CS 6604: Data Mining Fall 2007

Lecture 17 — Wednesday, October 17, 2007

Lecture: Naren Ramakrishnan *Scribe: Don Conry*

1 Review

As a review of some basic concepts in probability, it is convenient to define a few terms:

1. A *sample space* is the set of all possible outcomes of a trial or experiment. For a die roll, the sample space would be the set $\{1, 2, 3, 4, 5, 6\}$.
2. A *σ-algebra* of a set X is a collection of subsets of X, and is closed under the complement and union set operators. For example, a trivial σ-algebra for any set X is the set containing the empty set and X itself.
3. A *random variable* is a variable that takes on values from a sample space, based on some distribution. These variables can represent a random trial (such as a die toss) and its results.
4. A *distribution* defines a probability or likelihood that a random variable takes each of the values from a sample space.

2 Probability

Probability can be defined as a "measure" over a random variable with associated sample space and σ-algebra; however there are several schools of thought on precisely how to define this term. Two main schools are frequentists and subjectivists (or Bayesians). Frequentists (due to work by Neyman and Pearson) view probability simply as the ratio of the number of times an event occurs to all instances of a trial; in other words:

$$P(event) = \frac{\textbf{\# of occurrences of an event}}{\textbf{\# of all occurrences}} \tag{1}$$

A limitation of this view is that such an analysis requires a *repeatable* event. Unfortunately, some events are not easily repeated (e.g., the probability that Hillary Clinton will be president).

2.1 Subjectivists & Bayes' rule

Subjectivism is based on work by Thomas Bayes, and models probability on the idea of *belief*. Bayesian probability is defined as the degree of belief in a statement; this belief is subsequently modified as events are observed. This approach has the advantage that no prior knowledge about an event's history is required to state a probability. For most things we've seen so far, results from frequentists and subjectivists would agree. For data mining, we'll be focusing on subjectivism, as it is more interesting.

In Bayes' work, the notion of belief is identified by the *prior* and the *posterior*, while *likelihood* is concerned with data. Recall from probability that for two events A & B, $P(A, B) = P(A) \times P(B|A) = P(B) \times P(A|B)$. From this equality, Bayes derived the following rule:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \tag{2}$$

The reason why Bayes' rule is so useful is because it turns out that some conditional probabilities are easier to compute than others. Often the easier computation is fed into Bayes' rule to avoid the more difficult computation. Additionally, many times a comparison between two conditional probabilities is required: in these cases, the $P(A)$ denominator is common to both equations, and therefore divides out of the comparison ratio. For this reason, it is common to generalize and say that $P(B|A)$ is proportional to the numerator, written as $P(B|A) \propto P(B) \times P(A|B)$. In this equation, $P(B|A)$ is considered the *posterior*, $P(B)$ is the *prior*, and $P(A|B)$ is the *likelihood*. For data mining, generally the B parameter is considered the *hypothesis*, while the A parameter is the *data*, and Bayes' rule can be written as $P(\text{hypothesis}|\text{data}) \propto P(\text{hypothesis}) \times P(\text{data}|\text{hypothesis})$.

2.2 An example: coin flip

Consider a coin flip trial. Let it be parameterized by the probability of a heads. To find the probability of tails, we can simply subtract the probability of heads from one, since the two sides are complementary events. A prior over the probability of heads can be represented as a distribution, for example:

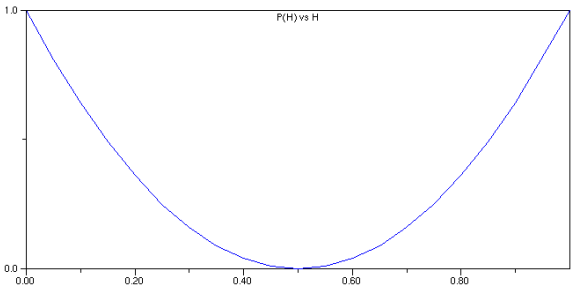

Figure 1: A loaded coin.

The coin with a prior distribution as in Fig 1 is a *loaded* coin, i.e. very likely to come up heads most of the time, or tails most of the time. Observe the scale on the y-axis; since we are only interested in a 'proportional' value, the curve doesn't quite integrate to 1.

Also recall that we are dealing with continuous-valued variables, so it is improper to ask for the probability of heads having a specific value (such as zero); it is more pertinent to deal with ranges of values instead.

Let us consider a uniform prior indicating complete ignorance, i.e., $P(H)$ is defined as 1 for $0 <= H <= 1$ and 0 otherwise, as shown in Fig 2. This initial prior must now be updated with incoming data. Suppose the first coin flip comes up heads; then $P(\text{data}|\text{hypothesis})$ is H. A new hypothesis is formed as follows: $P(\text{hypothesis}|\text{data}) \propto P(\text{hypothesis}) \times P(\text{data}|\text{hypothesis}) = 1 \times H = H$. For the second flip that comes up tails, this probability is $(1 - H)$. Therefore, $P(\text{hypothesis}|\text{data}) \propto P(\text{hypothesis}) \times P(\text{data}|\text{hypothesis}) = H \times (1 - H) = H(1 - H)$. In general, after N total flips with R of these flips coming up heads, the hypothesis becomes $H^R(1 - H)^{N-R}$.

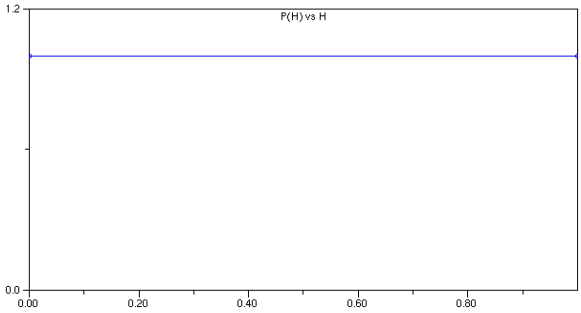

Figure 2: A simple distribution as a prior.

Keeping in mind that the function values are proportional (i.e. not normalized), any initial prior can be used, and the resulting hypotheses will converge to the correct distribution. The H-value that maximizes P can be found by setting the derivative of P to be 0 and solving for H. We can simplify the calculation by instead taking the derivative of $\log P$ (since this is a monotone function, the maximum value will occur at the same H-value as the original function). Since $\log P = R \log H + (N - R) \log (1 - H)$, the derivative is $\frac{R}{H} - \frac{(N-R)}{(1-H)}$; setting this to 0 yields $\frac{R}{H} - \frac{(N-R)}{(1-H)} = 0$, $\frac{R}{H} = \frac{(N-R)}{(1-H)}$, $R - RH = HN - RH$, $R = HN$, and finally $H = \frac{R}{N}$. The second derivative should also be verified as greater than 0 at this point, to ensure the point is a maximum and not a minimum.

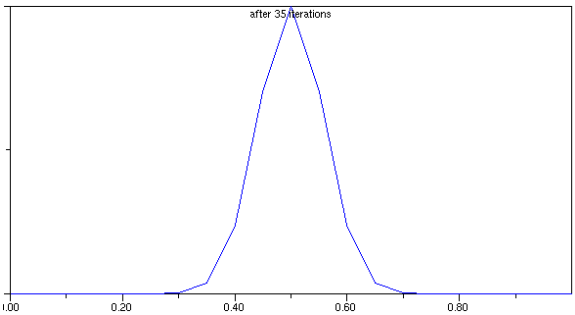

Figure 3: The improved prior curve after 70 flips, 35 heads & 35 tails.

2.3 Variance

To evaluate the variance of the prior curve, we can use the first three terms of the Taylor series expanded around the maximum at $H = R/N$. Consider the formula for these first three terms:

$$f(x) = f(\frac{R}{N}) + \frac{df}{dx}(\frac{R}{N}) * (x - \frac{R}{N}) + \frac{d^2 f}{dx^2}(\frac{R}{N}) * \frac{(x - \frac{R}{N})^2}{2} \tag{3}$$

Note that the 2nd term goes away (since the derivative of P is 0 at $H = \frac{R}{N}$) and we are left with a Gaussian function, i.e. the result is $\frac{\frac{R}{N}(1-\frac{R}{N})}{N}$ which is proportional to $\frac{1}{\sqrt{2\pi} * \sigma} * (e^{\frac{-(x-N)^2}{2\sigma^2}})$. Since the mean of the original function $\mu = N/R$, the variance can be written as $var(N) = \frac{\mu(1-\mu)}{N}$. Consequently, it requires

more data to determine that the coin is fair than it does to find it unfair.

3 Predicting the future

How do we use the posterior distributions? For instance, is a head more likely or a tail more likely? For a frequentist, the approach is to find a point at which the curve is maximum and see what it corresponds to. A Bayesian needs to integrate over all H-values $P(\text{heads}|H) * P(H)$. This latter method gives all of the H-values some degree of influence (weighted by their individual probabilities) on the final prediction; this is called *marginalization* or integrating out nuisance parameters (in this case, the variable H).