4190.101
Discrete Mathematics

Chapter 7 Discrete Probability

Gunhee Kim

[Based on S. Srinivasa Rao's Lecture Notes]

Bayes' Theorem

Section 7.3

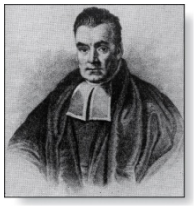
Section Summary

- Bayes' Theorem
- Generalized Bayes' Theorem
- Bayesian Spam Filters
- A.I. Applications (*optional, not currently included in the overheads*)

Motivation for Bayes' Theorem

- Bayes' theorem allows us to use probability to answer questions such as the following:
  - Given that someone tests positive for having a particular disease, what is the probability that they actually do have the disease?
  - Given that someone tests negative for the disease, what is the probability, that in fact they do have the disease?
- Bayes' theorem has applications to medicine, law, artificial intelligence, engineering, and many diverse other areas.

Bayes' Theorem



Thomas Bayes
(1702-1761)

- **Bayes' Theorem**: Suppose that *E* and *F* are events from a sample space S such that $p(E) \neq 0$ and $p(F) \neq 0$. Then:

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

- **Example**: We have two boxes. The first box contains two green balls and seven red balls. The second contains four green balls and three red balls. Bob selects one of the boxes at random. Then he selects a ball from that box at random. If he has a red ball, what is the probability that he selected a ball from the first box?
  - Let *E* be the event that Bob has chosen a red ball and *F* be the event that Bob has chosen the first box.
  - By Bayes' theorem the probability that Bob has picked the first box is:

$$p(F|E) = \frac{(7/9)(1/2)}{(7/9)(1/2) + (3/7)(1/2)} = \frac{7/18}{38/63} = \frac{49}{76} \approx 0.645.$$

Derivation of Bayes' Theorem

- Recall the definition of the conditional probability $p(E|F)$:

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

- From this definition, it follows that:

$$p(E|F) = \frac{p(E \cap F)}{p(F)} \quad , \quad p(F|E) = \frac{p(E \cap F)}{p(E)}$$

- Equating the two formulas for $p(E \cap F)$ shows that

$$p(E|F)p(F) = p(F|E)p(E)$$

- Solving for $p(E|F)$ and for $p(F|E)$ tells us that

$$p(E|F) = \frac{p(F|E)p(E)}{p(F)} \quad , \quad p(F|E) = \frac{p(E|F)p(F)}{p(E)}$$

continued →

Derivation of Bayes' Theorem

- Note that $p(E) = p(E|F)p(F) + p(E|\overline{F})p(\overline{F})$

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

because $E = E \cap S = E \cap (F \cup \overline{F}) = (E \cap F) \cup (E \cap \overline{F})$

and $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$

By the definition of conditional probability,

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

- Hence, $$p(F|E) = \frac{p(E|F)p(F)}{p(E|F)p(F) + p(E|\overline{F})p(\overline{F})}$$ ◄

Applying Bayes' Theorem

- **Example**: Suppose that one person in 100,000 has a particular disease. There is a test for the disease that gives a positive result 99% of the time when given to someone with the disease. When given to someone without the disease, 99.5% of the time it gives a negative result. Find
  a) the probability that a person who test positive has the disease.
  b) the probability that a person who test negative does not have the disease.
- Should someone who tests positive be worried?

Applying Bayes' Theorem

- **Solution**: Let *D* be the event that the person has the disease, and *E* be the event that this person tests positive. We need to compute $p(D|E)$ from $p(\mathrm{D})$, $p(E|D)$, $p(E|\overline{D})$, $p(\overline{D})$.

$$p(D) = 1/100,000 = 0.00001 \quad p(\overline{D}) = 1 - 0.00001 = 0.99999$$

$$p(E|D) = .99 \quad p(\overline{E}|D) = .01 \quad p(E|\overline{D}) = .005 \quad p(\overline{E}|\overline{D}) = .995$$

$$p(D|E) = \frac{p(E|D)p(D)}{p(E|D)p(D) + p(E|\overline{D})p(\overline{D})}$$

$$= \frac{(0.99)(0.00001)}{(0.99)(0.00001) + (0.005)(0.99999)}$$

$$\approx 0.002$$

Can you use this formula to explain why the resulting probability is surprisingly small?

So, don't worry too much, if your test for this disease comes back positive.

Applying Bayes' Theorem

– What if the result is negative?

$$p(\overline{D}|\overline{E}) = \frac{p(\overline{E}|\overline{D})p(\overline{D})}{p(\overline{E}|\overline{D})p(\overline{D}) + p(\overline{E}|D)p(D)}$$

$$= \frac{(0.995)(0.99999)}{(0.995)(0.99999) + (0.01)(0.00001)}$$

$$\approx 0.9999999$$

So, the probability you have the disease if you test negative is

$$p(D|\overline{E}) \approx 1 - 0.9999999 = 0.0000001.$$

– So, it is extremely unlikely you have the disease if you test negative.

Generalized Bayes' Theorem

- **Generalized Bayes' Theorem**: Suppose that *E* is an event from a sample space *S* and that F_1, F_2, …, F_n are mutually exclusive events such that

$$\bigcup_{i}^{n} F_i = S.$$

- Assume that $p(E) \neq 0$. Then for $i = 1, 2, \ldots, n$,

$$p(F_j|E) = \frac{p(E|F_j)p(F_j)}{\sum_{i=1}^{n} p(E|F_i)p(F_i)}.$$

  - *Exercise* 17 *for the proof*.

Bayesian Spam Filters

- How do we develop a tool for determining whether an email is likely to be spam?
- If we have an initial set *B* of spam messages and set *G* of non-spam messages. We can use this information along with Bayes' law to predict the probability that a new email message is spam.
- We look at a particular word *w*, and count the number of messages where it occurs in *B* and in *G*; $n_B(w)$ and $n_G(w)$.
  - Estimated probability that an email containing *w* is spam: $p(w) = n_B(w)/|B|$
  - Estimated probability that an email containing *w* is non-spam: $q(w) = n_G(w)/|G|$

continued →

Bayesian Spam Filters

- Let *S* be the event that the message is spam, and *E* be the event that the message contains the word *w*.
- Using Bayes' Rule, $p(S|E) = \dfrac{p(E|S)p(S)}{p(E|S)p(S) + p(E|\overline{S})p(\overline{S})}$

Assuming that it is equally likely that an arbitrary message is spam and is not spam; i.e., $p(S) = ½$.

$$p(S|E) = \frac{p(E|S)}{p(E|S) + p(E|\overline{S})}$$

Note: If we have data on the frequency of spam messages, we can obtain a better estimate for p(s).

$$r(w) = \frac{p(w)}{p(w) + q(w)}$$

Using our empirical estimates of $p(E|S)$ and $p(E|\bar{S})$.

r(*w*) estimates the probability that the message is spam. We can class the message as spam if *r*(*w*) is above a threshold.

Bayesian Spam Filters

- **Example**: We find that the word "Rolex" occurs in 250 out of 2000 spam messages and occurs in 5 out of 1000 non-spam messages. Estimate the probability that an incoming message is spam. Suppose our threshold for rejecting the email is 0.9.
- **Solution**: $p(Rolex)$ = 250/2000 =.0125 and $q(Rolex)$ = 5/1000 = 0.005.

$$r(Rolex) = \frac{p(Rolex)}{p(Rolex) + q(Rolex)} = \frac{0.125}{0.125 + .005} = \frac{0.125}{0.125 + .005} \approx 0.962$$

We class the message as spam and reject the email!

Bayesian Spam Filters using Multiple Words

- Accuracy can be improved by considering more than one word as evidence.
- Consider the case where E_1 and E_2 denote the events that the message contains the words w_1 and w_2 respectively.
- We make the simplifying assumption that the events are independent. And again we assume that $p(S) = ½$.

$$p(S|E_1 \cap E_2) = \frac{p(E_1|S)p(E_2|S)}{p(E_1|S)p(E_2|S) + p(E_1|\overline{S})p(E_2|\overline{S})}$$

$$r(w_1, w_2) = \frac{p(w_1)p(w_2)}{p(w_1)p(w_2) + q(w_1)q(w_2)}$$

Bayesian Spam Filters using Multiple Words

- **Example**: We have 2000 spam messages and 1000 non-spam messages. The word "stock" occurs 400 times in the spam messages and 60 times in the non-spam. The word "undervalued" occurs in 200 spam messages and 25 non-spam.
- **Solution**: $p(stock) = 400/2000 = .2$, $q(stock) = 60/1000 = .06$, $p(undervalued) = 200/2000 = .1$, $q(undervalued) = 25/1000 = .025$

$$r(stock, undervalued) = \frac{p(stock)p(undervalued)}{p(stock)p(undervalued) + q(stock)q(undervalued)}$$

$$= \frac{(0.2)(0.1)}{(0.2)(0.1) + (0.06)(0.025)} \approx 0.930$$

If our threshold is .9, we class the message as spam and reject it.

Bayesian Spam Filters using Multiple Words

- In general, the more words we consider, the more accurate the spam filter. With the independence assumption if we consider *k* words:

$$p(S|\bigcap_{i=1}^{k} E_i) = \frac{\Pi_{i=1}^{k} p(E_i|S)}{\Pi_{i=1}^{k} p(E_1|S) + \Pi_{i=1}^{k} p(E_i|\overline{S})}$$

$$r(w_1, w_2, ...w_n) = \frac{\Pi_{i}^{k} p(w_i)}{\Pi_{i=1}^{k} p(w_i) + \Pi_{i=1}^{k} q(w_i)}$$

We can further improve the filter by considering pairs of words as a single block or certain types of strings.

Expected Value and Variance

Section 6.4

Section Summary

- Expected Value
- Linearity of Expectations
- Average-Case Computational Complexity
- Geometric Distribution
- Independent Random Variables
- Variance
- Chebyshev’s Inequality

Expected Value

- **Definition**: The *expected value* (or *expectation* or *mean*) of the random variable *X*(*s*) on the sample space *S* is equal to
$$E(X) = \sum_{x \in S} p(s)X(s).$$
- **Example-Expected Value of a Die**: Let X be the number that comes up when a fair die is rolled. What is the expected value of *X*?
- **Solution**: The random variable X takes the values 1, 2, 3, 4, 5, or 6. Each has probability 1/6. It follows that
$$E(X) = \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \cdots \frac{1}{6} \cdot 6 = \frac{21}{6} = \frac{7}{2}.$$

Expected Value

- **Theorem 1**: If X is a random variable and $p(X = r)$ is the probability that X = r, so that

$$p(X = r) = \sum_{s \in S, X(s) = r} p(s), \quad \text{then}$$

$$E(X) = \sum_{r \in X(S)} p(X = r) r.$$

- **Proof**: Suppose that X is a random variable with range $X(S)$ and let $p(X = r)$ be the probability that X takes the value r. Consequently, $p(X = r)$ is the sum of the probabilities of the outcomes s such that $X(s) = r$. Hence,

$$E(X) = \sum_{r \in X(S)} p(X = r) r.$$

◀

Expected Value

- **Theorem 2**: The expected number of successes when *n* mutually independent Bernoulli trials are performed, where, the probability of success on each trial, $p = np$.
- **Proof**: Let *X* be the random variable equal to the number of success in *n* trials. By Theorem 2 of section 7.2, $p(X = k) = C(n,k)p^k q^{n-k}$. Hence,

$$E(X) = \sum_{k=1}^{n} kp(X = k)$$ by Theorem 1

continued →

Expected Value

$$E(X) = \sum_{k=1}^{n} kp(X = k)$$ from previous page

$$= \sum_{k=1}^{n} kC(n,k)p^k q^{n-k}$$ by Theorem 2 in Section 7.2

$$= \sum_{k=1}^{n} nC(n-1,k-1)p^k q^{n-k}$$ by Exercise 21 in Section 6.4

$$= np \sum_{k=1}^{n} C(n-1,k-1)p^{k-1} q^{n-k}$$ factoring *np* from each term

$$= np \sum_{j=0}^{n-1} C(n-1,j)p^j q^{n-1-j}$$ shifting index of summation with $j = k - 1$

$$= np(p + q)^{n-1}$$ by the binomial theorem

$$= np.$$ because $p + q = 1$

- We see that the expected number of successes in *n* mutually independent Bernoulli trials is *np*.

Linearity of Expectations

- The following theorem tells us that expected values are linear. For example, the expected value of the sum of random variables is the sum of their expected values.
- **Theorem 3**: If X_i, $i = 1, 2, \ldots, n$ with n a positive integer, are random variables on S, and if a and b are real numbers, then
  - (i) $E(X_1 + X_2 + \ldots + X_n) = E(X_1) + E(X_2) + \ldots + E(X_n)$
  - (ii) $E(aX + b) = aE(X) + b$.
  - *See the text for the proof*

Linearity of Expectations

- **Expected value in the hatcheck problem**: A new employee started a job checking hats, but forgot to put the claim check numbers on the hats. So, the *n* customers just receive a random hat from those remaining. What is the expected number of hat returned correctly?
- **Solution**: Let X be the random variable that equals the number of people who receive the correct hat. Note that $X = X_1 + X_2 + \cdots + X_n$, where $X_i = 1$ if the i-th person receives the hat and $X_i = 0$ otherwise.
  - Because it is equally likely that the checker returns any of the hats to the ith person, it follows that the probability that the ith person receives the correct hat is $1/n$. Consequently (by Theorem 1), for all i

    $$E(X_i) = 1 \cdot p(X_i = 1) + 0 \cdot p(X_i = 0) = 1 \cdot 1/n + 0 = 1/n .$$
  - By the linearity of expectations (Theorem 3), it follows that:

    $$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n) = n \cdot 1/n = 1.$$
- Consequently, the average number of people who receive the correct hat is exactly 1. (Surprisingly, this answer remains the same no matter how many people have checked their hats!)

Linearity of Expectations

- **Expected number of inversions in permutation**: The ordered pair (*i*,*j*) is an *inversion* in a permutation of the first *n* positive integers if *i* < *j*, but *j* precedes *i* in the permutation.
- **Example**: There are six inversions in the permutation of 3, 5, 1, 4, 2

  (1, 3), (1, 5), (2, 3), (2, 4), (2, 5), (4, 5).

  Find the average number of inversions in a random permutation of the first *n* integers.
- **Solution**: Let $I_{i,j}$ be the RV on the set of all permutations of the first n positive integers with $I_{i,j} = 1$ if (i,j) is an inversion of the permutation and $I_{i,j} = 0$ otherwise. If X is another RV equal to the number of inversions in the permutation, then

$$X = \sum_{1 \le i < j \le n} I_{i,j}.$$

  - Since it is equally likely for *i* to precede j in a randomly chosen permutation as it is for *j* to precede *i*, we have: $E(I_{i,j}) = 1 \cdot p(I_{i,j} = 1) + 0 \cdot p(I_{i,j} = 0) = 1 \cdot 1/2 + 0 = ½$, for all (i,j).
  - Because there are $\binom{n}{2}$ pairs *i* and *j* with $1 \le i < j \le n$, by the linearity of expectations (Theorem 3),

$$E(X) = \sum_{1 \le i < j \le n} E(I_{i,j}) = \binom{n}{2} \cdot \frac{1}{2} = \frac{n-1}{2} \cdot \frac{1}{2}.$$

- Consequently, it follows that there is an average of $n(n-1)/4$ inversions in a random permutation of the first *n* positive integers.

Average-Case Computational Complexity

- The average-case computational complexity of an algorithm can be found by computing the expected value of a random variable.
- Let the sample space of an experiment be the set of possible inputs a_j, $j = 1, 2, \ldots, n$, and let the random variable X be the assignment to a_j of the number of operations used by the algorithm when given a_j as input.
- Assign a probability $p(a_j)$ to each possible input value a_j.
- The expected value of X is the average-case computational complexity of the algorithm.

$$E(X) = \sum_{j=1}^{n} p(a_j)X(a_j).$$

Average-Case Complexity of Linear Search

- What is the average-case complexity of linear search (described in Chapter 3) if the probability that x is in the list is p and it is equally likely that x is any of the n elements of the list?

procedure *linear search*(x: integer, $a_1, a_2, \ldots, a_n$: distinct integers)
$i := 1$
while ($i \leq n$ and $x \neq a_i$)
    $i := i + 1$
    if $i \leq n$ **then** *location* := i
        else *location* := 0
return *location*{*location* is the subscript of the term that equals x, or is 0 if x is not found}

continued →

Average-Case Complexity of Linear Search

- **Solution**: There are $n + 1$ possible types of input: one type for each of the n numbers on the list and one additional type for the numbers not on the list. Recall that:
  - $2i + 1$ comparisons are needed if x equals the i-th element of the list.
  - $2n + 2$ comparisons are used if x is not on the list.
- The probability that x equals a_i is p/n and the probability that x is not in the list is $q = 1 - p$. The average-case case computational complexity of the linear search algorithm is:

$$
\begin{aligned}
E &= 3p/n + 5p/n + \ldots + (2n + 1)p/n + (2n + 2)q \\
&= (p/n)(3 + 5 + \ldots + (2n + 1)) + (2n + 2)q \\
&= (p/n)((n + 1)^2 - 1) + (2n + 2)q \quad (\textit{Example 2 from Section } 5.1) \\
&= p(n + 2) + (2n + 2)q.
\end{aligned}
$$

  - When x is guaranteed to be in the list, p = 1, $q = 0$, so that $E = n + 2$.
  - When p is ½ and $q = ½$, then E = $(n + 2)/2 + n + 1 = (3n + 4)/2$.
  - When p is ¾ and $q = ¼$ then E = $(n + 2)/4 + (n + 1)/2 = (5n + 8)/4$.
  - When x is guaranteed not to be in the list, p = 0, $q = 1$, then $E = 2n + 2$.

The Geometric Distribution

- **Example**: Suppose the probability that a coin comes up tails is *p*. What is the expected number of flips until this coin comes up tails?
- **Solution**:
  - The sample space is {T, HT, HHT, HHHT, HHHHT, …}.
  - We know the probability p(T) = p, p(HT) = (1-p)p, p(HHT) = $(1-p)^2$p
  - Let *X* be the random variable equal to the number of flips in an element of the sample space; *X*(*T*) = 1, *X*(*HT*) = 2, *X*(*HHT*) = 3, etc.
  - Form Theorem 1 of expected value,

$$E(X) = \sum_{j=1}^{\infty} j \cdot p(X = j) = \sum_{j=1}^{\infty} j(1-p)^{j-1}p = p\sum_{j=1}^{\infty} j(1-p)^{j-1}$$

$$= \frac{p}{(1-(1-p))^2} = \frac{1}{p}$$ (from table 2 in section 2.4)

The Geometric Distribution

- **Definition 2**: A random variable X has *geometric distribution with parameter* p if $p(X = k) = (1 - p)^{k-1}p$ for $k = 1,2,3,\ldots$, where p is a real number with $0 \leq p \leq 1$.

- **Theorem 4**: If the random variable X has the geometric distribution with parameter p, then $E(X) = 1/p$.

Independent Random Variables

- **Definition 3**: The random variables X and Y on a sample space S are independent if

  $p(X = r_1 \text{ and } Y = r_2) = p(X = r_1) \cdot p(Y = r_2)$.

- **Theorem 5**: If X and Y are independent variables on a sample space S, then $E(XY) = E(X)E(Y)$.

Variance

- **Deviation**: The *deviation* of X at s ∈ S is $X(s) - E(X)$, the difference between the value of X and the mean of X.
- **Definition 4**: Let X be a random variable on the sample space S. The *variance* of X, denoted by V(X) is

$$V(X) = \sum_{s \in S} (X(s) - E(X))^2 p(s).$$

- That is $V(X)$ is the weighted average of the square of the deviation of X. The standard deviation of X, denoted by $\sigma(X)$ is defined to be $\sqrt{V(X)}$

Variance

- **Theorem 6**: If X is a random variable on a sample space S, then $V(X) = E(X^2) - E(X)^2$.
  - See text for the proof
- **Corollary 1**: If X is a random variable on a sample space S and $E(X) = \mu$, then $V(X) = E((X - \mu)^2)$.
  - See text for the proof
- **Example:** What is the variance of the random variable X, where $X(t) = 1$ if a Bernoulli trial is a success and $X(t) = 0$ if it is a failure, where p is the probability of success and q is the probability of failure?
- **Solution**: Because X takes only the values 0 and 1, it follows that $X^2(t) = X(t)$. Hence,

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p) = pq.$$

Variance

- **Variance of the value of a die:** What is the variance of a random variable X, where X is the number that comes up when a fair die is rolled?
- **Solution**: We have $V(X) = E(X^2) - E(X)^2$.

  $E(X) = 1/6(1 + 2 + 3 + 4 + 5 + 6) = 7/2$

  $E(X^2) = 1/6(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2) = 91/6.$

  We conclude that

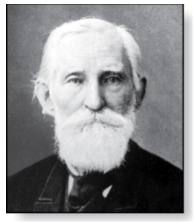
$$V(X) = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Variance

Irenée-Jules Bienaymé
(1796-1878)

- **Bienaymé's Formula**: If X and Y are two independent random variables on a sample space S, then $V(X + Y) = V(X) + V(Y)$. Furthermore, if X_i, $i = 1,2, \ldots,n$, with n a positive integer, are pairwise independent random variables on S, then
$$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n).$$
  - *see text for the proof*
- **Example**: Find the variance of the number of successes when n independent Bernoulli trials are performed, where on each trial, p is the probability of success and q is the probability of failure.
- **Solution**: Let X_i be the random variable with $X_i((t_1, t_2, \ldots., t_n)) = 1$ if trial t_i is a success and $X_i((t_1, t_2, \ldots., t_n)) = 0$ if it is a failure. Let $X = X_1 + X_2 + \ldots. X_n$. Then X counts the number of successes in the n trials.
  - By Bienaymé 's Formula, it follows that $V(X) = V(X_1) + V(X_2) + \cdots + V(X_n)$.
  - By the previous example ,$V(X_i) = pq$ for $i = 1,2, \ldots,n$.
- Hence, $V(X) = npq$.

Chebyshev's Inequality



Pafnuty Lvovich Chebyshev (1821-1894)

- **Chebyschev's Inequality**: Let X be a random variable on a sample space S with probability function p. If r is a positive real number, then
$$p(|X(s) - E(X)| \geq r) \leq V(X)/r^2.$$
  - *See text for the proof*
- **Example**: Suppose that X is a random variable that counts the number of tails when a fair coin is tossed n times. Note that X is the number of successes when n independent Bernoulli trials, each with probability of success ½ are done. Hence, (by Theorem 2) $E(X) = n/2$ and (by Example 18) $V(X) = n/4$.
- By Chebyschev's inequality with $r = \sqrt{n}$,
$$p(|X(s) - n/2| \geq \sqrt{n}) \leq (n/4)(\sqrt{n})^2 = ¼.$$
- This means that the probability that the number of tails that come up on n tosses deviates from the mean, $n/2$, by more than $\sqrt{n}$ is no larger than ¼.