Basic Probability

- Conditional – or Posterior - Probability
  - P(A|B) == the probability that A will occur given that B has occurred
  - P(A|B) = P(A and B both occur) / P(B)
    - For example: in a sentence from the “Wizard of Oz” what is the probability of the next word being “witch” given that the previous word was “wicked”
      - P(witch | wicked)
- Prior Probability
  - The probability of something given no information
  - P(“the”) = 0.07 in English text

Bayes Rule

- Bayes Theorem
  - $P(A|B) = (P(B|A)/P(B))\ P(A)$
    - $P(B|A)/P(B)$ is called the Likelihood of A given B
- Example
  - 2 bowls of cookies
    - Bowl A: 30 Chocolate chip, 10 plain
    - Bowl B: 20 Chocolate chip, 20 plain
  - Suppose randomly pick a bowl and then randomly select a cookie from that bowl. When you do so, you get a chocolate chip
  - What is the probability that you picked from bowl A?
    - $P(A \mid cc)$
  - Know: $P(cc|A) = 0.75$, $P(cc|B) = 0.5$, $P(cc)=0.675$, $P(A) = 0.5$
  - $P(A|cc) = P(cc|A)*P(A)/P(cc) = 0.75*0.5/0.675 = 0.6$

Bayes and IR

- Want to compute
  - P(Q=true | document D)
    = (P(D|Q=true) * P(Q=true)) / P(D)
- So, by Bayes would need:
  - P(D | Q=true), P(Q=true), P(D)
  - P(Q=true) = probability of picking a relevant document from among all documents. (This is the same for all documents)
  - P(D) = P(t1) * ... for each word in D
    - *This we know, but it drops out as a constant term*
    - *Fortunate as VERY near 0.*

- P(D | Q=true)
  - Assumes independence
    - So P(D|Q=true) = P(ti1|Q=true) * P(t2|Q=true)*...
  - For each word ti in D
    - pi = P(Q=true | ti), qi = P(Q!=true | ti)
    - Then let ci = log ((pi(i-qi)) / (qi(1-pi))
      - Roughly, ci is how predictive of relevancy a given word is.
        - The ***relevance weight***
      - Simple model says pi=1 iff word ti in in Q,
        - problem – this leads to division by 0 and is just wrong
          - When would this be correct?
      - Problem – words are all equally weighted.
        - Is this really a problem?
  - Summing the ci gives an indicator of relevance of the document. Roughly, if >0 then relevant
  - Problem, where do you get all of the pi and qi
  - The sum gives a total ordering of documents.

Naïve Bayes Model

- A very popular approach to spam filtering
- Partially because it has a simple mechanism for incorporating user feedback
- Partially because it works pretty well
- Partially because this model allows a tunable threshold for splitting spam and non-spam

Naïve Bayes

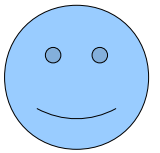
- $$\frac{(P(isSpam) * P(D \vee isSpam))}{\sum_{x \in yes, no} (P(isSpam = x) * P(D \vee isSpam = x))}$$
  - Note that this is not quite the Bayes formulation I gave earlier. Why?
- Assuming independence this is
- $$\frac{(P(isSpam) * \prod_{w \in D} P(w\ given\ isSpam))}{(\sum_{x \in yes, no} (P(isSpam = x) * \prod_{w \in D} P(w\ given\ isSpam = x)))}$$

Naïve Bayes for Spam

- By Heaps law the number of terms in the product will be very large, and by Zipfs law most of these terms will be rare.
- Statistics about rare terms will be noisy (implied by the law of large numbers)
- Therefore, if use every term the estimate will be very noisy
- Therefore, should use only those terms that are reliable
  - Which ones are reliable?

Feature Selection

- Terms that are sufficiently frequent
  - This gives reliable terms but many of these terms will be uninformative (Luhn's idea)
- Use only terms that are highly predictive
  - Mutual Information
  - $\sum_{y \in present, absent} \sum_{x \in (yes, no)} P(word = y, isSpam = x) * \log\left(\frac{(P(word = y, isSpam = x))}{(P(word = y) * P(isSpam = x))}\right)$
  - P(word=y, isSpam=x)
    - For discrete variables
    - = P(word=y | isSpam=x) * P(isSpam=x)
    - = P(isSpam=x | word=y) * P(word=y)
  - Note that MI gives a probabilistic foundation for Luhn's idea.



TF-IDF vs Probabilistic Models

- IDF term looks a lot like divisor in Bayes rule
- TF term looks a lot like – what?
  - It is related to P(ti | D) but this is not used.
- What in Probabilistic model corrects for the benefit of big documents?
  - Unlike cosine distance which is strictly non-negative the ci terms may be less than 0. How?
  - So, VS only accumulates evidence in favor according to the intersection of document and query. Probabilistic accumulates evidence for and against according to all words in document.