Course 67564 2022/23

Bayesian Machine Learning

Recitation 3: Estimating the Gaussian Distribution

Prof. Yair Weiss *TA: Roy Friedman*

Up until now, we only discussed the general properties of distributions, especially those of the Gaussian distribution. But in real life, we are often given a set of data points:

$$\mathcal{D}=\{x_1,\cdots,x_N\}\equiv\{x_i\}_{i=1}^N \tag{0.1}$$

which are assumed to have been drawn independently from some distribution. Under this assumption, we want to either find the parameters that explain the data the best (the frequentist view) or to update our beliefs regarding the distribution we think the points are drawn from (the Bayesian view). In this chapter we will see how to estimate the parameters of the Gaussian distribution under both views.

1 Frequentist versus Bayesian Views

These two philosophies are often framed as contradictory, but they don't necessarily have to compete with each other. We will see that many times the frequentist problem can be framed using the Bayesian philosophy (and vice versa). Understanding both sides of the argument will help build a stronger foundation and intuition for machine learning in general.

1.1 Frequentist Machine Learning

The frequentist outlook can further be split into two categories - classical and probabilistic. The first outlook is closer to what is regularly taught in intro2ML type classes, while the second is somewhat more structured. The core characteristics of the frequentist philosophy (both probabilistic and not) can be summarized as:

1. There is some true set of parameters θ which model the data; θ is **not** a probabilistic object
2. We collect data points $\mathcal{D}$ from this "true model" and want to *estimate* θ from these data points using a loss function $L\left(\mathcal{D};\,\theta\right)$

Classical Machine Learning

In the most general form, frequentist machine learning requires two objects: a set of parameters θ that define a hypothesis class and a loss function $L\left(\mathcal{D};\,\theta\right)$. Typically, the loss function is chosen in such a way that it is minimized by the wanted outcome, although this can be hard to control in many real life applications.

Framing the problem of machine learning in this manner results in a deterministic algorithm, whose correctness is then left to be proven. Examples of such algorithms are *decision trees*, *support vector machines* (SVMs) and *k-means*, whose definition and solution are inherently not probabilistic.

Probabilistic Machine Learning

This form of frequentist machine learning is slightly more structured, framing the problem probabilistically. In this form, a stochastic model for the generation of the data is assumed. This model has parameters θ and a distribution attached to these parameters $p\left(\mathcal{D};\,\theta\right)$ which controls how likely it is for us to have observed the data under a specific choice of the parameters θ. Here, while θ appears in a density function, it is *not a probabilistic object*; that is what the semi-colon (the ";" sign) is meant to convey. The solution to this problem is then to find the parameters θ that created the data points in $\mathcal{D}$.

Many times, the classical and probabilistic views are connected. Many classical algorithms can be reframed in a probabilistic framework and vice versa. However, the probabilistic outlook allows us to take into explicitly take into account the stochastic nature of the data, which allows for a slightly more focused view of the problem.

Maximum Likelihood Estimation

One of the most common probabilistic criterions used to estimate the parameters is called the likelihood. Given a data set $\mathcal{D}$, we define the likelihood as:

$$L\left(\theta\right)\triangleq p\left(\mathcal{D}\,;\theta\right)=\prod_{i=1}^{N}p\left(x_{i}\,;\theta\right) \tag{1.1}$$

Notice, of course, that the likelihood is a function of the parameters θ. In this definition, we used the fact that the points were drawn *i.i.d.* from $p\left(\cdot\,;\theta\right)$ in order to multiply their probabilities - if they weren't drawn independently, we couldn't have done this!

Having defined this criterion, the natural step in order to estimate the parameters θ is to maximize the likelihood:

$$\hat{\theta}_{ML}\triangleq\arg\max_{\theta}L\left(\theta\right) \tag{1.2}$$

after all, if θ maximizes the likelihood, it is the most likely set of parameters to describe the distribution[1]. This estimate is called the *maximum likelihood estimate* (MLE) of the distribution and we will denote it by $\hat{\theta}_{ML}$ (the $\hat{}$ is to remember that it is an estimate and the $_{ML}$ is to remember that it maximizes the likelihood).

Also, usually the log-likelihood is maximized instead of the likelihood, defined as:

$$\ell\left(\theta\right)\triangleq\log L\left(\theta\right)=\sum_{i=1}^{N}\log p\left(x_{i}\,;\theta\right) \tag{1.3}$$

The result is the same (since the logarithm is a strictly monotonically increasing function), however this includes maximizing a sum instead of a product, which is usually easier.

1.2 Bayesian Machine Learning

The Bayesian philosophy assumes that we have some knowledge about the distribution the points were drawn from ahead of time, i.e. we assume that the parameters themselves have some distribution $p\left(\theta\right)$. This distribution is usually called the *prior distribution*, because we assume we have some prior knowledge. This means that there is no single true value for θ, rather that a distribution of θs could have given rise to the data. That is, unlike the frequentist view where θ *is by definition not probabilistic*, under the Bayesian view we assume that there is some distribution over θs.

In this new outlook, instead of trying to find the θ that generated the data, we will try to update our knowledge regarding which values θ could have had to create the data. We will want to find is the *posterior distribution* $p\left(\theta|\mathcal{D}\right)$, so called because we update our beliefs *after* the fact (in Latin "post" means "after", while "prior" means "before"). Using Bayes' law, we can describe this using the prior and likelihood distributions:

$$\overbrace{p\left(\theta|\mathcal{D}\right)}^{\text{posterior}}=\frac{\overbrace{p\left(\mathcal{D}|\theta\right)}^{\text{likelihood}}\overbrace{p\left(\theta\right)}^{\text{prior}}}{p\left(\mathcal{D}\right)}\propto\overbrace{p\left(\mathcal{D}|\theta\right)}^{\text{likelihood}}\overbrace{p\left(\theta\right)}^{\text{prior}} \tag{1.4}$$

Usually we assume that the data set is held constant, so $p\left(\mathcal{D}\right)$ does not affect the calculation of the posterior probability, which is why it is usually disregarded (or swallowed up by the $\propto$ sign). The likelihood term here $p\left(\mathcal{D}|\theta\right)$ is actually exactly the same as the frequentist likelihood $p\left(\mathcal{D}\,;\theta\right)$, only now we can properly condition on θ.

As mentioned, the posterior distribution is an updated version of our beliefs, and gives a new distribution over which values of θ are likely. That said, we can also extract point estimates (single estimates) of θ from the posterior:

1. The *maximum a-posteriori* (MAP) estimate is defined as: $\hat{\theta}_{MAP}\triangleq\arg\max_{\theta}p\left(\theta|\mathcal{D}\right)=\arg\max_{\theta}p\left(\theta\right)p\left(\mathcal{D}|\theta\right)$

2. The *minimum mean squared error* (MMSE) estimate is defined as: $\hat{\theta}_{MMSE}\triangleq\mathbb{E}\left[\theta|\mathcal{D}\right]$

[1] A more formal reason to use MLE is that it minimizes the KL-divergence with respect to the true model - see the Wikipedia page for MLE

1.3 Connections

If we assume an uninformative prior over θ, i.e. all values of θ are equally probable and the prior doesn't add any knowledge as to the choice of θ:

$$p(\theta) \propto 1 \tag{1.5}$$

then in this case the MAP estimate is:

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max_{\theta} p(\theta)\, p(\mathcal{D}|\theta) \\ &= \arg\max_{\theta} p(\mathcal{D}|\theta) \times \text{const} \\ &= \arg\max_{\theta} p(\mathcal{D}|\theta) = \hat{\theta}_{ML}\end{aligned} \tag{1.6}$$

So we see that the frequentist estimate $\hat{\theta}_{ML}$ is a special case of the Bayesian estimate $\hat{\theta}_{MAP}$!

Actually, we did something extremely fishy when we said that $p(\theta)$ is "uniform" - this isn't possible in many cases! A uniform probability over all of the real line $\mathbb{R}$ is impossible... so how can we even talk about this so called "uninformative prior"? While this is true, as long as the posterior $p(\theta|\mathcal{D})$ is well defined, the MAP and MMSE estimates will still exist. In this special case, the frequentist and Bayesian world views collide, and it will be useful to keep this fact in mind as we continue.

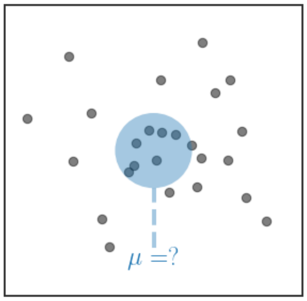


Figure 1: example of an estimation problem; data points are given, and their mean is probably in the blue area, but where?

2 Estimating the Gaussian Distribution

While Bayesian statistics is what interests us most in this course, many times it will prove easier to first go over the frequentist version as it is less mathematically involved. Only after we understand the ML solution, we will move on to the Bayesian treatment of the same, in the process revealing how they are related to each other.

The parameters of a Gaussian distribution are μ and Σ, so $\theta = \{\mu, \Sigma\}$. In the frequentist case we will estimate both, however the Bayesian treatment of Σ is a bit more complex and doesn't teach much, so we will ignore it for now.

2.1 MLE for the Gaussian Distribution

The log-likelihood of a data set $\mathcal{D} = \{x_i\}_{i=1}^N$ sampled from a Gaussian distribution is:

$$\begin{aligned}\log p(\mathcal{D}|\mu,\Sigma) &= \sum_i \log \mathcal{N}(x_i \,|\, \mu, \Sigma) \\ &= \sum_i \log \left[\frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left\{-\frac{1}{2}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu)\right\}\right] \\ &= \sum_i \left[-\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right] \\ &= -\frac{Nd}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_i (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\end{aligned} \tag{2.1}$$

Before we begin the process of finding the ML estimators[2] for μ and Σ, let's see another way of writing the log-likelihood. Notice that for any scalar a, we can write:

$$a = \text{trace}[a] \tag{2.2}$$

Since this is true for any scalar, we can apply this to the inner product of 2 vectors $x^T y$ (which is just a number), as well:

$$x^T y = \text{trace}\left[x^T y\right] = \text{trace}\left[y x^T\right] \tag{2.3}$$

[2] Bishop 2.3.4; Murphy 4.1.3

Now, recall that the Mahalanobis distance $(x-\mu)^T \Sigma^{-1}(x-\mu)$ is also a scalar, so we can use the above identity to rewrite it as:

$$\begin{aligned}
\ell(\mathcal{D}|\mu,\Sigma) &= -\frac{Nd}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\sum_i \text{trace}\left[\Sigma^{-1}(x_i-\mu)(x_i-\mu)^T\right] \\
&= -\frac{Nd}{2}\log 2\pi - \frac{N}{2}\log|\Sigma| - \frac{1}{2}\text{trace}\left[\Sigma^{-1}\sum_i (x_i-\mu)(x_i-\mu)^T\right] \qquad (2.4)
\end{aligned}$$

Finally, if we define $S \triangleq \frac{1}{N}\sum_i (x_i-\mu)(x_i-\mu)^T$ (which is almost the empirical covariance), we get a shorter form for the log-likelihood (which is sometimes used in the literature):

$$\ell(\mathcal{D}|\mu,\Sigma) = -\frac{N}{2}\left(d\log 2\pi + \log|\Sigma| + \text{trace}\left[\Sigma^{-1}S\right]\right) \qquad (2.5)$$

MLE for μ

We begin by finding the mean that maximizes the log-likelihood, by differentiating the log-likelihood:

$$\begin{aligned}
\frac{\partial\ell}{\partial\mu} &= -\frac{1}{2}\sum_i \frac{\partial}{\partial\mu}(x_i-\mu)^T \Sigma^{-1}(x_i-\mu) \\
&= -\frac{1}{2}\sum_i 2\Sigma^{-1}(x_i-\mu) \\
&= -\Sigma^{-1}\sum_i (x_i-\mu) \overset{!}{=} 0
\end{aligned}$$

By equating to 0 we can find the maxima:

$$\hat{\mu}_{\text{ML}} = \frac{1}{N}\sum_i x_i \qquad (2.6)$$

Here we write $\hat{\mu}_{\text{ML}}$ to show that it is the best *maximum likelihood estimator* for the data set. Another notation you may come across for an ML estimator is μ_{ML}. Notice that, unsurprisingly, the ML estimator for the mean of the Gaussian is the *empirical mean* or *sample mean* of the data.

MLE for Σ

Using the following definition of the derivatives (which are a bit harder to get directly on your own):

$$\frac{\partial}{\partial\Sigma}\log|\Sigma| = \frac{1}{|\Sigma|}\frac{\partial}{\partial\Sigma}|\Sigma| = \frac{1}{|\Sigma|}|\Sigma|\Sigma^{-1} = \Sigma^{-1} \qquad (2.7)$$

and

$$\frac{\partial}{\partial\Sigma}\text{trace}\left[\Sigma^{-1}S\right] = -\Sigma^{-1}S\Sigma^{-1} \qquad (2.8)$$

we can find the MLE for Σ. The full derivative of the log-likelihood be Σ is:

$$\begin{aligned}
&\frac{\partial\ell}{\partial\Sigma} = -\left(\Sigma^{-1} - \Sigma^{-1}S\Sigma^{-1}\right) \overset{!}{=} 0 \\
&\Rightarrow \Sigma^{-1} = \Sigma^{-1}S\Sigma^{-1} \\
&\Rightarrow I = \Sigma^{-1}S \\
&\Rightarrow \Sigma = S \\
&\Rightarrow \hat{\Sigma}_{\text{ML}} = \frac{1}{N}\sum_i (x_i-\mu)(x_i-\mu)^T \qquad (2.9)
\end{aligned}$$

Because $\hat{\mu}_{\text{ML}}$ is not dependent on $\hat{\Sigma}_{\text{ML}}$, we can first find the MLE for μ and then for Σ, so that:

$$\hat{\Sigma}_{\text{ML}} = \frac{1}{N}\sum_i (x_i-\hat{\mu}_{\text{ML}})(x_i-\hat{\mu}_{\text{ML}})^T \qquad (2.10)$$

Putting the two equations together, the MLE for a Gaussian distribution are:

$$\begin{aligned}\hat{\mu}_{\text{ML}} &= \frac{1}{N}\sum_i x_i \\ \hat{\Sigma}_{\text{ML}} &= \frac{1}{N}\sum_i \left(x_i - \hat{\mu}_{\text{ML}}\right)\left(x_i - \hat{\mu}_{\text{ML}}\right)^T\end{aligned} \tag{2.11}$$

2.2 1D Bayesian Inference

Recall that in the Bayesian treatment, we assume that the parameters are distributed in some manner. We begin by considering the 1D case for Gaussian distributions[3]:

$$p\left(x\right) = \frac{1}{Z}\exp\left[-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right] \tag{2.12}$$

For now, we will assume that we know the variance σ^2. We will assume a Gaussian prior over μ (if we want we can assume different priors as well, but let's stick with Gaussian priors for now):

$$p\left(\mu\right) = \mathcal{N}\left(\mu_0, \sigma_0^2\right) \tag{2.13}$$

Given a data set $\mathcal{D} = \{x_i\}_{i=1}^N$, the likelihood is:

$$p\left(\mathcal{D}|\mu\right) = \prod_{i=1}^N p\left(x_i|\mu\right) = \prod_{i=1}^N \mathcal{N}\left(x_i \mid \mu, \sigma^2\right) \tag{2.14}$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N \left(x_i-\mu\right)^2\right] \tag{2.15}$$

The posterior probability for μ will then be:

$$p\left(\mu|\mathcal{D}\right) = \frac{p\left(\mathcal{D}|\mu\right)p\left(\mu\right)}{p\left(\mathcal{D}\right)} \tag{2.16}$$

$$\propto p\left(\mathcal{D}|\mu\right)p\left(\mu\right) \tag{2.17}$$

Recall that the term $p\left(\mathcal{D}\right)$ is constant and only serves as a normalization, so for now we can ignore it.

Let's look at the product in equation 2.17 more closely:

$$p\left(\mathcal{D}|\mu\right)p\left(\mu\right) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^N \left(x_i-\mu\right)^2\right]\exp\left[-\frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right] \tag{2.18}$$

$$= \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^N \left(x_i-\mu\right)^2 + \frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)^2\right)\right] \tag{2.19}$$

Notice that the term in the exponent is *still quadratic in* μ. This means, of course, that this whole term is still a Gaussian distribution. Let's use the derivative trick from the previous recitation in order to find the distribution of μ exactly. Define:

$$\Delta \triangleq \frac{1}{2}\left(\frac{1}{\sigma^2}\sum_{i=1}^N \left(x_i-\mu\right)^2 + \frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)^2\right) \tag{2.20}$$

Recall, we can now differentiate Δ with respect to μ in order to find the mean and covariance of the posterior

[3] See Bishop 2.3.6 for more details

distribution:

$$\begin{aligned}\frac{\partial\Delta}{\partial\mu} &= \frac{1}{2}\left(\frac{1}{\sigma^2}\sum_i\frac{\partial}{\partial\mu}\left(x_i-\mu\right)^2+\frac{1}{\sigma_0^2}\frac{\partial}{\partial\mu}\left(\mu-\mu_0\right)^2\right)\\ &= \frac{1}{\sigma^2}\sum_i\left(\mu-x_i\right)+\frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)\\ &= \frac{1}{\sigma^2}\left(N\mu-\sum_i x_i\right)+\frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)\\ &= \frac{N}{\sigma^2}\left(\mu-\mu_{ML}\right)+\frac{1}{\sigma_0^2}\left(\mu-\mu_0\right)\\ &= \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\left(\mu-\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{N}{\sigma^2}\mu_{\text{ML}}+\frac{1}{\sigma_0^2}\mu_0\right)\right)\end{aligned} \tag{2.21}$$

where $\mu_{\text{ML}}=\frac{1}{N}\sum_i x_i$ is the ML estimate for μ, as we showed in section 2.1.

Defining:

$$\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\triangleq\frac{1}{\sigma_N^2} \tag{2.22}$$

the posterior of μ is equal to:

$$p\left(\mu|\mathcal{D}\right)=\mathcal{N}\left(\mu\,|\,\sigma_N^2\left(\frac{N}{\sigma^2}\mu_{\text{ML}}+\frac{1}{\sigma_0^2}\mu_0\right),\,\sigma_N^2\right) \tag{2.23}$$

where $\sigma_N^2=\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}=\left(\frac{N\sigma_0^2+\sigma^2}{\sigma_0^2\sigma^2}\right)^{-1}=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$ and $\mu_{\text{ML}}=\frac{1}{N}\sum_i x_i$. If we write all of this explicitly, we will get:

$$p\left(\mu|\mathcal{D}\right)=\mathcal{N}\left(\mu\,|\,\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\left(\frac{N}{\sigma^2}\mu_{\text{ML}}+\frac{1}{\sigma_0^2}\mu_0\right),\,\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\right) \tag{2.24}$$

Effects of sample size

It may be a good idea to get some intuition for the posterior we found. Let's look at the slightly simpler case of $\mu_0=0$ (but the analysis that follows is true for any μ_0). In this case, the posterior is:

$$\mathcal{N}\left(\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\cdot\frac{N}{\sigma^2}\mu_{\text{ML}},\,\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\right) \tag{2.25}$$

Let's see what happens when $N=0$. If we don't have any data, we should probably always fall back to the only thing we know: our prior. At $N=0$, we have:

$$N=0 \qquad \begin{array}{c}\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\cdot\frac{N}{\sigma^2}\mu_{\text{ML}}=0=\mu_0\\ \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}=\frac{\sigma_0^2\sigma^2}{\sigma^2}=\sigma_0^2\end{array} \tag{2.26}$$

so the posterior (naturally) falls back to the prior. If we look at the other extreme, $N\to\infty$, then there should be no ambiguity over the value of μ whatsoever:

$$N\to\infty \qquad \begin{array}{c}\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\cdot\frac{N}{\sigma^2}\mu_{\text{ML}}\to\frac{\sigma_0^2\sigma^2}{N\sigma_0^2}\cdot\frac{N}{\sigma^2}\mu_{\text{ML}}=\mu_{\text{ML}}\\ \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}\to\frac{\sigma_0^2\sigma^2}{N\sigma_0^2}=0\end{array} \tag{2.27}$$

If N is somewhere in between, then the term for the mean is (as we saw before):

$$\mu_N\triangleq\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}\left(\frac{N}{\sigma^2}\mu_{\text{ML}}+\frac{1}{\sigma_0^2}\mu_0\right) \tag{2.28}$$

This is a *weighted mean* of the two values μ_{ML} and μ_0(you can find a demo for this behavior here). We can look at the number of samples needed in order for μ_N to be *exactly* between the ML estimate and the prior by giving equal weight to both terms:

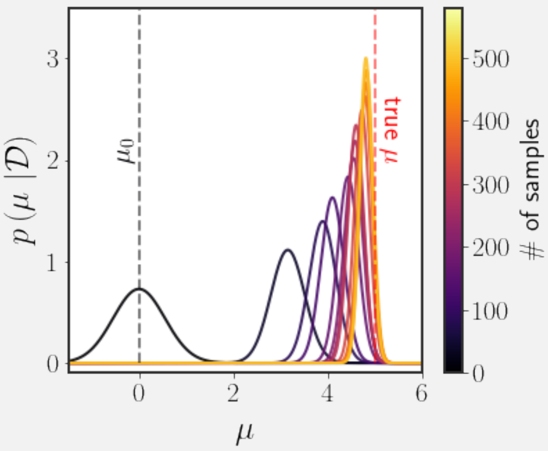
$$\frac{N}{\sigma^2}\overset{!}{=}\frac{1}{\sigma_0^2}\Rightarrow\hat{N}=\frac{\sigma^2}{\sigma_0^2} \tag{2.29}$$

So, when the variance of the prior is very small, which is like saying "we are very sure that μ is close to μ_0", then a lot of samples are needed in order to move μ_N away from the prior μ_0. If, on the other hand, the variance of the prior is very large, which may mean we are very unsure that μ_0 is correct, then few points are needed in order to move the mean from the prior mean. Finally, if the sample variance (σ^2) is very large, then we need to get a lot of data to be sure that the MLE is correct, while if it is very small, then we need very few points in order to be sure of the MLE.

Because the posterior is so dependent on the number of samples, it is sometimes written (like in equation 2.28) as:

$$p\left(\mu|\mathcal{D}\right) = \mathcal{N}\left(\mu\,|\,\mu_N, \sigma_N^2\right) \tag{2.30}$$

with the intention behind this notation being "this is the posterior mean after having sampled N points".



Posteriors for the mean after different amount of data are seen. When $N = 0$, the posterior is equal to the prior. As more points are observed, the posterior is pulled towards the true value that generated the points. Brighter posteriors are those with more observed points.

2.3 MAP and MMSE Estimates for μ

If we want to find the MAP estimate for μ under the prior we defined in section 2.2, we need to find:

$$\hat{\mu}_{\text{MAP}} = \arg\max_{\mu} p\left(\mu|\mathcal{D}\right) = \arg\max_{\mu} \mathcal{N}\left(\mu\,|\,\mu_N, \sigma_N^2\right) \tag{2.31}$$

Of course, the Gaussian distribution only has one maxima, which is the mean of the distribution. So the MAP estimate of μ is simply the mean:

$$\hat{\mu}_{\text{MAP}} = \mathbb{E}\left[p\left(\mu|\mathcal{D}\right)\right] = \mu_N \tag{2.32}$$

where μ_N is given explicitly in equation 2.28. Notice that (in this case) this is also the MMSE estimate:

$$\hat{\mu}_{MAP} = \mathbb{E}\left[p\left(\mu|\mathcal{D}\right)\right] = \hat{\mu}_{MMSE} \tag{2.33}$$

The fact that the MAP and MMSE estimates are the same is unique to the Gaussian distribution, in general they will differ quite a bit!

2.4 Multivariate Gaussian

Now that we understood the basic premise of the Bayesian inference for μ in 1D, we can start all over again for the multivariate case. We assume, again, that:

$$x \sim \mathcal{N}\left(\mu, \Sigma\right) \tag{2.34}$$

where the covariance matrix Σ is known. We also assume a prior over μ of the form:

$$\mu \sim \mathcal{N}\left(\mu_0, \Sigma_0\right) \tag{2.35}$$

The likelihood for a data set $\mathcal{D}$ is:

$$p\left(\mathcal{D}|\mu\right) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(x_i-\mu\right)^T\Sigma^{-1}\left(x_i-\mu\right)\right] \tag{2.36}$$

and the posterior is:

$$p\left(\mu|\mathcal{D}\right) \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(x_i-\mu\right)^T\Sigma^{-1}\left(x_i-\mu\right)\right]\exp\left[-\frac{1}{2}\left(\mu-\mu_0\right)^T\Sigma_0^{-1}\left(\mu-\mu_0\right)\right] \tag{2.37}$$

Essentially nothing has changed from before; the term in the exponent is still quadratic in μ, so we can employ our tricks once again:

$$\begin{aligned}\frac{\partial}{\partial\mu}\Delta &= \Sigma^{-1}\sum_i\left(\mu-x_i\right)+\Sigma_0^{-1}\left(\mu-\mu_0\right)\\ &= N\Sigma^{-1}\left(\mu-\mu_{\mathrm{ML}}\right)+\Sigma_0^{-1}\left(\mu-\mu_0\right)\\ &= \left(N\Sigma^{-1}+\Sigma_0^{-1}\right)\left[\mu-\left(N\Sigma^{-1}+\Sigma_0^{-1}\right)^{-1}\left(N\Sigma^{-1}\mu_{\mathrm{ML}}+\Sigma_0^{-1}\mu_0\right)\right]\end{aligned}$$

where we used the same definition as before for μ_{ML} (the ML estimate for μ). The full posterior is given by:

$$\mu|\mathcal{D} \sim \mathcal{N}\left(\mu_N,\Sigma_N\right) \tag{2.38}$$

where:

$$\Sigma_N = \left[N\Sigma^{-1}+\Sigma_0^{-1}\right]^{-1} \tag{2.39}$$

$$\mu_N = \Sigma_N\left[N\Sigma^{-1}\mu_{\mathrm{ML}}+\Sigma_0^{-1}\mu_0\right] \tag{2.40}$$

The result is consistent with what we saw in 1D. The main difference here is that now we need to invert the matrix $N\Sigma^{-1}+\Sigma_0^{-1}$ in order to find Σ_N and μ_N. The MAP/MMSE estimates for the multivariate μ are again the mean of the posterior (since this is a Gaussian distribution as well):

$$\hat{\mu}_{\mathrm{MAP}} = \left[N\Sigma^{-1}+\Sigma_0^{-1}\right]^{-1}\left[N\Sigma^{-1}\mu_{\mathrm{ML}}+\Sigma_0^{-1}\mu_0\right] = \hat{\mu}_{MMSE} \tag{2.41}$$

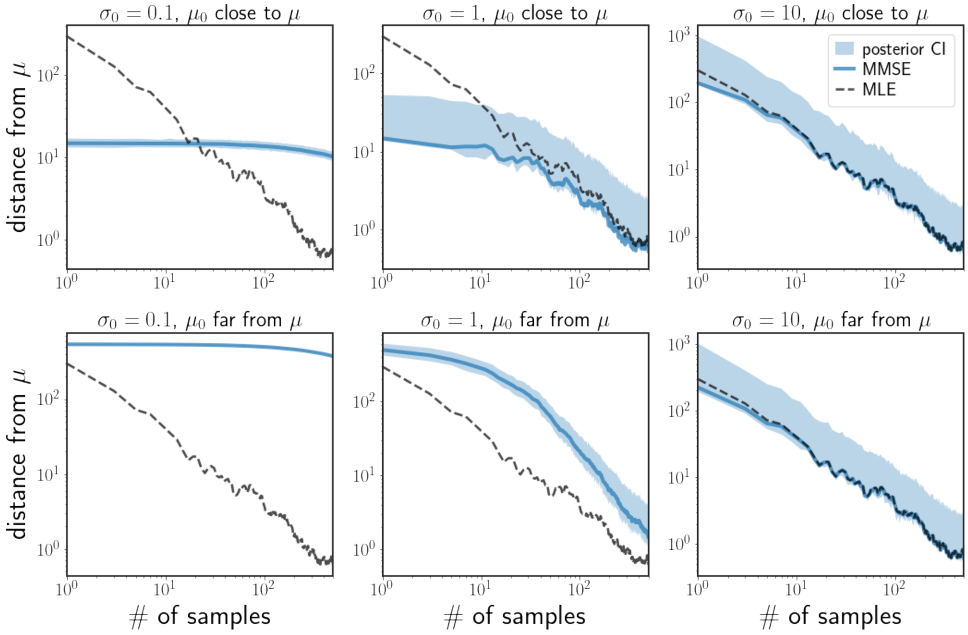


Figure 2: estimating the mean of a 30-dimensional Gaussian under different priors; to generate the data $\mu = 0$ was used, while the near prior mean was $\mu_0 = 1$ and the far prior mean was $\mu_0 = 10$. The shaded areas are the areas of the posterior with total probability of 95%. This simple example shows that the prior can positively or negatively affect the performance of the estimation.

2.5 Choices of Priors

Bayesian machine learning is often described in terms of “known priors”. However, many times we don’t actually have an explicit prior we can choose, which is the main criticism against the Bayesian approach. The problem is that when there isn’t a good prior, researchers amount to choosing arbitrary distributions for their priors.

Figure 2 illustrates what happens when Gaussian priors of different kinds are chosen. When the variance of the prior is low, i.e. $\Sigma_0 = I\sigma_0^2$ with small σ (left column), then many samples are needed to change the posterior distribution. When the prior mean is well calibrated to the generating distribution, this translates to a better MMSE than the ML estimate with few samples but doesn’t get better when more samples are introduced. However, when the prior mean is far from the generating distribution, then the estimate will always be quite bad. The other end of the spectrum is when σ is large (right column), in which case it doesn’t really matter what the prior mean is since the posterior mean is more or less equal to the ML estimate.

The more interesting case is when μ_0 is well calibrated and Σ_0 is moderate (middle column, top). In this setting, the MMSE gives a much better estimate than the MLE, *especially* in low sample-size settings - in this case, more than an order of magnitude. However, when the prior is bad, the MMSE estimate will always be worse than the MLE (middle column, bottom).