A Tutorial on Simple Bayesian Inference and Classification Using Sampling

T. Nathan Mundhenk

University of Southern California

Department of Computer Science
**Bayesian inference using sampling**

Sampling is a very common method used to get at what is happening in a system when one does not know all the underlying parameters of it. For instance, deriving the probabilities of Black Jack does not require sampling, we know all the rules of the game and we have a small deck of countable cards. Therefore, we do not need to sample a blackjack game in order to know whether someone should stand on 20. However, many things in real life are extremely complex and we don’t know what’s really going on to cause an event to happen more often than not. We don’t know why some percent of the population dies every year from cancer. Instead, we sample the population of people and can derive your likelihood of dying from cancer. If we knew exactly what caused cancer, we wouldn’t need to sample the population. We would know that some molecule X combines with some gene Y and 25% of the time you die from cancer. If we see X and Y then we can derive your probability of dying from cancer the same way we would compute your odds at blackjack. However, we do not know these things, so what we do is take samples to try to get a better idea of what is really going on. So we do not know the real cause of people dying from cancer, but we can sample and find that for instance people who eat fish have a lower probability of dying from cancer. While the eating of fish may have nothing to do directly with cancer, somehow maybe fish oils reduce the amount of molecule X or maybe people who like to eat fish also tend not to carry gene Y. At any rate, we measure what we can see and make the most logical inference from the data we have at hand.

**Using Bayesian inference and Sampling in AI**

An AI may not know exactly what is going on in the world, but we want it to make the best decision it can with the data it has at hand. For instance, a robot can walk around the world taking samples of images, sounds, smells and whatever else it can perceive. The robot may perceive a fish smell, which may mean it is most likely in a fish market. However, it may be inland in which case it is most likely in a seafood restaurant. Robots can learn these inferences by taking samples of the world. If it smells fish 10 times and 9 times it’s in a seafood restaurant, it might make a very simple inference that 90% of the time when it smells fish, it’s in a seafood restaurant.

However, what happens if it has a simple sensor that instead of giving us a “yes/no” answer on the smell of fish, it tells us how fishy things smell. This might be much more useful. It might turn out that readings on the fish ‘o’ meter between 75 and 100 indicate a seafood restaurant, while readings between 50 and 75 indicate that it is riding on a New York subway. We can learn these rules of inference by walking around with the fish ‘o’ meter and measuring whiffs of air. We then record the reading from the fish ‘o’ meter with the location the sample was taken from. If we are careful in how we take our samples, for instance from several random locations, after we have enough samples we can infer that given a certain reading on the fish ‘o’ meter, what our location is most
likely to be. Thus, when our robot gets a reading of 84, using Bayesian inference from our samples, we might believe that our robot has a 90% chance of being in a seafood restaurant. It is important to note the word “believe” here because with sampling, one never knows for sure what the true probability is unless you are able to sample all possible samples. This is actually a more advanced topic and statisticians are always working on ways to try to figure out the reliability of their inferences. For now, we will put the reliability question aside, but we will remember that it is a very important issue.

A First Quick Easy Approach to Bayesian Inference from Sampling

There are many ways to make reasonable Bayesian inferences from samples. The easiest, and perhaps most common way to do this is by making the assumption that the underlying process is Gaussian. Putting aside the complex math of what it means for a process to be Gaussian, what it most simply means is that we believe that if we take a bunch of samples, then compute the average sample, any future samples will most likely be like the average samples. Additionally, as samples become less like the average sample, they should be less common. So for instance, if the average height of a human male is 5’9”, then most adult males should be around 5’9” and very few should be 4’9” or 7’1” tall. Here we assume that we don’t know what really makes people as tall as they are, but that the underlying process is Gaussian in nature. This allows one to use Gauss’s equations for deriving probabilities.
The cool thing about using this method is that we only need to take a bunch of samples, but if we come upon a new sample that’s not totally like any sample we have previously taken, we can still derive its probability. So for instance, if we sample the height of 50 adult males, we might not have a man who is 5’6” tall in our sample, but we can still derive the probability that an adult male will be 5’6” tall. In general, the Gaussian process works like curve fitting in that by assuming a certain shape to the distribution of our measurements, we can fill in the blanks.

To see how we do this, let’s see an example.

We have a population of Smurfs and Trolls. We have a made a Dance Club, but we only want to admit Smurfs since Trolls are well known trouble makers, and for the most part cannot dance. We have put a robot at the front door to our club and it will eject any Troll it detects entering. It will do this by looking at the height of the guest entering. We believe that Trolls are much taller than Smurfs. So we take measurements of height from several Trolls and several Smurfs. The robot’s job is to determine, given the height of the guest whether it is more likely to be a Smurf or a Troll.

(1) The first thing to do is to take samples and find the mean \( \bar{x} \) height. This is just the average taken by summing all the heights and dividing by the number of samples:

\[
\bar{x} = \frac{\sum x_i}{N}
\]

Here, \( x_i \) is each height sample and \( N \) is the number of samples. So we might have 50 Smurfs, we sum their heights and divide by 50 and we have the average height of all Smurfs.
(2) In order to get at the real distribution of heights so that we know what the probability of different heights is, we compute what is called the standard deviation. This can be thought of as the average tendency of measures to be different from the mean. If the standard deviation is very high, then the measures we have taken are very spread out. As an example, the number of fingers a human has would have a very small standard deviation since by and far, most humans have 10 fingers. A few unfortunate souls might have fewer or in some cases more fingers than 10, but its uncommon. However, a persons height could be thought of as having a much higher standard deviation since while most adult males are 5’9”, it is easy to see that many men are much taller or shorter than that. We can compute the standard deviation \( S \) as:

\[
(1.2) \quad S = \sqrt{\frac{\sum_{i} x_i^2}{N} - \bar{x}^2}
\]

Thus, we take and compute the sum of squares of the heights of all the smurfs and subtract from that the mean height of all the smurfs. If the height of the smurfs varies a whole lot, then \( S \) will turn out to be large.

(3) Once we have the mean and standard deviation, we can compute the base probability using the Gaussian probability density distribution:

\[
(1.3) \quad p(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\bar{\mu})^2}{2\sigma^2}}
\]

Thus, given mean \( \bar{x} \) and standard deviation \( S \), if we take a new measurement \( x \), then the probability of seeing this measurement is \( p(x) \). It is important to reiterate that we believe the probability to be \( p(x) \) and that we don’t necessarily know what that actual probability really is. However, if we took our samples carefully, and
the underlying process is gaussian, then (1.3) gives us a very good unbiased estimation of what the probability really is. Thus, we plug in the mean and standard deviation of the height of the smurfs we have sampled. We then plug in \( x \) as a new measurement from a new smurf. \( p(x) \) is then the probability we estimate of seeing a smurf that is \( x \) tall.

(4) We have an estimate of the probability for heights, but we need to use Bayes rule to allow us to make a clear decision about whether we believe that a guest is a smurf or a troll from its height. We can do this by computing \( \bar{x} \) and \( S \) for both trolls and smurfs then derive two \( p(x) \)’s , one for trolls and one for smurfs. We write this as \( p(x|j) \) or in this case \( p(\text{height} | \text{creature}) \). This notation can be read as “The probability of observing a sampled height given the type of creature”. Thus, \( p(2” | \text{smurf}) \) can be read as “The probability of measuring a height of 2” given that we are measuring the height of a smurf”

(5) Up till now, we have concentrated on general probability for a class. That is, we have not looked at how to compare trolls and smurfs, but have instead only looked at each individually. That is, we forgot about how many trolls there are and what the likelihood of seeing a troll really was. Instead, we looked at the distribution of the troll’s heights. In order to complete the Bayesian inference and make a good judgment, we need to know what the likelihood in general of seeing a troll is regardless of height. This is important because if there are 1,000,000 smurfs and only 100 trolls, unless the heights are very different between smurfs and trolls, it is most logical to assume that we are almost always seeing a smurf no matter what the height is. Thus, to make a Bayesian decision we need two basic pieces of information. First we need to know the prior probability \( P(j) \) of observing a \( j \) which is either a smurf or a troll. Second, we need to know the class conditional probability \( p(x|j) \) of observing a height \( x \) given some creature \( j \). Here is how we might do this:

We compute the mean heights for both trolls and smurfs:

\[
\overline{\text{height}}_{\text{smurf}} = \frac{\sum_{\text{smurfs}} \text{height}_{\text{smurf}}}{N_{\text{smurf}}}
\]

(1.4)

\[
\overline{\text{height}}_{\text{troll}} = \frac{\sum_{\text{trolls}} \text{height}_{\text{troll}}}{N_{\text{troll}}}
\]

(1.5)

Then we compute the standard deviation of both trolls and smurfs

\[
S_{\text{smurf}} = \sqrt{\frac{\sum_{\text{smurfs}} \text{height}_{\text{smurf}}^2}{N_{\text{smurf}}} - \overline{\text{height}}_{\text{smurf}}^2}
\]

(1.6)
Finally we compute the class conditional probability of observing a height given either a smurf or a troll

\[ p(\text{height}|\text{smurf}) = \frac{1}{\sqrt{2\pi S^2_{\text{smurf}}}} e^{-\frac{(\text{height} - \text{height}_{\text{smurf}})^2}{2S^2_{\text{smurf}}}} \]  

\[ p(\text{height}|\text{troll}) = \frac{1}{\sqrt{2\pi S^2_{\text{troll}}}} e^{-\frac{(\text{height} - \text{height}_{\text{troll}})^2}{2S^2_{\text{troll}}}} \]  

Finally, we compute the prior probability of observing a troll or a smurf regardless of height as:

\[ P(j) = \sum_{i \in I} \frac{N_{j \in I}}{N_{i \in I}} \]  

\[ P(\text{smurf}) = \frac{N_{\text{smurf}}}{N_{\text{smurf}} + N_{\text{troll}}} \]  

\[ P(\text{troll}) = \frac{N_{\text{troll}}}{N_{\text{smurf}} + N_{\text{troll}}} \]  

In general, we have the most important parts done. However, let's take a look at Bayes rule in this context:

\[ P(j|x) = \frac{p(x|j)P(j)}{p(x)} \]

We notice that in order to figure out the final probability, we need the marginal probability \( p(x) \) which is the base probability of height. Basically, \( p(x) \) is there to make sure that if we summed up all the \( P(j|x) \) they would equal one. This is simply

\[ p(x) = \sum_{j=1}^{M} p(x|j)P(j) \]

as such we see we can compute this as:
(1.15) \[ p(\text{height}) = \sum_{\text{creatures}} p(\text{height}|\text{creature})P(\text{creature}) \]

which would be:

(1.16) \[ p(\text{height}) = p(\text{height}|\text{smurf})P(\text{smurf}) + p(\text{height}|\text{troll})P(\text{troll}) \]

Finally…. We have everything we need to determine if a creature is most likely a troll or a smurf based upon its height. Thus, we can now ask, if we observe a creature that is 2” tall how likely is it to be a smurf?

(1.17) \[ P(\text{smurf}|2") = \frac{p(2"|\text{smurf})P(\text{smurf})}{p(2")} \]

Next we ask what is the likelihood of that we have a troll given the observation of 2”?

(1.18) \[ P(\text{troll}|2") = \frac{p(2"|\text{troll})P(\text{troll})}{p(2")} \]

The end result is that if:

(1.19) \[ P(\text{troll}|2") < P(\text{smurf}|2") \]

then we can decide that we are most likely observing a smurf. It should be noted that we have only stated that one is more likely than the other. We have not stated how much more likely one is than the other. We will leave that alone for now. What we will finish with is that by determining that the probability of one creature is higher than the other, we have a decision boundary. Thus if we measure some height \( x \) we can take one of two actions in this case:

(1.20) \[ P(\text{troll}|x) > P(\text{smurf}|x) \]

in which case we are observing a troll and should eject it from the club. Or we might have:

(1.21) \[ P(\text{troll}|x) < P(\text{smurf}|x) \]

in which case we are observing a smurf and should let it into the club. Thus, the simplest rule we can use is that the highest probability wins.