Chris Piech
CS109

Lecture Handout #4
April 10th, 2017

Conditional Probability and Bayes Theorem

An all knowing computer would be able to store what we call the "joint" probability of all possible combinations of events. That is not feasible.

Conditional Probability

In English, a conditional probability states "what is the chance of an event E happening given that I have already observed some other event F". It is a critical idea in machine learning and probability because it allows us to update our beliefs in the face of new evidence. The definition for calculating conditional probability is:

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

This equation implies that: $P(EF) = P(E|F)P(F)$ which we call the Chain Rule. Intuitively it states that the probability of observing events E and F is the probability of observing F, multiplied by the probability of observing E, given that you have observed F.

Here is the general form of the Chain rule:

$$P(E_1E_2\dots E_n) = P(E_1)P(E_2|E_1)\dots P(E_n|E_1\dots E_{n-1})$$

In the case where the sample space has equally likely outcomes:

$$P(E|F) = \frac{\text{\#outcomes in E consistent with F}}{\text{\#outcomes in S consistent with F}} = \frac{|EF|}{|SF|} = \frac{|EF|}{|F|}$$

Bayes Theorem

Very often we know a conditional probability in one direction, say P(E—F), but we would like to know the conditional probability in the other direction. Bayes Theorem provides a way to convert from one to the other. There are a lot of things called Bayes Theorem. Here are the two most common equations:

Most Common Version:

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

Expanded Version:

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

Conditional Probability Example 1

Machine Learning (sometimes called Data Science) is the love child of: probability, data and computers. Sometimes machine learning involves complex algorithms. But often it's just the core ideas of probability applied to large datasets.

As an example let us consider Netflix, a company that has thrived because of well thought out machine learning. One of the primary probabilities that they calculate is the probability that a user will watch a given movie given no other information about the user. We call this the prior.

Let E be the event that a user watches a given movie. We can approximate P(E) using the definition of probability from Friday's class:

$$P(E) = \lim_{n \to \infty} \frac{n(E)}{n}$$

Using this definition, we can approximate the probability by counting the number of users who watched the movie divided by the number of users who are on Netflix. Since the number of users on Netflix is huge, this is a good approximation.

Another probability that Netflix cares about is the conditional probability of a user watching a movie Life is Beautiful (call that event E) given that they have already watched another movie Amelie (Event F). Combining the definition of conditional probability with the limit approximation of probability we get:

$$\begin{aligned} P(E|F) &= \frac{P(EF)}{P(F)} = \frac{\frac{\text{\#people who watched both}}{\text{\#people on Netflix}}}{\frac{\text{\#people who watched } F}{\text{\#people on Netflix}}} \\ &= \frac{\text{\#people who watched both}}{\text{\#people who watched } F} \end{aligned}$$

Conditional Probability Example 2

From our class roster we know that the probability that a student is a Sophomore is 0.28. By asking the students in the lecture hall we can observe that the probability that someone is both a Sophomore and in attendance is 0.22. What is the conditional probability that a student attends lecture, given that they are a Sophomore?

Let A be the event that a student attends lecture. Let S be the event that a student is a Sophomore. By the definition of conditional probability:

$$P(A|S) = \frac{P(AS)}{P(S)} = \frac{0.22}{0.28} = 0.79$$

Bayes Theorem Example 1

For an example of Bayes Theorem (fully worked out and with intuition) see the demo on the course website: cs109.stanford.edu/demos/naturalBayes.html

Bayes Theorem Example 2

Email programs all come in with a spam filter that calculates probabilities to decide if an incomming email is spam (fake) or ham (not fake). We know that 60% of all email is spam. We also know from experience that 90% of all spam has a forged header where-as only 20% of non spam has a forged header. What is the probability that an email is spam given that it has a forged header.

Let E be the event that an email contains a forged header. Let F be the event that an email is spam.

$$\begin{aligned} P(F|E) &= \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^C)P(F^C)} \\ &= \frac{(0.9)(0.6)}{(0.9)(0.6) + (0.2)(0.4)} \approx 0.871 \end{aligned}$$