A Little Probability

. . .

. . .

Coding and Information Theory

Fall, 2004

Tom Carter

http://astarte.csustan.edu/~ tom/

October, 2004

Some probability background

- There are two notions of the *probability* of an event happening. The two general notions are:

  1. A *frequentist* version of probability:

     In this version, we assume we have a set of possible events, each of which we assume occurs some number of times. Thus, if there are N distinct possible events $(x_1, x_2, \ldots, x_N)$, no two of which can occur simultaneously, and the events occur with frequencies $(n_1, n_2, \ldots, n_N)$, we say that the probability of event x_i is given by

     $$P(x_i) = \frac{n_i}{\sum_{j=1}^{N} n_j}$$

     This definition has the nice property that

     $$\sum_{i=1}^{N} P(x_i) = 1$$

2. An *observer relative* version of probability:

   In this version, we take a statement of *probability* to be an assertion about the belief that a specific observer has of the occurrence of a specific event.

   Note that in this version of *probability*, it is possible that two different observers may assign different probabilities to the same event.

   Furthermore, the *probability* of an event, for me, is likely to change as I learn more about the event, or the context of the event.

3. In some (possibly many) cases, we may be able to find a reasonable correspondence between these two views of probability. In particular, we may sometimes be able to understand the *observer relative* version of the probability of an event to be an approximation to the *frequentist* version, and to view new knowledge as providing us a better estimate of the relative frequencies.

- I won't go through much, but some probability basics, where a and b are events:
  $P(not\ a) = 1 - P(a)$.
  $P(a\ or\ b) = P(a) + P(b) - P(a$ and $b)$.
  We will often denote $P(a$ and $b)$ by $P(a, b)$. If $P(a, b) = 0$, we say a and b are mutually exclusive.

- Conditional probability:

  $P(a|b)$ is the probability of a, given that we know b. The joint probability of both a and b is given by:

  $$P(a, b) = P(a|b)P(b).$$

  Since $P(a, b) = P(b, a)$, we have Bayes' Theorem:

  $$P(a|b)P(b) = P(b|a)P(a),$$

  or

  $$P(a|b) = \frac{P(b|a)P(a)}{P(b)}.$$

- If two events a and b are such that

$$P(a|b) = P(a),$$

we say that the events a and b are *independent*. Note that from Bayes' Theorem, we will also have that

$$P(b|a) = P(b),$$

and furthermore,

$$P(a, b) = P(a|b)P(b) = P(a)P(b).$$

This last equation is often taken as the definition of *independence*.

- A quick example:
  Suppose that you are asked by the government to help them understand the results of a "terrorist screening system" they are developing. They have been told that the system is 99.9% accurate. What is the probability that when the system identifies a potential "terrorist" that they have actually found one?

- You do some research, and find out that independent estimates put the number of actual "terrorists" in the US at around 250. The creators of the system assert that the system is 99.9% accurate. You push the question, and find that they say that one tenth of one percent of the time, the test falsely clears someone who is a "terrorist", and one tenth of one percent of the time the system falsely reports someone to be a "terrorist" when they are not. If the system identifies someone as a "terrorist," how seriously should the government take the identification? Given this much information, what can you calculate as the probability the individual is a "terrorist"?

In general, there are four possible situations for an individual being identified:

1. Test positive (Tp), and are a "terrorist" (T).

2. Test negative (Tn), and are not a "terrorist" (NT).

3. Test positive (Tp), and are not a "terrorist" (NT).

4. Test negative (Tn), and are a "terrorist" (T).

We would like to calculate the probability someone is a "terrorist" (T) given, that they have been identified as such by the system (Tp):

$$P(T|Tp).$$

We can do this using Bayes' Theorem.

We can calculate:

$$P(T|Tp) \ = \ \frac{P(Tp|T) * P(T)}{P(Tp)}.$$

We need to figure out the three items on the right side of the equation. We can do this by using the information given.

Suppose the screening test was done on 250,000,000 people in the US. Out of these $2.5 * 10^8$ people, we expect there to be 250 people who are "terrorists", and 249,999,750 people who are not. According to the creators of the system, we would expect the test results to be:

- Test positive (Tp), and are "terrorists" (T): 250 people.

- Test negative (Tn), and are not "terrorists" (NT):

  $0.999 * 249,999,750 = 249,749,750$ people.

- Test positive (Tp), and are not "terrorists" (NT):

  $0.001 * 249,999,750 = 250,000$ people.

- Test negative (Tn), and are "terrorists" (T): 0 people.

Now let's put the the pieces together:

$$
\begin{aligned}
P(T) &= \frac{250}{250,000,000} \\
&= 10^{-6} \\
P(Tp) &= \frac{250 + 250,000}{250,000,000} \\
&= \frac{250,250}{250,000,000} \\
&= 0.001001 \\
P(Tp|T) &= 0.999
\end{aligned}
$$

Thus, our calculated probability that someone identified as a "terrorist" actually is one:

$$\begin{aligned} P(T|Tp) &= \frac{P(Tp|T) * P(T)}{P(Tp)} \\ &= \frac{0.999 * 10^{-6}}{0.001001} \\ &= \frac{9.99 * 10^{-7}}{1.001 * 10^{-3}} \\ &= 9.98002 * 10^{-4} \\ &< 10^{-3} = .001 \end{aligned}$$

In other words, an individual identified by the system as a "terrorist", with a test that is promised to be 99.9% correct, has less that one chance in 1000 of actually being one! Another way of saying it is that for every one "terrorist" that is actually identified, 1000 innocent people are incorrectly identified as being one.

- There are a variety of questions we could ask now, such as, how accurate would the system have to be for there to be a greater than 50% probability that someone identified as a "terrorist" actually is one?

  For this, we need fewer false positives than true positives. Thus, in the example, we would need fewer than 250 false positives out of the 249,999,750 people who are not. In other words, the proportion of those who are not "terrorists" for whom the system would have to be correct would need to be greater than:

$$\frac{249,999,500}{249,999,750} = 99.9999\%!!$$

- Another question we could ask is, "How prevalent would "terrorists" have to be in order for a 99.9% accurate test (0.1% false positive and 0.1% false negative) to give a greater than 50% probability of actually being a "terrorist" when identified as one?"

  Again, we need fewer false positives than true positives. We would therefore need the actual occurrence to be greater than 1 in 1000 (each false positive would be matched by at least one true positive, on average) – in other words, there would have to be about 250,000 "terrorists" in the US!

- Another example: consider another situation, with a test that is not so accurate. Suppose the test were 80% accurate (20% false positive and 20% false negative). Suppose that we are testing for a condition expected to affect 1 person in 100. What would be the probability that a person testing positive actually has the condition?

  We can do the same sort of calculations.

Let's use 1000 people this time. Out of this sample, we would expect 10 to have the condition.

- Test positive (Tp), and have the condition (Ha):

$$0.80 * 10 = 8 \text{ people.}$$

- Test negative (Tn), and don't have the condition (Na):

$$0.80 * 990 = 792 \text{ people.}$$

- Test positive (Tp), and don't have the condition (Na):

$$0.20 * 990 = 198 \text{ people.}$$

- Test negative (Tn), and have the condition (Ha):

$$0.20 * 10 = 2 \text{ people.}$$

Now let's put the the pieces together:

$$\begin{aligned} P(Ha) &= \frac{1}{100} \\ &= 10^{-2} \\ P(Tp) &= \frac{8+198}{10^3} \\ &= \frac{206}{10^3} \\ &= 0.206 \\ P(Tp|Ha) &= 0.80 \end{aligned}$$

Thus, our calculated probability that a person testing positive actually has the condition is:

$$\begin{aligned} P(Ha|Tp) &= \frac{P(Tp|Ha) * P(Ha)}{P(Tp)} \\ &= \frac{0.80 * 10^{-2}}{0.206} \\ &= \frac{8 * 10^{-3}}{2.06 * 10^{-1}} \\ &= 3.883495 * 10^{-2} \\ &< .04 \end{aligned}$$

In other words, one who has tested *positive*, with a test that is 80% correct, has less that one chance in 25 of actually having this condition. (Imagine for a moment, for example, that this is a drug test being used on employees of some corporation . . .)

- We could ask the same kinds of questions we asked before:

  1. How accurate would the test have to be to get a better than 50% chance of actually having the condition when testing positive?

     (99%)

  2. For an 80% accurate test, how frequent would the condition have to be to get a better than 50% chance?

     (1 in 5)

- Some questions:

  1. Are these examples realistic? If not, why not?

  2. What sorts of things could we do to improve our results?

  3. Would it help to repeat the test? For example, if the probability of a false positive is 1 in 100, would that mean that the probability of two false positives on the same person would be 1 in 10,000 ($\frac{1}{100} * \frac{1}{100}$)? If not, why not?

  4. In the case of a medical condition such as a genetic anomaly, it is likely that the test would not be applied randomly, but would only be ordered if there were other symptoms suggesting the anomaly. How would this affect the results?