CS598JHM: **Advanced NLP** (Spring '10)

Lecture 1: Introduction

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

http://www.cs.uiuc.edu/class/sp10/cs598jhm

Why should you take this class?

Many recent NLP papers use **Bayesian methods**.
If you want to do research in NLP, you need to understand these papers!

You may wonder:
- What are Bayesian methods?
(or: what's Gibbs sampling, variational inference, LDA, HDP, conjugate priors?)
- Why are they used?
- How can you use them in your own research?

You may also wonder:
- What's the difference between generative and discriminative models, or generative and discriminative estimation?
(and why should I care?)

Why should you take this class?

CS546 has been canceled!
(and you need to fill your schedule)

What are we going to do in this class?

Graphical models are the foundation of much of machine learning used in NLP (and elsewhere).

Graphical models enable us to work with very complex probabilistic models, because they have efficient **inference** and **parameter estimation** methods.

You've (hopefully) all seen graphical models before, but it won't harm to go over the basics again and delve a bit deeper.

What are we going to do in this class?

Foundations:
Relevant parts of probability theory
Graphical models
Parameter estimation
Inference: variational methods and sampling

Applications:
Topic models (LDA etc.)
Finite-state methods: e.g. HMMs and extensions
Grammar-based methods

What are we going to do in this class?

Lectures:
Background material

Paper presentations:
Current research

Readings:
Background material and current papers

Homework:
More reading (and some writing)

Research project:
Applying what we're learning (and again some writing)

Grades etc.

Your grade will consist of:

- 50% **research project**
- 30% **class presentations**
- 10% (mini) **literature reviews**
- 10% **class participation**

The research project

The goal: write a research paper.
We may submit good papers to conferences/workshops.
You can work alone or in pairs.

The project can be related to your thesis research, but it needs to be clear that what you are doing for this class is something different from what you would have done anyway.

Grading criteria:
- technical soundness
- originality
- quality of writeup

You may have to give a couple of brief presentations in class about your project.

Class presentations

You will be expected to present **two conference papers** (or **one longer journal paper**) in class.

Grading criteria:
- Clarity of presentation
- Can you interpret the paper?

- We will upload your slides to the course website.

Literature reviews

There will be **five homeworks**, in which you will be asked to write a 1-2 page minireview of a conference paper.

What was done? Why? How?
What would be a good follow-up paper?

Class website; email

Slides, pointers to reading material etc. will be posted at:

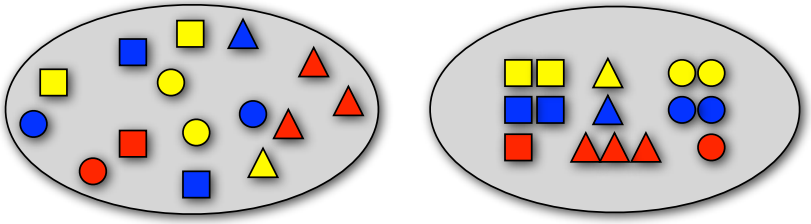
http://www.cs.uiuc.edu/class/sp10/cs598jhm

We'll also have a mailing list.

Basic Probability Theory

Sampling with replacement

Pick a random shape, then put it back in the bag.



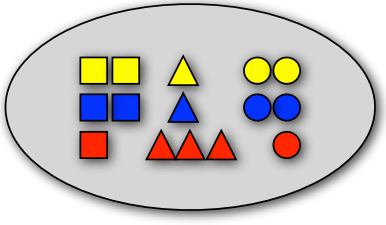
P(■) = 2/15
P(blue) = 5/15
P(blue | □) = 2/5

P(■) = 1/15
P(red) = 5/15
P(□) = 5/15

P(■ or △) = 2/15
P(△ | red) = 3/5

Sampling with replacement

Pick a random shape, then put it back in the bag.
What sequence of shapes will you draw?



P(●△▲■)
= 1/15 × 1/15 × 1/15 × 2/15
= 2/50625

P(△●●▲)
= 3/15 × 2/15 × 2/15 × 3/15
= 36/50625

P(■) = 2/15
P(■) = 2/15
P(■) = 1/15

P(△) = 1/15
P(▲) = 1/15
P(▲) = 3/15

P(●) = 2/15
P(●) = 2/15
P(●) = 1/15

Sampling with replacement

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

P(of) = 3/66
P(Alice) = 2/66
P(was) = 2/66
P(to) = 2/66

P(her) = 2/66
P(sister) = 2/66
P(,) = 4/66
P(') = 4/66

Sampling with replacement

beginning by, very Alice but was and?
reading no tired of to into sitting
sister the, bank, and thought of without
her nothing: having conversations Alice
once do or on she it get the book her had
peeped was conversation it pictures or
sister in, 'what is the use had twice of
a book''pictures or' to

P(of) = 3/66
P(Alice) = 2/66
P(was) = 2/66
P(to) = 2/66

P(her) = 2/66
P(sister) = 2/66
P(,) = 4/66
P(') = 4/66

In our model, P(English) = P(word salad)

Conditioning on the previous word

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

P(w$_{i+1}$= of | w$_i$=tired) = 1
P(w$_{i+1}$= of | w$_i$=use) = 1
P(w$_{i+1}$= sister | w$_i$=her) = 1
P(w$_{i+1}$= beginning | w$_i$=was) = 1/2
P(w$_{i+1}$= reading | w$_i$=was) = 1/2
P(w$_{i+1}$= bank | w$_i$=the) = 1/3
P(w$_{i+1}$= book | w$_i$=the) = 1/3
P(w$_{i+1}$= use | w$_i$=the) = 1/3

Probability theory: terminology

Trial: picking a shape, predicting a word

Sample space Ω: the set of all possible outcomes
(all shapes; all words in *Alice in Wonderland*)

Event ω ⊆ Ω: an actual outcome (a subset of Ω)
(predicting '*the*', picking a triangle)

The probability of events

Kolmogorov axioms:

$$0 \leq P(\omega \subseteq \Omega) \leq 1$$
$$P(\emptyset) = 0 \text{ and } P(\Omega) = 1$$
$$\sum_{\omega_i \subseteq \Omega} P(\omega_i) = 1 \quad \text{if } \forall j \neq i : \omega_i \cap \omega_j = \emptyset \text{ and } \bigcup_i \omega_i = \Omega$$

The ω_i form a **partition** of Ω

Random variables

A random variable X is a function which maps every element in the sample space to some value.

Discrete random variables:
- heads or tails,
- age (in years), day of the week,
- number of letters/vowels/*e*'s in a word,
- part of speech of a word
- the word itself

Continuous random variables:
- size, height, weight,….

Discrete probability distributions: Throwing a coin

Bernoulli distribution:

Probability of success (=head,yes) in single yes/no trial

- The probability of *head* is *p*.
- The probability of *tail* is *1–p*.

Binomial distribution:

Prob. of the number of heads in a sequence of yes/no trials

The probability of getting exactly *k* heads in *n* independent yes/no trials is:

$$P(k \text{ heads},\ n-k \text{ tails}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Discrete probability distributions: Rolling a die

Categorical distribution:

Probability of getting one of *N* outcomes in a single trial

- The probability of category/outcome c_i is p_i $(\sum p_i = 1)$

Multinomial distribution:

Probability of observing each possible outcome c_i exactly X_i times in a sequence of *n* yes/no trials

$$P(X_1 = x_i, \ldots, X_N = x_N) = \frac{n!}{x_1! \cdots x_N!} p_1^{x_1} \cdots p_N^{x_N} \quad \text{if } \sum_{i=1}^{N} x_i = n$$

Expectation

The **expectation** $E(X)$ of a *discrete* random variable X with probability $p(x=X)$ is

$$E(X) = \sum_x p(x)x$$

$E(X)$ is also called the **expected value**, **mean** or **average**.

The expectation $E(X)$ of a *continuous* random variable X with probability density function $p(x=X)$ is

$$E(X) = \int_{-\infty}^{+\infty} p(x)x\,dx$$

X=number of vowels/word

Words:
lab, think, eat, come, mail, student, book, lecture, coffee, facebook

$$p(X=1)=0.2 \quad p(X=2)=0.5$$
$$p(X=3)=0.2 \quad p(X=4)=0.1$$

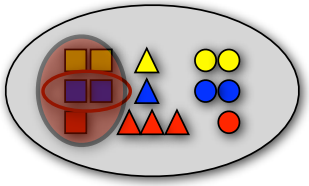
$$\begin{aligned} E(X) &= 1\times 0.2 + 2\times 0.5 + 3\times 0.2 + 4\times 0.1 \\ &= 2.2 \end{aligned}$$

The expected value of a random variable may not correspond to any possible value!

Joint and Conditional Probability

The **conditional probability of X given Y**, $P(X|Y)$, is defined in terms of the probability of Y, $P(Y)$, and the **joint probability of X and Y**, $P(X,Y)$:

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$



P(blue | ■) = 2/5

The chain rule

Conversely, the joint probability $P(X,Y)$ can also be expressed in terms of the conditional probability $P(X|Y)$

$$P(X,Y) = P(X|Y)P(Y)$$

This leads to the so-called **chain rule**:

$$\begin{aligned} P(X_1, X_2, \ldots, X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)....P(X_n|X_1, ...X_{n-1}) \\ &= P(X_1)\prod_{i=2}^{n} P(X_i|X_1 \ldots X_{i-1}) \end{aligned}$$

Independence

Two random variables X and Y are independent if

$$P(X,Y) = P(X)P(Y)$$

If X and Y are independent, then $P(X|Y) = P(X)$:

$$\begin{aligned} P(X|Y) &= \frac{P(X,Y)}{P(Y)} \\ &= \frac{P(X)P(Y)}{P(Y)} \quad (X,\ Y \text{ independent}) \\ &= P(X) \end{aligned}$$

Bayes' theorem

If the events $A_1,...A_k$ form a partition of the space Ω such that $P(A_j) > 0$ for $j=1,...,k$, and B is any event such that $P(B) > 0$, then, for $i=1,....,k$, then:

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{k} P(A_j)P(B|A_j)}$$

Example

We have some widgets,
- 20% were produced by machine M1
- 30% were produced by machine M2
- 50% were produced by machine M3

Unfortunately,
- 1% of the items produced by M1 are defective
- 2% of the items produced by M2 are defective
- 3% of the items produced by M3 are defective

We have one defective widget.
What's the probability that it was produced by M2?

Priors and posteriors

$$P(A_i|B) = \frac{P(A_i)P(B|A_i)}{\sum_{j=1}^{k} P(A_j)P(B|A_j)}$$

$P(A_i)$ is the **prior** probability of A_i

$P(A_i|B)$ is the **posterior** probability of A_i

Bayesian statistics

- We use θ to represent (the parameters of) probability distributions
- Probabilities θ represent degrees of belief
- Data D provides evidence for or against our beliefs. We update our belief θ based on the evidence we see:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{\int P(\theta)P(D|\theta)d\theta}$$

- For a given set of data D, $P(D|\theta)$ is the **likelihood** of θ