Probability concepts explained: Bayesian inference for parameter estimation.

Jonny Brooks-Bartlett Follow
Jan 5, 2018 · 14 min read

Introduction

In the previous blog post I covered the maximum likelihood method for parameter estimation in machine learning and statistical models. In this post we'll go over another method for parameter estimation using Bayesian inference. I'll also show how this method can be viewed as a generalisation of maximum likelihood and in what case the two methods are equivalent.

Some fundamental knowledge of probability theory is assumed e.g. marginal and conditional probability. These concepts are explained in my first post in this series. Additionally, it also helps to have some basic knowledge of a Gaussian distribution but it's not necessary.

Bayes' Theorem

Before introducing Bayesian inference, it is necessary to understand Bayes' theorem. Bayes' theorem is really cool. What makes it useful is that it allows us to use some knowledge or belief that we already have (commonly known as the *prior*) to help us calculate the probability of a

related event. For example, if we want to find the probability of selling ice cream on a hot and sunny day, Bayes' theorem gives us the tools to use prior knowledge about the likelihood of selling ice cream on any other type of day (rainy, windy, snowy etc.). We'll talk more about this later so don't worry if you don't understand it just yet.

Mathematical definition

Mathematically Bayes' theorem is defined as:

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

where *A* and *B* are events, *P(A|B)* is the conditional probability that event *A* occurs given that event *B* has already occurred (*P(B|A)* has the same meaning but with the roles of *A* and *B* reversed) and *P(A)* and *P(B)* are the marginal probabilities of event *A* and event *B* occurring respectively.

Example

Mathematical definitions can often feel too abstract and scary so let's try to understand this with an example. One of the examples that I gave in the introductory blog post was about picking a card from a pack of traditional playing cards. There are 52 cards in the pack, 26 of them are red and 26 are black. *What is the probability of the card being a 4 given that we know the card is red?*

To convert this into the math symbols that we see above we can say that event A is the event that the card picked is a 4 and event B is the card being red. Hence, *P(A|B)* in the equation above is *P(4|red)* in our example, and this is what we want to calculate. We previously worked out that this probability is equal to 1/13 (there 26 red cards and 2 of those are 4's) but let's calculate this using Bayes' theorem.

We need to find the probabilities for the terms on the right hand side. They are:

1. P(B|A) = P(red|4) = 1/2
2. P(A) = P(4) = 4/52 = 1/13
3. P(B) = P(red) = 1/2

When we substitute these numbers into the equation for Bayes' theorem above we get 1/13, which is the answer that we were expecting.

How does Bayes' Theorem allow us to incorporate prior beliefs?

Above I mentioned that Bayes' theorem allows us to incorporate prior beliefs, but it can be hard to see how it allows us to do this just by looking at the equation above. So let's see how we can do that using the ice cream and weather example above.

Let A represent the event that we sell ice cream and B be the event of the weather. Then we might ask *what is the probability of selling ice cream on any given day given the type of weather?* Mathematically this is written as P(A=ice cream sale | B = type of weather) which is equivalent to the left hand side of the equation.

P(A) on the right hand side is the expression that is known as the **prior.** In our example this is P(A = ice cream sale), i.e. the (marginal) probability of selling ice cream regardless of the type of weather outside. P(A) is known as the prior because we might already know the marginal probability of the sale of ice cream. For example, I could look at data that said 30 people out of a potential 100 actually bought ice cream at some shop somewhere. So my P(A = ice cream sale) = 30/100 = 0.3, *prior to me knowing anything about the weather*. This is how Bayes' Theorem allows us to incorporate prior information.

Caution: I mentioned above that I could find data from a shop to get prior information, but there is nothing stopping me from making up a completely subjective prior that is not based on any data whatsoever. It's possible for someone to come up with a prior that is an informed guess from personal experience or particular domain knowledge but it's important to know that the resulting calculation will be affected by this choice. I'll go into more detail regarding how the strength of the prior belief affects the outcome later in the post.

Bayesian Inference

Definition

Now we know what Bayes' theorem is and how to use it, we can start to answer the question *what is Bayesian inference?*

Firstly, (statistical) **inference** is the process of deducing properties about a population or probability distribution from data. We did this in my previous post on maximum likelihood. From a set of observed data points we determined the maximum likelihood estimate of the mean.

Bayesian inference is therefore just the process of deducing properties about a population or probability distribution from data *using Bayes' theorem.* That's it.

Using Bayes' theorem with distributions

Until now the examples that I've given above have used single numbers for each term in the Bayes' theorem equation. This meant that the answers we got were also single numbers. However, there may be times when single numbers are not appropriate.

In the ice cream example above we saw that the prior probability of selling ice cream was 0.3. However, what if 0.3 was just my best guess but I was a bit uncertain about this value. The probability could also be 0.25 or 0.4. In this case a distribution of our prior belief might be more appropriate (see figure below). This distribution is known as the **prior distribution**.

2 distributions that represent our prior probability of selling ice on any given day. The peak value of both the blue and gold curves occur around the value of 0.3 which, as we said above, is our best guess of our prior probability of selling ice cream. The fact that f(x) is non-zero of other values of x shows that we're not completely certain that 0.3 is the true value of selling ice cream. The blue curve shows that it's likely to be anywhere between 0 and 0.5, whereas the gold curve shows that it's likely to be anywhere between 0 and 1. The fact that the gold curve is more spread out and has a smaller peak than the blue curve means that a prior probability expressed by the gold curve is "less certain" about the true value than the blue curve.

In a similar manner we can represent the other terms in Bayes' Theorem using distributions. We mostly need to use distributions when we're dealing with models.

Model form of Bayes' Theorem

In the introductory definition of Bayes' Theorem above I've used events A and B but when the model form of Bayes' theorem is stated in the literature different symbols are often used. Let's introduce them.

Instead of event A, we'll typically see Θ, this symbol is called Theta. Theta is what we're interested in, it represents the set of parameters. So if we're trying to estimate the parameter values of a Gaussian distribution then Θ represents both the mean, μ and the standard deviation, σ (written mathematically as $\Theta = \{\mu, \sigma\}$).

Instead of event B, we'll see *data* or $\boldsymbol{y} = \{y1, y2, \ldots, yn\}$. These represent the data, i.e. the set of observations that we have. I'll explicitly use *data* in the equation to hopefully make the equation a little less cryptic.

So now Bayes' theorem in model form is written as:

$$P(\boldsymbol{\Theta}|data) = \frac{P(data|\boldsymbol{\Theta}) \times P(\boldsymbol{\Theta})}{P(data)}.$$

We've seen that P(Θ) is the prior distribution. It represents our beliefs about the true value of the parameters, just like we had distributions representing our belief about the probability of selling ice cream.

*P(*Θ|*data)* on the left hand side is known as the **posterior distribution.** This is the distribution representing our belief about the parameter values after we have calculated everything on the right hand side taking the observed data into account.

P(data| Θ*)* is something we've come across before. If you made it to the end of my previous post on maximum likelihood then you'll remember that we said *L(data; μ, σ)* is the likelihood distribution (for a Gaussian distribution). Well *P(data|* Θ*)* is exactly this, it's the **likelihood distribution** in disguise. Sometimes it's written as $\mathscr{L}$(Θ; *data)* but it's the same thing here**.**

Therefore we can calculate the *posterior distribution* of our parameters using our *prior beliefs* updated with our *likelihood*.

This gives us enough information to go through an example of parameter inference using Bayesian inference. But first…

Why did I completely disregard P(data)?

Well, apart from being the marginal distribution of the data it doesn't really have a fancy name, although it's sometimes referred to as the **evidence**. Remember, we're only interested in the parameter values but *P(data)* doesn't have any reference to them. In fact, *P(data)* doesn't even evaluate to a distribution. It's just a number. we've already observed the data so we can calculate *P(data)*. In general, it turns out that calculating *P(data)* is **very hard** and so many methods exist to calculate it. This blog post by Prasoon Goyal explains several methods of doing so.

The reason why *P(data)* is important is because the number that comes out is a normalising constant. One of the necessary conditions for a probability distribution is that the sum of all possible outcomes of an event is equal to 1 (e.g. the total probability of rolling a 1, 2, 3, 4, 5 or 6 on a 6-sided die is equal to 1). The normalising constant makes sure that the resulting posterior distribution is a true probability distribution by ensuring that the sum of the distribution (I should really say integral because it's usually a continuous distribution but that's just being too pedantic right now) is equal to 1.

In some cases we don't care about this property of the distribution. We only care about where the peak of the distribution occurs, regardless of whether the distribution is normalised or not. In this case many people write the model form of Bayes' theorem as

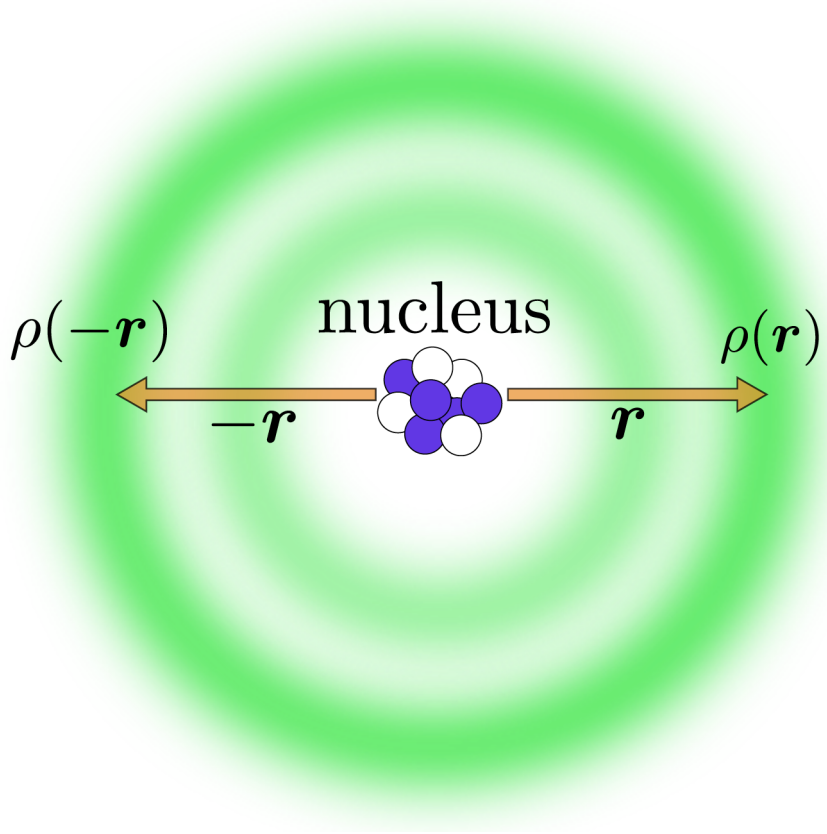
$$P(\boldsymbol{\Theta}|data) \propto P(data|\boldsymbol{\Theta}) \times P(\boldsymbol{\Theta})$$

where ∝ means "proportional to". This makes it explicit that the true posterior distribution is not equal to the right hand side because we haven't accounted for the normalisation constant *P(data)*.

Bayesian inference example

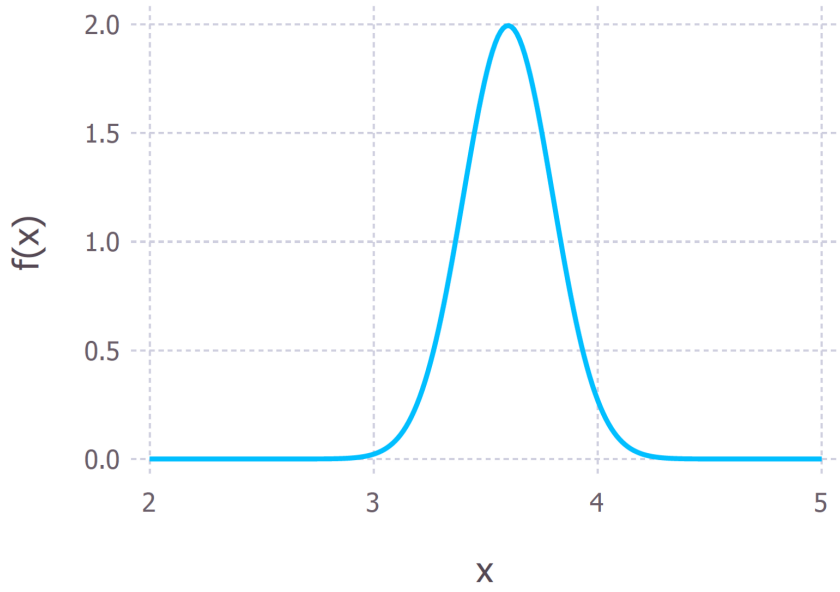
Well done for making it this far. You may need a break after all of that theory. But let's plough on with an example where inference might

come in handy. The example we're going to use is to work out the length of a hydrogen bond. You don't need to know what a hydrogen bond is. I'm only using this as an example because it was one that I came up with to help out a friend during my PhD (we were in the Biochemistry department which is why it was relevant at the time).



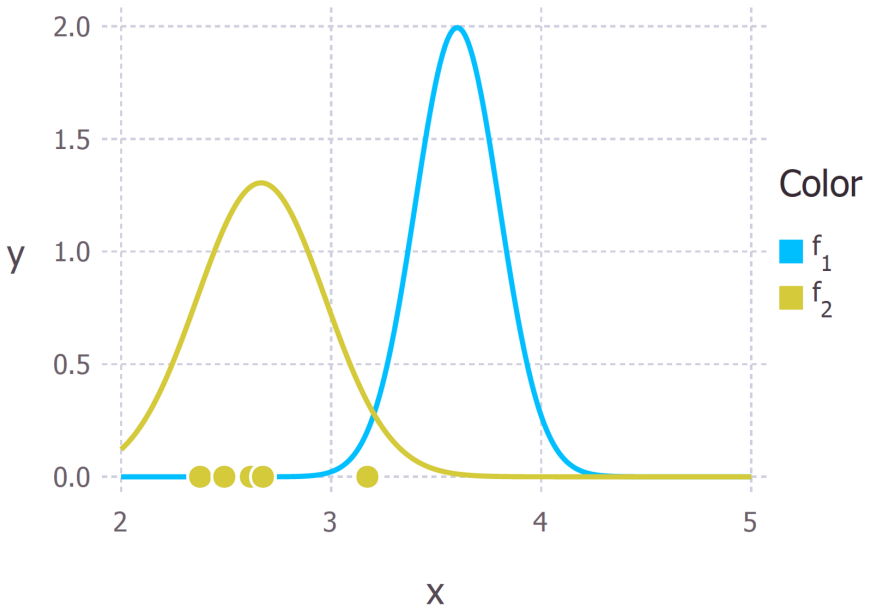
I've included this image because I think it looks nice, helps to break up the dense text and is kind of related to the example that we're going to go through. Don't worry, you don't need to understand the figure to understand what we're about to go through on Bayesian inference. In case you're wondering, I made the figure with Inkscape.

Let's assume that a hydrogen bond is between 3.2Å—4.0Å (A quick check on Google gave me this information. The Ångström, Å, is a unit of distance where 1Å is equal to 0.1 nanometers, so we're talking about very tiny distances). This information will form my prior. In terms of a probability distribution, I'll reformulate this as a Gaussian distribution with mean $\mu = 3.6$Å and standard deviation $\sigma = 0.2$Å (see figure below).



Our prior probability for the length of a hydrogen bond. This is represented by a Gaussian distribution with mean μ = 3.6Å and standard deviation σ = 0.2Å.

Now we're presented with some data (5 data points generated randomly from a Gaussian distribution of mean 3Å and standard deviation 0.4Å to be exact. In real world situations these data will come from the result of a scientific experiment) that gives measured lengths of hydrogen bonds (gold points in Figure 3). We can derive a likelihood distribution from the data just like we did in the previous post on maximum likelihood. Assuming that the data were generated from a process that can be described by a Gaussian distribution we get a likelihood distribution represented by the gold curve in the figure below. Notice that the maximum likelihood estimate of the mean from the 5 data points is less than 3 (about 2.8Å)

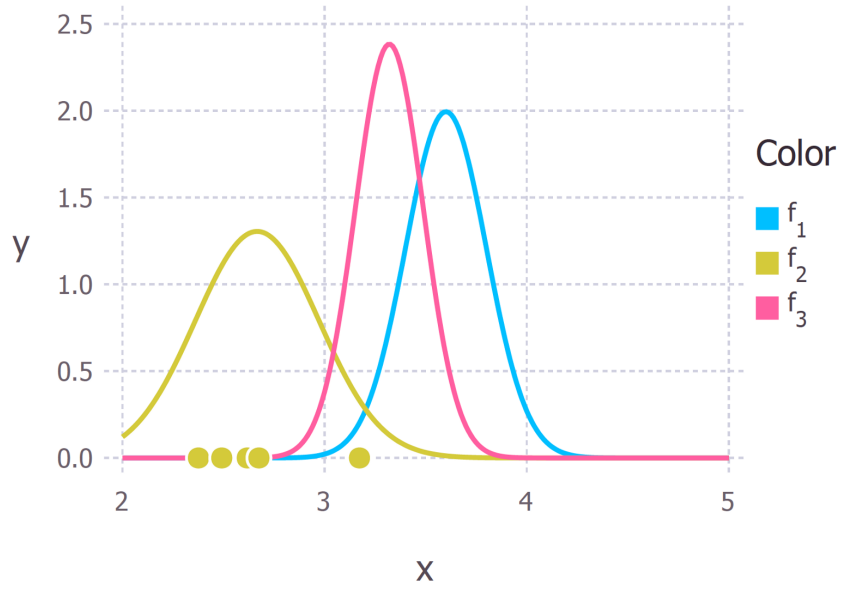


Prior probability for the distance of a hydrogen bond in blue and the likelihood distribution in gold derived from the 5 gold data points.

Now we have 2 Gaussian distributions, blue representing the prior and gold representing the likelihood. We don't care about the normalising constant so we have everything we need to calculate the unnormalised posterior distribution. Recall that the equation representing the probability density for a Gaussian is

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

So we have to multiply 2 of these. I wont go through the maths here because it gets very messy. If you're interested in the maths then you can see it performed in the first 2 pages of this document. The resulting posterior distribution is shown in pink in the figure below.



The posterior distribution in pink generated by multiplying the blue and gold distributions.

Now we have the posterior distribution for the length of a hydrogen bond we can derive statistics from it. For example, we could use the expected value of the distribution to estimate the distance. Or we could calculate the variance to quantify our uncertainty about our conclusion. One of the most common statistics calculated from the posterior distribution is the mode. This is often used as the estimate of the true value for the parameter of interest and is known as the **Maximum a posteriori probability estimate** or simply, the **MAP** estimate. In this case the posterior distribution is also a Gaussian distribution, so the mean is equal to the mode (and the median) and the MAP estimate for the distance of a hydrogen bond is at the peak of the distribution at about 3.2Å.

Concluding remarks

Why am I always using Gaussians?

You'll notice that in all my examples that involve distributions I use Gaussian distributions. One of the main reasons is that it makes the maths a lot easier. But for the Bayesian inference example it required calculating the product of 2 distributions. I said this was messy and so I didn't go through the maths. But even without doing the maths myself, I knew that the posterior was a Gaussian distribution. This is because the Gaussian distribution has a particular property that makes it easy to work with. It's *conjugate* to itself with respect to a Gaussian likelihood function. This means that if I multiply a Gaussian prior distribution with a Gaussian likelihood function, I'll get a Gaussian posterior function. The fact that the posterior and prior are both from the same

distribution family (they are both Gaussians) means that they are called **conjugate distributions.** In this case the prior distribution is known as a **conjugate prior.**

In many inference situations likelihoods and priors are chosen such that the resulting distributions are conjugate because it makes the maths easier. An example in data science is Latent Dirichlet Allocation (LDA) which is an unsupervised learning algorithm for finding topics in several text documents (referred to as a corpus). A very good introduction to LDA is can be found here in Edwin Chen's blog.

In some cases we can't just pick the prior or likelihood in such a way to make it easy to calculate the posterior distribution. Sometimes the likelihood and/or the prior distribution can look horrendous and calculating the posterior by hand is not easy or possible. In these cases we can use different methods to calculate the posterior distribution. One of the most common ways is by using a technique called Markov Chain Monte Carlo methods. Ben Shaver has written a brilliant article called A Zero-Math Introduction to Markov Chain Monte Carlo Methods that explains this technique in a very accessible manner.

What happens when we get new data?

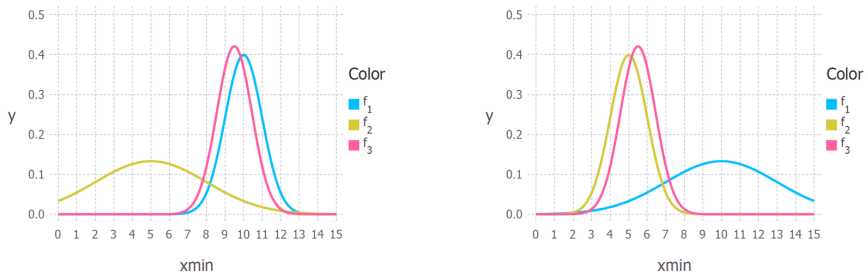
One of the great things about Bayesian inference is that you don't need lots of data to use it. 1 observation is enough to update the prior. In fact, the Bayesian framework allows you to update your beliefs iteratively in realtime as data comes in. It works as follows: you have a prior belief about something (e.g. the value of a parameter) and then you receive some data. You can update your beliefs by calculating the posterior distribution like we did above. Afterwards, we get even more data come in. So our posterior becomes the new prior. We can update the new prior with the likelihood derived from the new data and again we get a new posterior. This cycle can continue indefinitely so you're continuously updating your beliefs.

The Kalman filter (and it's variants) is a great example of this. It's used in many scenarios, but possibly the most high profile in data science are its applications to self driving cars. I used a variant called the Unscented Kalman filter during my PhD in mathematical protein crystallography, and contributed to an open source package implementing them. For a good visual description of Kalman Filters check out this blog post: How a Kalman filter works, in pictures by Tim Babb.

Using priors as regularisers

The data that we generated in the hydrogen bond length example above suggested that 2.8Å was the best estimate. However, we may be at risk of overfitting if we based our estimate solely on the data. This would be a huge problem if something was wrong with the data collection process. We can combat this in the Bayesian framework using priors. In our example using a Gaussian prior centred on 3.6Å resulted in a posterior distribution that gave a MAP estimate of the hydrogen bond length as 3.2Å. This demonstrates that our prior can act as a regulariser when estimating parameter values.

The amount of weight that we put on our prior vs our likelihood depends on the relative uncertainty between the two distributions. In the figure below we can see this graphically. The colours are the same as above, blue represents the prior distribution, gold the likelihood and pink the posterior. In the left graph in the figure you can see that our prior (blue) is much less spread out than the likelihood (gold). Therefore the posterior resembles the prior much more that the likelihood. The opposite is true in the graph on the right.

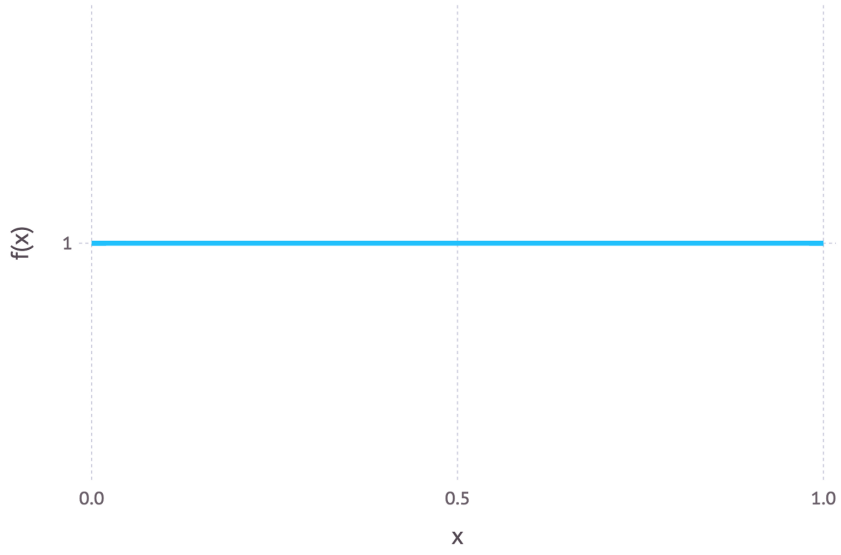


Therefore if we wish to increase the regularisation of a parameter we can choose to narrow the prior distribution in relation to the likelihood.

Michael Green has written an article called The truth about Bayesian priors and overfitting that covers this in more detail and gives advice on how to set priors.

When is the MAP estimate equal to the maximum likelihood estimate?

The MAP estimate is equal to the MLE when the prior distribution is uniform. An example of a uniform distribution is shown below.



Uniform distribution

What we can see is that the uniform distribution assigns equal weight to every value on the x-axis (it's a horizontal line). Intuitively it represents a lack of any prior knowledge about which values are most likely. In this case all of the weight is assigned to the likelihood function, so when we multiply the prior by the likelihood the resulting posterior exactly resembles the likelihood. Therefore, the maximum likelihood method can be viewed as a special case of MAP.

. . .

When I started writing this post I didn't actually think that it would be anywhere near this long so thank you so much for making it this far. I really do appreciate it. As always, if there is anything that is unclear or I've made some mistakes in the above feel free to leave a comment. In the next post in this series I will probably try to cover marginalisation for working out *P(data)*, the normalising constant that I ignored in this post. Unless of course there is something else that someone would like me to go over ;)

Thank you for reading.