}{p(y)} - \sum_{x,y} p(x,y)\log_2 p(x,y)$$
$$= \sum_{x,y} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$$

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 random variables

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