Lecture 7: Probabilistic Information Retrieval,

Bayes' Rule

$$p(A \mid D) = \frac{p(D \mid A)}{p(A)p(D)}$$

A is a Boolean-valued random variable (e.g. the document is relevant)

- p(A) = prior probability of hypothesis A – PRIOR
- p(D) = prior probability of data D – EVIDENCE
- p(D|A) = probability of D given A - LIKELIHOOD
- P(A|D) = probability of A given D – POSTERIOR

Boolean hypotheses: Odds

$$O(A \mid D) = \frac{p(A \mid D)}{p(\neg A \mid D)} = \frac{p(A \mid D)}{1 - p(A \mid D)}$$

Probability Ranking Principle (PRP)

Long version, van Rijsbergen (1979:113-114):

- "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."

Short version:

- Have representative training document collection and training queries, learn p(relevant|document,query)
- Ranking of new documents w.r. t. new queries (similar to, or in, the training collection) with this function is bound to better than anything else

Task 1.4, 2.3

Probability Ranking Principle (PRP)

Let *d* be a document in the collection

- *R* represents relevance of doc w.r.t. given (fixed) query
- *NR* represents non-relevance of doc w.r.t. given (fixed) query

Relevance is binary variable with values {NR,R}

Need to find p(*R*|*d*,*q*) - probability that a document *d* is relevant given query *q*

- This distribution lives in a HUGE space

How can this be done?

- Reformulate using Bayes' rule to distributions easier to learn

Probability Ranking Principle (PRP)

$$p(R \mid d,q) = \frac{p(d \mid R,q)\, p(R \mid q)}{p(d \mid q)}$$

$$p(NR \mid d,q) = \frac{p(d \mid NR,q)\, p(NR \mid q)}{p(d \mid q)}$$

$$p(R \mid d,q) + p(NR \mid d,q) = 1$$

$$O(R \mid d,q) = \frac{p(R \mid d,q)}{p(NR \mid d,q)}$$

p(*R*|*q*), p(*NR*|*q*) - prior probability of retrieving a (non-)relevant document given query *q*

(constant w. r. t. *d*)

PRP in action: Rank all documents by p(*R*|*d*,*q*) or O(*R*|*d*,*q*)