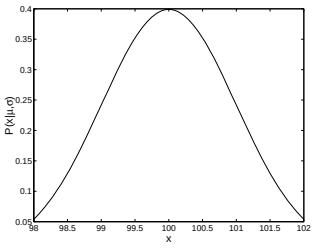
Computational Perception & Scene Analysis
15/85-485/785

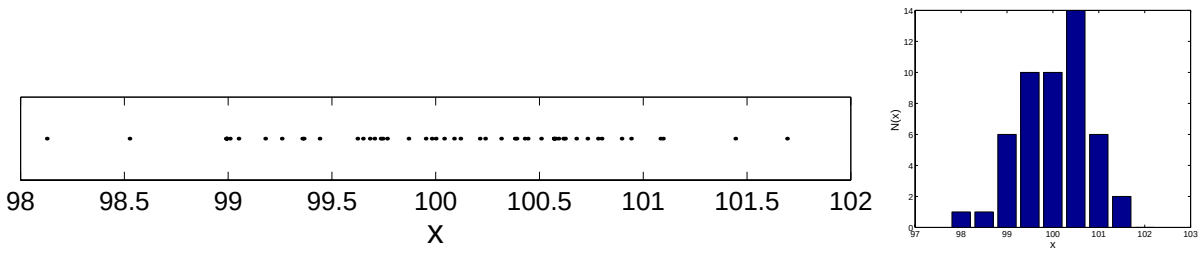
A tutorial on classical and Bayesian statistics

1 Introduction

Imagine drawing fifty samples $x_1 ... x_{50}$ from a Gaussian distribution with some mean μ and variance σ. For $\mu = 100$ and $\sigma = 1$ the probability distribution function (pdf) $P(x|\mu, \sigma)$ looks like this:



Here's a scatterplot of 50 samples, and a histogram of their distribution:



The histogram is of course somewhat distorted compared to the pdf it was sampled from. If we didn't know μ, the mean of the Gaussian from which the samples were drawn, how could we estimate it from the samples?

The classical answer is to compute some function of the samples to serve as an estimator of the underlying quantity you're interested in. In this case, the mean of the samples is the usual such estimator (and it equals 100.0565 for the data shown above).

It's not always clear that this is the best approach. For one thing, boiling 50 numbers worth of data down to a single summary statistic throws away a lot of information. We'll consider a different approach, using the information in the samples to derive a *distribution* of all possible values of μ and their relative likelihoods given the evidence, rather than just a single summary estimate of μ.

In classical statistics, a pdf is used to describe the probabilities that observations of a *random variable* such as x will take on various values. In Bayesian statistics, a pdf can also be used to represent beliefs about *any variable* — such as μ — whose true value is uncertain. In this case, the pdf represents the likelihood

that the *true value* of μ has various values. (A random variable like x, in contrast, does not have a true value.) Here we will consider how to manipulate such pdfs and how to update their values in the face of new information.

2 Bayes' rule; maximum likelihood estimation

The basic tool for these purposes is *Bayes' rule*:

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

(This is easily derived from the product rule for conjunctive probabilities $P(a,b) = P(a|b)P(b)$ together with the fact that $P(a,b) = P(b,a)$.)

Bayes' rule provides a recipe for *updating* your beliefs about the true value of some quantity a when you obtain new information b. This process is called *conditionalizing* on b. If $P(a)$ is the pdf representing your beliefs about the true value of a before observing b — known as the *prior* distribution — then the result of applying Bayes rule is the updated distribution $P(a|b)$, known as the *posterior*. The denominator in Bayes' rule, $P(b)$, the probability of the observation, is just a constant normalizing factor. Since $P(b) = \int P(b|a)P(a)da$, dividing by it ensures that the pdf $P(a|b)$ is normalized: $\int P(a|b)da = 1$.

We can see how this works concretely by returning to the example of the Gaussian distribution. We'll start by assuming that the standard deviation $\sigma = 1$ is known, and all we're trying to determine from the samples $x_1...x_{50}$ is a posterior distribution $P(\mu|x_1...x_{50})$ for the true value of the mean μ. Here the fifty samples play the role of the observation b, but there are several ways of organizing the actual update. Of course, we could apply Bayes' rule once with b being the joint event of all fifty observations. But Bayes' rule can also be used iteratively: we can update our distribution once for each individual observation x_n, then use the resulting posterior $P(\mu|x_1...x_n)$ as the new prior in the next update with observation x_{n+1}. This can be useful in practice because it allows one to take data into account incrementally as it becomes available. Both of these procedures result in the same ultimate posterior, as does updating with any permutation of the observations $x_1...x_{50}$. (You can verify this easily using the commutativity of multiplication and the independence of the observations.) We'll work through some incremental updates here.

We also have to choose the initial prior $P(\mu)$. If we had any information about μ other than the samples, we could reflect it with our choice of prior. Absent that, it seems reasonable to assume that all values of μ are equally likely; this is known as the *uniform prior*. Under a uniform prior, Bayes' rule takes a particularly simple form:

$$P(\mu|x_1) \propto P(x_1|\mu)$$

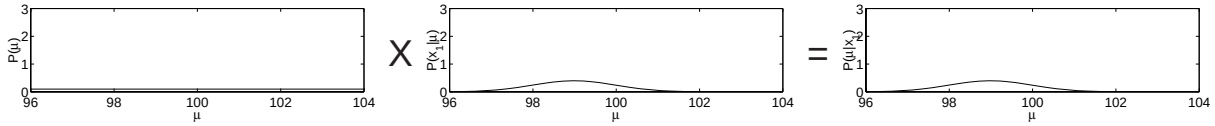
since the rest of the terms are constant for all μ

This equation asserts that the posterior $P(\mu|x_1)$ is proportional to the *likelihood function* $P(x_1|\mu)$. This is the function that tells us, for each possible choice for the true mean μ, how likely would the observation x_1 have been. We can compute this using the standard Gaussian equation:

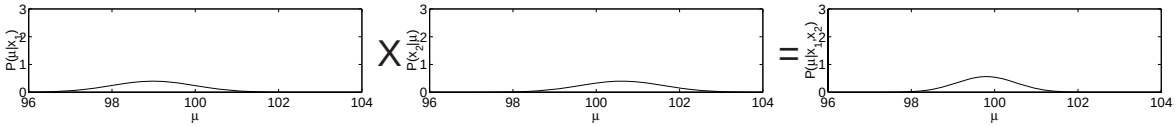
$$P(x_1|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/(2\sigma^2)} \tag{1}$$

by plugging in the observed value x_1 and $\sigma = 1$ and then looking at how the likelihood $P(x_1|\mu,\sigma)$ varies as a function of the mean μ. Note this is somewhat different from how you are probably used to using a pdf equation like this: by plugging in a pair of values for μ and σ and looking at the probability of different possible observations x_1 (this, for instance, is how we made the first figure in this review). Here the observation x_1 is fixed and we want to see its probability under different possible choices of μ. In fact, μ and x_1 are used symmetrically in the equation, so holding x_1 fixed and varying μ just produces a Gaussian, with mean x_1 and standard deviation $\sigma = 1$.

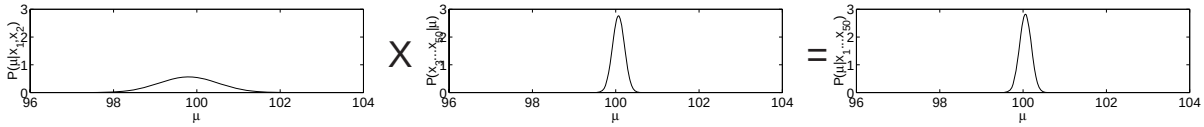
So applying Bayes rule to the first observation in the data shown above — which was $x_1 = 98.9894$ — produces a Gaussian posterior. (We illustrate here the prior times the likelihood producing the normalized posterior.)



We can repeat the process to update our beliefs about μ in light of the second data point, $x_2 = 100.6145$. This time, though, we have a nonuniform prior: $P(\mu|x_1)$, which is just the Gaussian posterior pictured above. We multiply this by the likelihood function $P(\mu|x_2)$, another Gaussian, and renormalize to obtain the posterior $P(\mu|x_1, x_2)$. (Multiplying two functions is multiplying them at each point. To compute the posterior at the point $\mu = 100.0$, $P(\mu = 100.0|x_1, x_2)$, we multiply the prior at that point, $P(\mu = 100.0|x_1)$, by the likelihood at that point, $P(x_2|\mu = 100.0)$. Repeat this for all μ and renormalize to obtain the posterior.) Since the product of two Gaussians is also Gaussian, our posterior is still Gaussian. It is somewhat more sharply peaked, since we now have more information about the true value of μ:



We can conditionalize on the remaining 48 sample points all at once to arrive at the final posterior $P(\mu|x_1...x_{50})$. It is a very sharp Gaussian — reflecting the fact that we have enough information to be fairly certain about the true value of μ — and it is centered at the sample mean of 100.0565. This point is also called the *maximum likelihood estimate* of μ because it is the choice of μ that maximizes the likelihood $P(x_1...x_{50}|\mu)$ of the data. (In the illustration, we have renormalized the likelihood $P(x_3...x_{50}|\mu)$ to make it visible.)



3 Marginalization

One final point. If we hadn't been told the value of σ, but rather just a *distribution* over possible values $P(\sigma)$, how would we have proceeded, and how would our beliefs be different? How do we get from the distributions we know — $P(x|\mu, \sigma)$ and $P(\sigma)$ — to the one we need, $P(x|\mu)$? We can *marginalize* out the unknown quantity, by summing over all of its possible values, weighted by their probabilities. (This is essentially a weighted average.) The familiar discrete rule is:

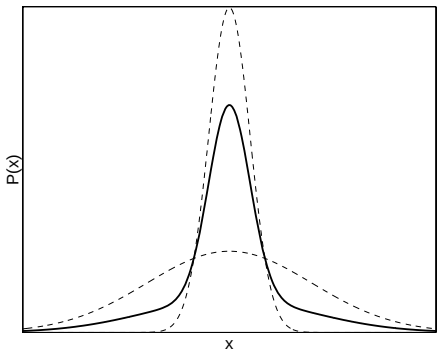
$$P(a) = \sum_b P(b)P(a|b)$$

In the Gaussian problem

$$P(x|\mu) \sum_\sigma P(\sigma)P(x|\mu, \sigma)$$

The continuous case just substitues an integral for the sum. The answer, $P(x|\mu)$ is known as the *marginal distribution.*

As a concrete example, suppose our values x were temperature measurements taken from a sensor. When this sensor is functioning, it returns values distributed with variance $\sigma = 1$, but it may be broken, in which case it will return less reliable samples distributed with variance $\sigma = 4$. If we know that the sensor is broken with probability 40%, we can *marginalize* over the two possible values for the standard deviation .by constructing a weighted sum of the dashed-line normal distributions $P(x|\mu, 1)$ and $P(x|\mu, 4)$. This tells us $P(x|\mu)$ (the solid line), the overall probability distribution of x, taking into account the possibility that the sensor is broken. As you would expect, since the observations x are less likely to accurately reflect μ, the marginal distribution is more spread out and has fatter tails than the likelihood function that assumed $\sigma = 1$ (which is the sharper of the dashed distributions above):



Using the marginal in place of the Gaussian for our likelihood function we can proceed as before, conditionalizing on the 50 observations. The results look similar, but our posterior pdf is sightly broader and less sharply peaked, reflecting the fact that since we are unsure of the true value of σ, we are not as confident in the reliability of our data as a reflection of μ. Here is the final posterior $P(\mu|x_1...x_{50})$; the dashed line is our previous result assuming $\sigma = 1$ and the solid line is the new result assuming the sensor may be broken: