CIS 833 – Information Retrieval and Text Mining

Lecture 11

Probabilistic Models

September 29, 2015

Credits for slides: Hofmann, Mihalcea, Mobasher, Mooney, Schutze.

Assignments

- HW3 due October 2^{nd} (extended)
- PA1 due October 16^{th} (extended)
- Exam 1 – October 13^{th}

Classes of Retrieval Models

- Boolean models (set theoretic)
  - Extended Boolean
- Vector space models (algebraic)
  - Generalized VS
  - Latent Semantic Indexing
- Probabilistic models
  - Inference Networks
  - Belief Networks

Exact match

Ranking - "Best" match

Required Reading

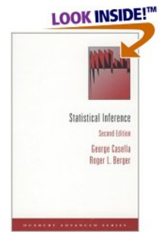
Probability Review

- Textbook Reading
  - Chapter 11: 11.1 - Review of basic probability theory

Probabilistic Retrieval Models

- Chapter 11: 11.2-11.4 - Probabilistic retrieval models

Probability & Statistics References

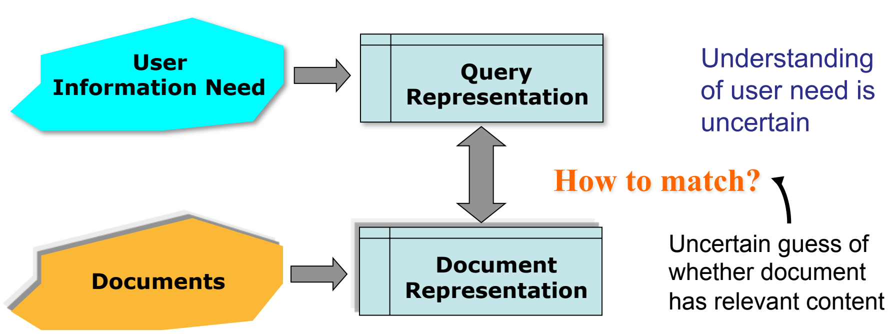


Statistical Inference (Hardcover)
George Casella, Roger L. Berger

Basic Statistics: http://www.statsoft.com/textbook/stbasic.html
(correlations, tests, frequencies, etc.)

Electronic Statistics Textbook: StatSoft
http://www.statsoft.com/textbook/stathome.html
(from basic statistics to ANOVA to discriminant analysis, clustering, regression data mining, machine learning, etc.)

Why probabilities in IR?



In traditional IR systems, matching between each document and query is attempted in a semantically imprecise space of index terms.

Probabilities provide a principled foundation for uncertain reasoning.

Can we use probabilities to quantify our uncertainties?

Definition of Probability

- *Frequentist* interpretation: the probability of an event is the proportion of the time events of same kind will occur in the long run.
- Examples
  - – the probability my flight to Chicago will be on time
  - – the probability my ticket will win the lottery
  - – the probability it will snow tomorrow
- Always a number in the interval [0,1]
  - 0 means “never occurs”
  - 1 means “always occurs”
- The *Bayesian* view of probability is related to degree of belief. It is a measure of the plausibility of an event given incomplete knowledge.

http://www.behind-the-enemy-lines.com/2008/01/are-you-bayesian-or-frequentist-or.html

Sample Spaces

- *Sample space:* a set of possible outcomes for some event
- Examples
  - – flight to Chicago: {on time, late}
  - – lottery: {ticket 1 wins, ticket 2 wins,…,ticket n wins}
  - – weather tomorrow:

    {rain, not rain} or

    {sun, rain, snow} or

    {sun, clouds, rain, snow, sleet} or…

Random Variables

- *Random variable*: a variable representing the outcome of an experiment
- Example
  - X represents the outcome of my flight to Chicago
  - We write the probability of my flight being on time as

    $\Pr(X = on\text{-}time)$ or $P(X = on - time)$

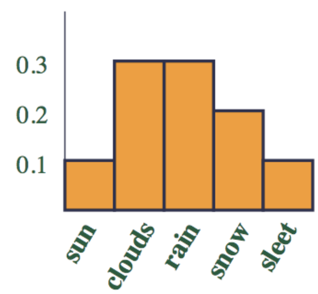
  - When it's clear which variable we're referring to, we may use the shorthand $\Pr(on\text{-}time)$ or $P(on - time)$

Probability Distributions

- If X is a random variable, the function given by $\Pr(X = x)$ for each x is the *probability distribution* of X
- Requirements:

$$\Pr(x) \geq 0 \text{ for every } x$$

$$\sum_x \Pr(x) = 1$$



A histogram plots the number of times or the frequency with which each value of a given variable is observed.

Probability Distribution = Frequency Histogram

Joint Distributions

- *Joint probability distribution:* the function given by
  $\Pr(X = x, Y = y)$
- Read "*X* equals *x* and *Y* equals *y*"
- Example

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

probability that it's sunny and my flight is on time

Marginal Distributions

- The *marginal distribution* of *X* is defined by

$$\Pr(x) = \sum_y \Pr(x, y)$$

- "The distribution of *X* ignoring other variables"
- This definition generalizes to more than two variables, e.g.

$$\Pr(x) = \sum_y \sum_z \Pr(x, y, z)$$

Marginal Distribution Example

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

Marginal Distribution Example

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distribution for X

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

Conditional Distributions

- The *conditional distribution* of *X* given *Y* is defined as:

$$\Pr(X = x \mid Y = y) = \frac{\Pr(X = x, Y = y)}{\Pr(Y = y)}$$

- "the distribution of *X* given that we know *Y* "

Conditional Distribution Example

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

Conditional Distribution Example

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

conditional distribution for X given Y=on-time

x	$\Pr(X = x \mid Y = \textit{on-time})$
sun	0.20/0.45 = 0.444
rain	0.20/0.45 = 0.444
snow	0.05/0.45 = 0.111

Independence

- Two random variables, X and Y, are *independent* if

$$\Pr(x, y) = \Pr(x) \times \Pr(y) \text{ for all } x \text{ and } y$$

Independence Example #1

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$\Pr(Y = y)$
on-time	0.45
late	0.55

Are *X and Y independent here?*

Independence Example #1

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, on-time	0.20
rain, on-time	0.20
snow, on-time	0.05
sun, late	0.10
rain, late	0.30
snow, late	0.15

marginal distributions

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$\Pr(Y = y)$
on-time	0.45
late	0.55

Are *X and Y independent here? NO.*

Independence Example #2

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Northwest	0.03
rain, fly-Northwest	0.05
snow, fly-Northwest	0.02

marginal distributions

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$\Pr(Y = y)$
fly-United	0.9
fly-Northwest	0.1

Are *X and Y independent here?*

Independence Example #2

joint distribution

x, y	$\Pr(X = x, Y = y)$
sun, fly-United	0.27
rain, fly-United	0.45
snow, fly-United	0.18
sun, fly-Northwest	0.03
rain, fly-Northwest	0.05
snow, fly-Northwest	0.02

marginal distributions

x	$\Pr(X = x)$
sun	0.3
rain	0.5
snow	0.2

y	$\Pr(Y = y)$
fly-United	0.9
fly-Northwest	0.1

Are *X and Y independent here? YES.*

Conditional Independence

- Two random variables *X* and *Y* are *conditionally independent* given *Z* if

$$\Pr(X \mid Y, Z) = \Pr(X \mid Z)$$

"once you know the value of *Z*, knowing *Y* doesn't tell you anything about *X* "

- Alternatively

$$\Pr(x, y \mid z) = \Pr(x \mid z) \times \Pr(y \mid z) \text{ for all } x, y, z$$

Conditional Independence Example

Flu	Fever	Vomit	Pr
true	true	true	0.04
true	true	false	0.04
true	false	true	0.01
true	false	false	0.01
false	true	true	0.009
false	true	false	0.081
false	false	true	0.081
false	false	false	0.729

Fever and Vomit are not independent: e.g. $\Pr(fever, vomit) \neq \Pr(fever) \times \Pr(vomit)$

Fever and Vomit are conditionally independent given Flu:

$$\Pr(fever, vomit \mid flu) = \Pr(fever \mid flu) \times \Pr(vomit \mid flu)$$

$$\Pr(fever, vomit \mid \neg flu) = \Pr(fever \mid \neg flu) \times \Pr(vomit \mid \neg flu)$$

etc.

Bayes Rule

$$\Pr(x \mid y) = \frac{\Pr(y \mid x)P(x)}{\Pr(y)} = \frac{\Pr(y \mid x)\Pr(x)}{\sum_x \Pr(y \mid x)\Pr(x)}$$

This theorem is extremely useful

- There are many cases when it is hard to estimate $\Pr(x \mid y)$ directly, but it's not too hard to estimate $\Pr(y \mid x)$ and $\Pr(x)$

$$\Pr(y) = \sum_x \Pr(x,y) = \sum_x \Pr(y \mid x)\Pr(x)$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London, **53:370-418**

Bayes Theorem Example

- MDs usually aren't good at estimating Pr(*Disorder* | *Symptom*)
- They're usually better at estimating Pr(*Symptom* | *Disorder*)
- If we can estimate Pr(*Fever* | *Flu*) and Pr(*Flu*) we can use Bayes' Theorem to do diagnosis

$$\Pr(flu \mid fever) = \frac{\Pr(fever \mid flu)\Pr(flu)}{\Pr(fever \mid flu)\Pr(flu) + \Pr(fever \mid \neg flu)\Pr(\neg flu)}$$

Expected Values

- The *expected value* of a random variable that takes on numerical values is defined as:

  $$E[X] = \sum_x x \times \Pr(x)$$

  This is the same thing as the *mean.*
- We can also talk about the expected value of a function of a random variable

  $$E[g(X)] = \sum_x g(x) \times \Pr(x)$$

Expected Value Examples

$$E\left[Shoesize\right] = 5 \times \Pr(Shoesize = 5) + \ldots + 14 \times \Pr(Shoesize = 14)$$

Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

$$E\left[gain(Lottery)\right] =$$

Expected Value Examples

$$E[Shoesize] = \\ 5 \times \Pr(Shoesize = 5) + \ldots + 14 \times \Pr(Shoesize = 14)$$

Suppose each lottery ticket costs $1 and the winning ticket pays out $100. The probability that a particular ticket is the winning ticket is 0.001.

$$E\left[gain(Lottery)\right] = \\ gain(\text{winning})\Pr(\text{winning}) + gain(\text{losing})\Pr(\text{losing}) = \\ (\$100 - \$1) \times 0.001 - \$1 \times 0.999 = \\ -\$0.90$$

Likelihood

- We often speak of the probability of some data *D* given some distribution or model *M*: *Pr*(*D*|*M*).
- We call *Pr*(*D*|*M*) the **likelihood of *D* given *M*.**
- **Note this is not the same as the probability of *M* given *D*, but they are related by Bayes rule.**

$$\Pr(M \mid D) = \frac{\Pr(D \mid M)\Pr(M)}{\Pr(D)} = \alpha \Pr(D \mid M)\Pr(M)$$

- Here α is a normalization constant.
- Sometimes we use the word "hypothesis" instead of model.

Likelihood (Continued)

$$\Pr(M \mid D) = \frac{\Pr(D \mid M)\Pr(M)}{\Pr(D)} = \alpha \Pr(D \mid M)\Pr(M)$$

- We can compute Pr(M|D) from Pr(D|M) if we also have Pr(M), i.e., a prior probability distribution over models.
- Often, we are interested in the maximum likelihood model given data.
- Often talk about log likelihood (typically base 2).

Odds Ratio

- We may be interested in the relative probabilities of two models M1 and M2, given data, or the ratio Pr(M1|D)/Pr(M2|D).
- If the prior probabilities of the models are the same (e.g., uniform prior), this is the same as the relative probabilities of the likelihoods: Pr(D|M1)/Pr(D|M2).
- This ratio of likelihoods is called the **odds ratio.**

Probabilistic IR topics

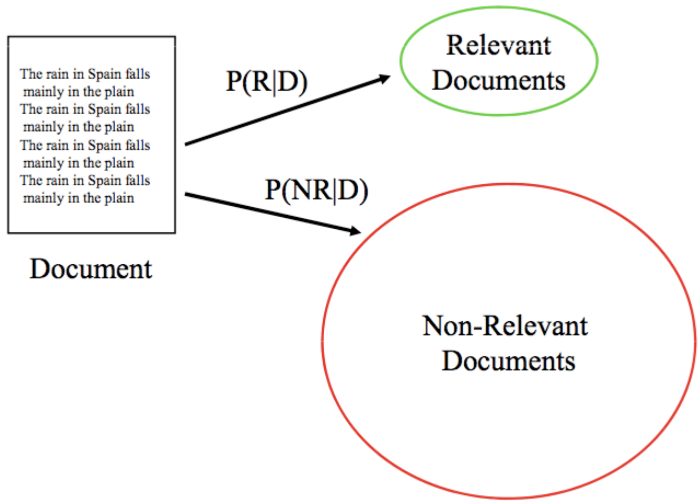
- Classical probabilistic retrieval model
  - Probability ranking principle, etc.
- Language model approach to IR
  - An important emphasis in recent work
- Bayesian networks for text retrieval
- (Naïve) Bayesian Text Categorization

Probabilistic methods are one of the oldest but also one of the currently hottest topics in IR.

Basic Probabilistic Retrieval Model

- Retrieval is modeled as a classification process
- Two classes for each query: the *relevant* and *non-relevant* documents
- Given a particular document D, calculate the probability of belonging to the relevant class, retrieve if greater than probability of belonging to non-relevant class
  - i.e., retrieve if $P(R|D) > P(NR|D)$
- Equivalently, rank by *likelihood ratio* $P(D|R) \div P(D|NR)$
- Different ways of estimating these probabilities lead to different probabilistic models

Basic Probabilistic Model



The Probability Ranking Principle

> "If a reference retrieval system's response to each request is a ranking of the documents in the collection in order of decreasing probability of relevance to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data have been made available to the system for this purpose, the overall effectiveness of the system to its user will be the best that is obtainable on the basis of those data."

- [1960s/1970s] S. Robertson, W.S. Cooper, M.E. Maron; van Rijsbergen (1979:113); Manning & Schütze (1999:538)

Probability Ranking Principle

Let d be a document in the collection.
Let R represent **relevance** of a document w.r.t. given (fixed) query and let NR represent **non-relevance.**

R={0,1} vs. NR/R

Need to find P($R|d$) - probability that a document d is **relevant.**

$$P(R \mid d) = \frac{P(d \mid R)P(R)}{P(d)}$$

$$P(NR \mid d) = \frac{P(d \mid NR)P(NR)}{P(d)}$$

P(R),P(NR) - prior probability of retrieving a (non) relevant document

$$P(R \mid d) + P(NR \mid d) = 1$$

P($d|R$), P($d|NR$) - probability that if a relevant (non-relevant) document is retrieved, it is d.

Probability Ranking Principle (PRP)

- Simple case: no selection costs or other utility concerns that would differentially weight errors
- ***Bayes' Optimal Decision Rule***
  - ***d*** is **relevant** iff $P(R|d) > P(NR|d)$
- PRP in action: Rank all documents by $P(R|d)$
- Theorem:
  - Using the PRP is optimal, in that it minimizes the loss (Bayes risk) under 1/0 loss
  - Provable if all probabilities correct, etc. [e.g., Ripley 1996]

Probability Ranking Principle – With Costs

- Assuming retrieval costs:
  - Let *d* be a document
  - *C* - cost of retrieving a <u>relevant</u> document
  - *C'* - cost of retrieving a <u>non-relevant</u> document
- Probability Ranking Principle: if

$$C \cdot p(R \mid d) + C' \cdot (1 - p(R \mid d)) \leq C \cdot p(R \mid d') + C' \cdot (1 - p(R \mid d'))$$

for all *d' not yet retrieved*, then *d* **is the next document to be retrieved**

Probability Ranking Principle

- How do we compute all those probabilities?
  - Do not know exact probabilities, have to use estimates
  - Binary Independence Retrieval (BIR) – the simplest model
- Questionable assumptions
  - "Relevance" of each document is independent of relevance of other documents – **duplicates** returned
  - Boolean model of relevance
  - The user has a single step information need
    - Seeing a range of results might let user refine query
  - BIR assumptions
    - Documents and queries are represented as binary term incidence vectors (0/1).
    - Terms in a document are independent.