Indian Institute of Technology, Kanpur
Probabilistic Machine Learning and Inference

Instructors: Piyush Rai
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

SCRIBE

4

Multi-parameter models and their Inference

1. Multi-parameter Models

In the previous scribes, we had looked at the inference of the mean and the variance / precision parameters of a Normal distribtuion, however we assumed only one of them to be unkown and inferred the other. That assumption made the inference a lot easier, as it was for only a single parameter.

However, most machine learning tasks often have more than one parameter, and hence multi-parameter problems are important to understand. For this scribe, we will consider only models with two unknown parameters, with a single label / prediction, however the idea remains the same for any other multi-parameter model. Some examples of such models are shown in Figure 1

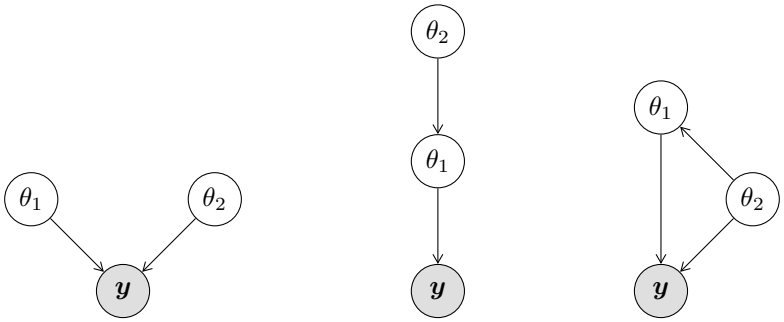

Figure 1: Plate Notations of multi-parameter models with two parameters

Suppose we had to infer both the mean and the precision of a Gaussian Distribution jointly (two-parameter model), how would we do it? We will look at this example and its inference in the next section, and then mention some other more general solutions to inference.

2. A Simple Multi-parameter Model

Assume a one-dimensional Gaussian Distribution with unknown mean μ and precision λ. We have n i.i.d. observations $X = \{x^1, x^2 \dots x^n\}$ sampled from this Gaussian *i.e*

$$\forall\, i \in [\, n \,],\ x^i \sim \mathcal{N}\left(x^i \mid \mu, \lambda^{-1}\right)$$

We already know how to perform MLE estimation for such a problem. We simply need to equate the derivative of the MLE w.r.t. μ and λ as we did for the coin-toss example (see Scribe 3). MLE estimation can, therefore, be a simple problem, even for multi-parameter models.

We wish to infer the mean and the precision of this Gaussian using fully Bayesian Inference. Therefore, we need to compute the joint posterior of μ and λ. Let us first state the likelihood function $\left(\mathbb{P}\left[\, X \mid \mu, \lambda \,\right]\right)$ and the joint prior on μ and λ.

Likelihood:
$$\mathbb{P}\left[\, X \mid \mu, \lambda \,\right] = \prod_{i=1}^{n} \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}\left(x^i - \mu\right)^2\right)$$

In order to make the posterior tractable, we need to have a conjugate prior. In order to get an idea for that, let us represent the likelihood in a different manner

$$\mathbb{P}\left[X \mid \mu, \lambda \right] \quad \propto \quad \left[\lambda^{\frac{1}{2}} \exp\left(-\frac{\lambda \mu^2}{2} \right) \right]^n \exp\left(\lambda\mu \sum_{i=1}^{n} x^i - \frac{\lambda}{2} \sum_{i=1}^{n} \left(x^i \right)^2 \right) \tag{1}$$

Looking at Equation 1, we can say that the likelihood looks like a multiple of a gaussian and a gamma distribution. A similar form for the prior might prove to be a conjugate for the likelihood. In fact, we can precisely have such a distribution for the prior, known as the Normal-Gamma or the Gaussian-Gamma (NG) distribution.

$$\mathrm{NG}\left(\mu, \lambda \mid \kappa_0, c, d \right) \quad \propto \quad \left[\lambda^{\frac{1}{2}} \exp\left(-\frac{\lambda \mu^2}{2} \right) \right]^{\kappa_0} \exp\left(\lambda\mu c - \lambda d \right)$$

Remark. The multi-variate version of the Normal-Gamma distribution is known as the Normal-Wishart Distribution.

Remark. If we are doing inference for variance rather than precision, then we can use Normal-Inverse Gamma or for the multi-variate version, Normal-Inverse Wishart Distribution.

It can be observed that the NG distribution has two terms, one which seems like a gaussian over μ and another which seems like a gamma distribution over λ. This can be better understood as when we marginalize over μ (*i.e.* integrate over μ), we will get only the second term, that is the gamma over λ, since the gaussian over μ will integrate to one. Therefore, the gaussian term is just the conditional of μ over λ and the gamma term is the marginal distribution of γ.

More formally, we can write the NG distribution in the following representation

$$\mathrm{NG}\left(\mu, \lambda \mid \mu_0, \kappa_0, \alpha_0, \beta_0, \right) \quad = \quad \mathbb{P}\left[\mu \mid \lambda \right] \mathbb{P}\left[\lambda \right] \quad = \quad \mathcal{N}\left(\mu \mid \mu_0, (\kappa_0 \lambda)^{-1} \right) \mathrm{Gamma}\left(\lambda \mid \alpha_0, \beta_0 \right) \tag{2}$$

where $\mu_0 = c/\kappa_0$, $\alpha_0 = 1 + \kappa_0/2$ and $\beta_0 = d - c^2/2\kappa$

Now, we can try to compute the posterior and show that the prior is indeed conjugate to the likelihood.

2.1 Computing the Posterior

$$\begin{aligned} \mathbb{P}\left[\mu, \lambda \mid X \right] &= \frac{\mathbb{P}\left[X \mid \mu, \lambda \right] \mathbb{P}\left[\mu, \lambda \right]}{\mathbb{P}\left[X \right]} \\ &\propto \mathbb{P}\left[X \mid \mu, \lambda \right] \mathbb{P}\left[\mu, \lambda \right] \\ &= \mathbb{P}\left[X \mid \mu, \lambda \right] \mathbb{P}\left[\mu \mid \lambda \right] \mathbb{P}\left[\lambda \right] \end{aligned}$$

The complete derivation is much more involved and can be found in the paper by Murphy [1]. However, we show the final form of the posterior as below. It is advised to try out the derivation as an exercise.

Posterior: $$\mathbb{P}\left[\mu, \lambda \mid X \right] \quad = \quad \mathrm{NG}\left(\mu_n, \kappa_n, \alpha_n, \beta_n \right) \tag{3}$$

where

$$\begin{aligned} \mu_n &= \frac{\kappa_0 \mu_0 + n\overline{x}}{\kappa_0 + n} \\ \kappa_n &= \kappa_0 + n \\ \alpha_n &= \alpha_0 + \frac{n}{2} \\ \beta_n &= \beta_0 + \frac{1}{2} \sum_{i=1}^{n} \left(x^i - \overline{x} \right)^2 + \frac{\kappa_0 n \left(\overline{x} - \mu_0 \right)^2}{2\left(\kappa_0 + n \right)} \\ \overline{x} &= \frac{1}{n} \sum_{i=1}^{n} x^i \end{aligned}$$

Hence, we have the posterior as a Normal-Gamma distribution, same as the prior.

2.2 Completing the Inference

We can also compute some other quantities of interest, such as the predictive posterior and the margin likelihood. The marginal likelihood will have the following form

$$\textbf{Marginal Likelihood:} \qquad \mathbb{P}\left[\, X \,\right] \;=\; \frac{\Gamma(\alpha_n)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha}{}_0}{\beta_n^{\alpha}{}_n} \left(\frac{\kappa_0}{\kappa_n}\right)^{\frac{1}{2}} (2\pi)^{-\frac{n}{2}}$$

For this case, we have the predictive posterior as a student t-distribution, however we will not look at the derivation, which although can be found in the Murphy's paper [1]

$$\textbf{Predictive Posterior:} \quad \mathbb{P}\left[\, x^* \,\middle|\, X \,\right] \;=\; \int_{\mu,\lambda} \mathbb{P}\left[\, x^* \,\middle|\, \mu, \lambda \,\right] \mathbb{P}\left[\, \mu, \lambda \,\middle|\, X \,\right] \;=\; t_{2\alpha_n}\left(x^* \,\middle|\, \mu_n, \frac{\beta_n(\kappa_n+1)}{\alpha_n \kappa_n}\right) \tag{4}$$

We can also compute the marginal posteriors of μ and λ

$$\begin{aligned}
\mathbb{P}\left[\, \lambda \,\middle|\, X \,\right] \;&=\; \int_{\mu} \mathbb{P}\left[\, \mu, \lambda \,\middle|\, X \,\right] \;=\; \text{Gamma}\left(\lambda \,\middle|\, \alpha_n, \beta_n\right) \\
\mathbb{P}\left[\, \mu \,\middle|\, X \,\right] \;&=\; \int_{\lambda} \mathbb{P}\left[\, \mu, \lambda \,\middle|\, X \,\right] \;=\; t_{2\alpha_n}\left(\mu \,\middle|\, \mu_n, \beta_n/(\alpha_n \kappa_n)\right)
\end{aligned}$$

3. Handling the other cases

We analysed a very simple multi-parameter model, for which we could find a conjugate prior. However, for most of the machine learning problems, this is not the case *i.e.* the posterior is generally intractable. Such a method would obviously fail in such a situation.

To conquer these, we use sampling methods such as Monte Carlo Markov Chain (MCMC) and Variational Bayesian Inference (VI). These methods can approximate inference techniques using much weaker conditions than conjugacy, such as for MCMC methods, local conjugacy (more on this later) is required, whereas VI can be used even without that.
Hence, the solution we discussed although powerful is infeasible for most of the problems. We will look at stronger methods as mentioned earlier later on in the series.

References

[1] Kevin P. Murphy *Conjugate Bayesian Analysis of the Gaussian Distribution, 2007*