Naive Bayes

Alex S.*

1 Introduction

Probability is a tricky word—usually meaning the likelyhood of something occuring—or how frequent something is. Obviously, if something happens frequently, then its probability of happening is high. As you may have guessed, coming up with non-circular definition of probability (that doesn't involve 'frequency' or 'likelyhood') is an exercise in futility, and you shouldn't lose sleep over it (on second though, losing sleep over something philosophical may not be a total waste).

1.1 Basics

Suppose we have a set of events, $\Gamma = \{e_1, e_2, \ldots, e_n\}$. A probability function $P(E)$ assigns a real number, 0 to 1, to each subset $E \subseteq \Gamma$. Naturally then, $P(\Gamma)$ is 1, and $0 \leq P(E) \leq 1$ for every $E \subseteq \Gamma$.

For dice[1], $\Gamma = \{1, 2, 3, 4, 5, 6\}$, since any single throw can land on some number 1 through 6. Throwing a single die isn't very fun—each number has an equal chance of showing up, namely 1 out of 6. Now, consider throwing two: the outcomes may be:

$$
\begin{aligned}
2 &= \{1,1\} \\
3 &= \{1,2\} \text{ or } \{2,1\} \\
4 &= \{1,3\} \text{ or } \{2,2\} \text{ or } \{3,1\} \\
5 &= \{1,4\} \text{ or } \{2,3\} \text{ or } \{3,2\} \text{ or } \{4,1\} \\
6 &= \{1,5\} \text{ or } \{2,4\} \text{ or } \{3,3\} \text{ or } \{4,2\} \text{ or } \{5,1\} \\
7 &= \{1,6\} \text{ or } \{2,5\} \text{ or } \{3,4\} \text{ or } \{4,3\} \text{ or } \{5,2\} \text{ or } \{6,1\} \\
8 &= \{2,6\} \text{ or } \{3,5\} \text{ or } \{4,4\} \text{ or } \{5,3\} \text{ or } \{6,2\} \\
9 &= \{3,6\} \text{ or } \{4,5\} \text{ or } \{5,4\} \text{ or } \{6,3\} \\
10 &= \{4,6\} \text{ or } \{5,5\} \text{ or } \{6,4\} \\
11 &= \{5,6\} \text{ or } \{6,5\} \\
12 &= \{6,6\}
\end{aligned}
$$

*alex@theparticle.com

[1]Small cubes used in gaming or in determining by chance.

That sure is a lot of outcomes; 36 in fact. Now, each subset has a 1 in 36 chance of occuring. For example, if you throw two dice, your chances of getting a "2", or $P(2)$ are 1/36. Your chances of getting "11", or $P(11)$ are 2/36 (since there are two subsets that add up to 11, namely, $\{5,6\}$ and $\{6,5\}$). What about $P(7)$? We can get that any number of ways:

$$\{1,6\} \text{ or } \{2,5\} \text{ or } \{3,4\} \text{ or } \{4,3\} \text{ or } \{5,2\} \text{ or } \{6,1\}$$

There are six ways of getting a "7". Each one of those has a 1/36 chance of coming up, thus $P(7) = 6/36$.

1.2 Set Operations

For subsets $E, F \subseteq \Gamma$ we have:

$$P(E \cup F) = P(E) + P(F) - P(E \cap F)$$

Going back to the dice example: what are the chances of us throwing a "7" where one of the die comes up as "1"? Well, what's the probability of one of the die coming up "1"? Here are the possibilities:

$$\{1,1\},\{1,2\},\{1,3\},\{1,4\},\{1,5\},\{1,6\},\{2,1\},\{3,1\},\{4,1\},\{5,1\},\{6,1\}$$

That makes 11/36. We already know that chance of getting a "7" is 6/36. To add the probabilities gets us:

$$P(one\_die\_is\_1) + P(7) = 11/36 + 6/36 = 17/36$$

But wait... we counted some of them twice. $\{1,6\}$ and $\{6,1\}$ show up for both $P(one\_die\_is\_1)$ and $P(7)$, so we must subtract them... So the end result is:

$$P(one\_die\_is\_1) + P(7) - P(\{1,6\} \text{ or } \{6,1\}) = 11/36 + 6/36 - 2/36 = 15/36$$

1.3 Conditional

Events tend to occur one after the other. A *conditional* probability captures just that. $P(E|F)$ represents the probability of E given F. In other words, if F has already happened, what is the probability that E will happen? A simple relation for it is

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

obviously events can be *independent* if say $P(E|F) = P(E)$, or when probabilities are zero.

1.4 Bayes Rule

Thomas Bayes (1702-1761) gave rise to a new form of statistical reasoning—the inversion of probabilities. We can view it as

$$Posterior = Prior \times Likelihood$$

where *Posterior* is the probability that the *hypothesis* is true given the evidence. *Prior* is the probability that the hypothesis was true *before* the evidense (ie: an assumption). *Likelihood* is the probability of obtaining the observed evidense given that the hypothesis is true.

We can write it as:

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

It turns out to be pretty easy to derive, ie:

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

we multiply each side by its respective denominator:

$$P(E|F)P(F) = P(E \cap F) \qquad P(F|E)P(E) = P(F \cap E)$$

or in other words,

$$P(E|F)P(F) = P(F|E)P(E)$$

then we just divide by $P(F)$, and get

$$P(E|F) = \frac{P(F|E)P(E)}{P(F)}$$

Neat, no?

2 Naive Bayes Classifier

SEE: http://en.wikipedia.org/wiki/Naive_Bayes_classifier

We can use the Bayes rule to do document classification—commonly used to classify emails into spam/nospam categories. For this to work, we need (either assumed, or calculated) prior probabilities of certain word occuring in a certain document category, ie:

$$P(w_i|C)$$

where w_i is some word, and C is some document category (ie: spam, nospam, etc.). The probability of a given document D given a certain document category is:

$$P(D|C) = \Pi_i P(w_i|C)$$

note that none of the probabilities can be zero, otherwise the whole thing is zero. In practice, this is usually accomplished by specifying a very small number as the minimum probability (even if the word doesn't exist in a particular category).

Now we do that Bayes thing:

$$P(D|C) = \frac{P(D \cap C)}{P(C)} \qquad P(C|D) = \frac{P(D \cap C)}{P(D)}$$

we multiply each side by its respective denominator:

$$P(D|C)P(C) = P(D \cap C) \qquad P(C|D)P(D) = P(D \cap C)$$

or in other words,

$$P(C|D)P(D) = P(D|C)P(C)$$

then we just divide by $P(D)$, and get

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)}$$