CSE 321 Discrete Structures

Winter 2008
Lecture 20
Probability: Bernoulli, Random Variables, Bayes' Theorem

Announcements

- Readings
  - Probability Theory
    - 6.1, 6.2 (5.1, 5.2) Probability Theory
    - 6.3 (New material!) Bayes' Theorem
    - 6.4 (5.3) Expectation
  - Advanced Counting Techniques – Ch 7.
    - Not covered
  - Relations
    - Chapter 8 (Chapter 7)

Highlights from Lecture 19

- Conditional Probability
- Independence

Let E and F be events with p(F) > 0. The conditional probability of E given F, defined by p(E | F), is defined as:

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

The events E and F are independent if and only if $p(E \cap F) = p(E)p(F)$

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

Bayes' Theorem

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

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

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

$p(Y \mid D) \quad p(D \mid Y)$

$p(\overline{Y} \mid D) \quad p(D \mid \overline{Y})$

$p(\overline{Y} \mid \overline{D}) \quad p(\overline{D} \mid \overline{Y})$

$p(Y \mid \overline{D}) \quad p(\overline{D} \mid Y)$

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?

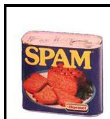
P(D|Y)

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

$p(D) = 0.00001$

$p(Y \mid D) = 0.99$

$p(\overline{Y} \mid \overline{D}) = 0.995$



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

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

Proving Bayes' Theorem

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

$$p(E \mid F) = \frac{p(E \cap F)}{p(F)} \qquad p(F \mid E) = \frac{p(E \cap F)}{p(E)}$$

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

Expectation

The expected value of random variable X(s) on sample space S is:

$$E(X) = \sum_{x \in S} p(s)X(s)$$

Expectation examples

Number of heads when flipping a coin 3 times

Sum of two dice

Successes in n Bernoulli trials with success probability p