Lecture 2: Counting Things

Methods in Computational Linguistics II
Queens College
(Based on slides by Andrew Rosenberg)

Overview

- Role of probability and statistics in computational linguistics
- Basics of Probability.

Training corpus and parameter estimation

- **Karlsson-on-the-Roof**

On **a** perfectly **ordinary** street in Stockholm, in **a** perfectly **ordinary** house, lives **a** perfectly **ordinary** family called Ericson. There is **a** perfectly **ordinary** Daddy and **a** perfectly **ordinary** Mommy and three perfectly **ordinary** children—Bobby, Betty, and Eric….

There is only one person in the entire house who is not **ordinary**—and that is Karlsson-on-the-Roof. He lives on the roof, Karlsson does. This alone is out of the **ordinary**. Things may be different in other parts of the world, but in Stockholm people hardly ever live in **a** little house of their own on top of a roof. But Karlsson does. He is **a** very small, very round, and very self-possessed gentleman—and he can fly! Anybody can fly by airplane or helicopter, but only Karlsson can fly all by himself. He simply turns **a** button in the middle of his tummy and, presto, the cunning little engine on his back starts up. Karlsson waits for **a** moment or two to let the engine warm up; then he accelerates, takes off, and glides on his way with all the dignity and poise of **a** statesman; that is, if you can picture **a** statesman with **a** motor on his back.

Parameter estimation

- What is p(ordinary)?
- What is p(ordinary|perfectly)?
  - *Recall the chain rule: p*(A,B)=p(A)· p(B|A)

    = p(B)· p(A|B)=

    So, p(A|B)=p(B,A)/p(B)

    p(A,B)=c(A,B)/c(B)

    p(ordinary|perfectly)=c(perfectly,ordinary)/c(perfectly)
- What is p(ordinary|a,perfectly)?

What is a probability?

- A degree of belief in a proposition.

- The likelihood of an event occurring.

- Probabilities range between 0 and 1.
- The probabilities of **all mutually exclusive events** sum to 1.

Random Variables

- A *discrete random variable* is a function that
  - takes discrete values from a countable domain and
  - maps them to a number between 0 and 1
  - Example: **Weather** is a discrete (propositional) random variable that has domain <sunny,rain,cloudy,snow>.
    - *sunny* is an abbreviation for *Weather = sunny*
    - *P(Weather=sunny)=0.72, P(Weather=rain)=0.1*, etc.
    - Can be written: *P(sunny)=0.72, P(rain)=0.1,* etc.
    - Domain values must be exhaustive and mutually exclusive
- Other types of random variables:
  - *Boolean random variable* has the domain <true,false>,
    - e.g., *Cavity* (special case of discrete random variable)
  - *Continuous random variable* as the domain of real numbers

Prior Probability

- *Prior (unconditional) probability*
  - corresponds to belief prior to arrival of any (new) evidence
  - ***P(sunny)=0.72, P(rain)=0.1*, etc.**
- *Probability distribution* gives values for all possible assignments:
  - Vector notation: Weather is one of <0.72, 0.1, 0.08, 0.1>
  - ***P****(Weather) = <0.72,0.1,0.08,0.1>*
  - Sums to 1 over the domain

Joint Probability

- Probability assignment to all combinations of values of random variables

	Toothache	¬ Toothache
Cavity	0.04	0.06
¬ Cavity	0.01	0.89

- **The sum of the entries in this table has to be 1**
- *Every question about a domain can be answered by the joint distribution*
- Probability of a proposition is the sum of the probabilities of atomic events in which it holds
  - P(cavity) = 0.1 [add elements of cavity row]
  - P(toothache) = 0.05 [add elements of toothache column]

Joint Probability Table

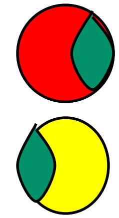
- How could we calculate P(A)?
  - Add up $P(A \wedge B)$ and $P(A \wedge \neg B)$.
- Same for P(B).
- How about $P(A \vee B)$?
  - Two options…
  - We can read $P(A \wedge B)$ from chart and find P(A) and P(B). $P(A \vee B)=P(A)+P(B)-P(A \wedge B)$
  - Or just add up the proper three cells of the table.

$P(A \wedge B)$

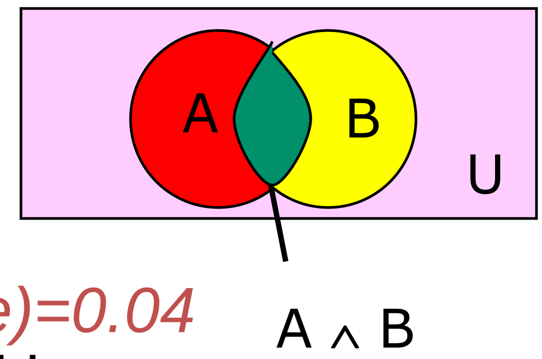
Each cell contains a 'joint' probability of both occurring.

	B	¬ B
A	0.35	0.02
¬ A	0.15	0.48

Conditional Probability



	Toothache	¬ Toothache
Cavity	0.04	0.06
¬ Cavity	0.01	0.89



- *P(cavity)=0.1* and *P(cavity ∧ toothache)=0.04* are both ***prior*** (unconditional) probabilities
- Once the agent has new evidence concerning a *previously unknown* random variable, e.g., toothache, we can specify ***a posterior*** (conditional) probability
  - **e.g., *P(cavity | toothache)***
- *P(A | B) = P(A ∧ B) / P(B)* [prob of A w/ U limited to B]
- *P(cavity | toothache) = 0.04 / 0.05 = 0.8*

Review of Notation

- What do these notations mean?

A ← Boolean Random Variable

$P(A)$ ← Unconditional Probability.
The notation P(A) is a shortcut for P(A=true).

$P(\neg A)$

$P(A \vee B)$ ← Probability of A or B: $P(A) + P(B) - P(A \wedge B)$

$P(A \wedge B)$ ← Joint Probability. Probability of A and B together.

$P(A \mid B)$ ← Probability of A given that we know B is true.

Product Rule

$$P(A \wedge B) = P(A|B) * P(B)$$

$$P(A|B) = P(A \wedge B) / P(B)$$

So, if we can find two of these values someplace (in a chart, from a word problem), then we can calculate the third one.

Using the Product Rule

- When there's a fire, there's a 99% chance that the alarm will go off.

  $P(A \mid F)$

- On any given day, the chance of a fire starting in your house is 1 in 5000.

  $P(F)$

- What's the chance of there being a fire and your alarm going off tomorrow?

  $P(A \wedge F) = P(A \mid F) * P(F)$

Conditioning

- Sometimes we call the 2nd form of the product rule the "conditioning rule" because we can use it to calculate a conditional probability from a joint probability and an unconditional one.

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Conditioning Problem

- Out of the 1 million words in some corpus, we know that 9100 of those words are "to" being used as a PREPOSITION.

  $P(\text{PREP} \wedge \text{"to"})$

- Further, we know that 2.53% of all the words that appear in the whole corpus are the word "to".

  $P(\text{"to"})$

- If we are told that some particular word in a sentence is "to" but we need to guess what part of speech it is, what is the probability the word is a PREPOSITION?

  What is $P(\text{PREP} \mid \text{"to"})$?

  Just calculate: $P(\text{PREP}|\text{"to"}) = P(\text{PREP} \wedge \text{"to"}) / P(\text{"to"})$

Marginalizing

What if we are told only joint probabilities about a variable H=h, is there a way to calculate an unconditional probability of H=h?

Yes, when we're told the joint probabilities involving <u>every single value</u> of the other variable…

$$P(H=h) = \sum_{d \in Domain(V)} P(H=h \wedge V=d)$$

Marginalizing Problem

- We have an AI weather forecasting program. We tell it the following information about this weekend… We want it to tell us the chance of rain.

- Probability that there will be rain and lightning is 0.23.
  $P(\text{rain=true} \wedge \text{lightning=true}) = 0.23$
- Probability that there will be rain and no lightening is 0.14.
  $P(\text{rain=true} \wedge \text{lightning=false}) = 0.14$

- What's the probability that there will be rain?
  $P(\text{rain=true})$?   Lightning is only ever true or false.
  $P(\text{rain=true}) = 0.23 + 0.14 = 0.37$

Chain Rule

- Is there a way to calculate a really big joint probability if we know lots of different conditional probabilities?

$$P(f_1 \wedge f_2 \wedge f_3 \wedge f_4 \wedge \ldots \wedge f_{n-1} \wedge f_n) = \begin{array}{l} P(f_1) \; * \\ P(f_2 \mid f_1) \; * \\ P(f_3 \mid f_1 \wedge f_2) \; * \\ P(f_4 \mid f_1 \wedge f_2 \wedge f_3) \; * \\ \ldots \\ \ldots \\ P(f_n \mid f_1 \wedge f_2 \wedge f_3 \wedge f_4 \wedge \ldots \wedge f_{n-1}) \end{array}$$

You can derive this using repeated substitution of the "Product Rule."
$P(A \wedge B) = P(A|B)\,P(B)$

Chain Rule Problem

- If we have a white ball, the probability it is a baseball is 0.76.

  $P(\text{ baseball} \mid \text{white} \wedge \text{ball })$

- If we have a ball, the probability it is white is 0.35.

  $P(\text{white} \mid \text{ball})$

- The probability we have a ball is 0.03.

  $P(\text{ball})$

- So, what’s the probability we have a white ball that is a baseball?

  $P(\text{white} \wedge \text{ball} \wedge \text{baseball}) = 0.76 * 0.35 * 0.03$

Bayes' Rule

Bayes' Rule relates conditional probability distributions:

$$P(h \mid e) = \frac{P(e \mid h) \; * \; P(h)}{P(e)}$$

or with additional conditioning information:

$$P(h \mid e \wedge k) = \frac{P(e \mid h \wedge k) * P(h \mid k)}{P(e \mid k)}$$

Bayes Rule Problem

- The probability I think that my cup of coffee tastes good is 0.80.

  $P(G) = .80$
- I add Equal to my coffee 60% of the time.

  $P(E) = .60$
- I think when coffee has Equal in it, it tastes good 50% of the time.

  $P(G|E) = .50$

- If I sip my coffee, and it tastes good, what are the odds that it has Equal in it?

  $P(E|G) = \quad P(G|E) * P(E) \quad / \quad P(G)$

Bayes' Rule

- $P(disease \mid symptom) = \dfrac{P(symptom \mid disease) * P(disease)}{P(symptom)}$

- Assess diagnostic probability from causal probability:
  - $P(Cause|Effect) = \dfrac{P(Effect|Cause) * P(Cause)}{P(Effect)}$
- Prior, Likelihood, Posterior

Bayes Example

- Imagine
  - disease = BirdFlu, symptom = coughing
  - *P(disease | symptom)* is different in BirdFlu-indicated country vs. USA
  - *P(symptom | disease)* should be the same
    - It is more useful to learn *P(symptom | disease)*

Conditioning

- *Idea:* Use *conditional probabilities* instead of joint probabilities
- $P(A) = P(A \wedge B) + P(A \wedge \neg B)$

  $= P(A \mid B) * P(B) + P(A \mid \neg B) * P(\neg B)$

  Example:

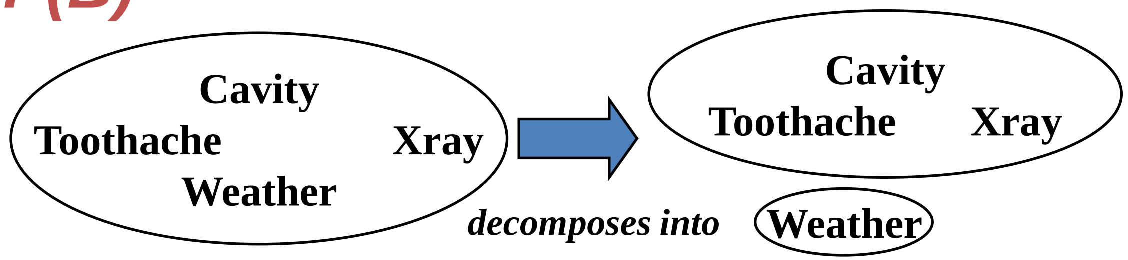
  $P(symptom) =$

  $P(symptom \mid disease) \quad * P(disease) +$

  $P(symptom \mid \neg disease) * P(\neg disease)$
- More generally: $P(Y) = \sum_z P(Y|z) * P(z)$
- Marginalization and conditioning are useful rules for derivations involving probability expressions.

Independence

- A and B are ***independent*** iff
  - $P(A \wedge B) = P(A) * P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$



$$P(T, X, C, W) = P(T, X, C) * P(W)$$

- Independence is essential for efficient probabilistic reasoning

Conditional Independence

- A and B are *<u>conditionally independent</u> given C* iff
  - $P(A \mid B, C) = P(A \mid C)$
  - $P(B \mid A, C) = P(B \mid C)$
  - $P(A \wedge B \mid C) = P(A \mid C) * P(B \mid C)$

- Toothache (T), Spot in Xray (X), Cavity (C)
  - None of these propositions are independent of one other
  - But:
    T and X are conditionally independent given C

Frequency Distribution

Word Tally

the	𝍸 𝍸 𝍸 𝍷𝍷𝍷𝍷
been	𝍸 𝍸 𝍷
message	𝍷𝍷𝍷𝍷
persevere	𝍷
nation	𝍸 𝍷𝍷𝍷

- Count up the number of occurrences of each member of a set of items.
- This counting can be used to calculate the probability of seeing any word.