CSE 321 Discrete Structures

February 26th, 2010

Lecture 20: Probability Theory

Poll

Makeup class on Wednesday, 3/3

- 4:30 – 5:30 ?
- 5:30 – 6:30 ?

Makeup class is NOT mandatory, but recommended !

Bernoulli Trials and Binomial Distribution

- Bernoulli Trial
  - Success probability p, failure probability q

The probability of exactly k successes in n independent Bernoulli trials is

$$\binom{n}{k} p^k q^{n-k}$$

Random Variables

A random variable is a function from a sample space to the real numbers

Baye's Theorem
Shanon's Expansion Formula

Baye's Theorem: $P(E \mid F) = P(F \mid E) * P(E) / P(F)$

Shanon's Expansion: $P(F) = P(F \mid E)*P(E) + P(F \mid \text{not}(E))*P(\text{not}(E))$

A Consequence

Suppose that E and F are events from a sample space S such that p(E) > 0 and p(F) > 0. Then

$$p(F \mid E) = \frac{p(E \mid F)p(F)}{p(E \mid F)p(F) + p(E \mid \overline{F})p(\overline{F})}$$

Note: book calls this Baye's theorem

False Positives, False Negatives

Let D be the event that a person has the disease

Let Y be the event that a person tests positive for the disease

What can go wrong ?

False positive: P(Y | not(D))

False negative: P(not(Y) | D)

Testing for disease

Disease is very rare: p(D) = 1/100,000

Testing is accurate:
False negative: 1%
False positive: 0.5%

Suppose you get a positive result, what do you conclude?

P(D|Y)

$$p(D \mid Y) = \frac{p(Y \mid D)p(D)}{p(Y \mid D)p(D) + p(Y \mid \overline{D})p(\overline{D})}$$

p(D) = 0.00001

p(Y | D) = 0.99

p(Y | not D) = 0.005

p(D | Y) =
= 0.99*0.00001/(0.99*0.00001 + 0 .005*0.99999)
= 0.0000099/0.00500985
= 0.00197…

Answer: 0.2 % !

Spam Filtering

From: Zambia Nation Farmers Union
[znfukabwe@mail.zamtel.zm]
Subject: Letter of assistance for school installation
To: Richard Anderson

Dear Richard,
I hope you are fine, Iam through talking to local headmen about the possible assistance of school installation. the idea is and will be welcome.
I trust that you will do your best as i await for more from you.
Once again
Thanking you very much
Sebastian Mazuba.

Bayesian Spam filters

- Classification domain
  - Cost of false negative
  - Cost of false positive
- Criteria for spam
  - v1agra, ONE HUNDRED MILLION USD
- Basic question: given an email message, based on spam criteria, what is the probability it is spam

Email message with phrase "Account Review"

- 250 of 20000 messages known to be spam
- 5 of 10000 messages known not to be spam
- Assuming 50% of messages are spam, what is the probability that a message with "Account Review" is spam

$$p(S \mid A) = \frac{p(A \mid S)p(S)}{p(A \mid S)p(S) + p(A \mid \overline{S})p(\overline{S})}$$