EECE 7397 Review

Jiahao Zhu

September 2020

1 Bayesian Inference

The maximum likelihood estimation gave point estimates for the parameters μ and Σ. Now we consider the Bayesian treatment for the Gaussian distribution. Let's introduce prior distribution over these parameters. Given $X = \{x_1, x_2, ..., x_n\}$, here are three cases for the parameters μ and Σ.

1.1 Case 1: Mean is Unknown and Precision is Known

Assume that

- μ: unknown
- σ^2: known

In this case, we consider the task of inferring the mean μ given a set of N observations X. The likelihood function, that is the probability of the observed data given μ, is given by

$$p(\mathbf{D} \mid \mu) = \prod_{n=1}^{N} p(x_n \mid \mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\{-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2\} \tag{1}$$

It shows that the likelihood function takes the form of the exponential of a quadratic form in μ. Suppose we choose a prior $p(\mu)$ given by Gaussian. Let's consider the posterior distribution $p(\mu \mid D)$. In previous section, we have showed that the corresponding posterior will be a product of two exponential parts of quadratic function of μ which means that it will be Gaussian. Now we take the prior distribution to be

$$p(\mu) = N(\mu \mid \mu_0, \sigma_0^2) \tag{2}$$

and the posterior distribution is given by

$$\begin{aligned} p(x \mid \mathbf{D}) &\propto p(\mathbf{D} \mid \mu) p(\mu) \\ &\propto \exp[-\frac{1}{2}(\frac{\sum_{n=1}^{N} (x_n - \mu)^2}{\sigma^2}) + \frac{(\mu - \mu_0)^2}{\sigma_0^2}] \end{aligned} \tag{3}$$

We will separate the expression into two parts: the quadratic part and the linear part. Let's first discuss the quadratic part. It is given by

$$-\frac{1}{2}(\frac{\sum_{n=1}^{N}\mu^2}{\sigma^2}+\frac{\mu^2}{\sigma_0^2})=-\frac{1}{2}(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2})\mu^2 \tag{4}$$

As we discussed above, the quadratic part is

$$-\frac{1}{2}\mu^T\Sigma^{-1}\mu \tag{5}$$

It follows that

$$\frac{1}{\sigma_N^2}=\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}=\frac{N\sigma_0^2+\sigma^2}{\sigma^2\sigma_0^2} \tag{6}$$

The linear part is

$$\begin{aligned}&-\frac{1}{2}(\frac{-2\sum_{n=1}^{N}x_n\mu}{\sigma^2}-\frac{2\mu\mu_0}{\sigma_0^2})\\&=\frac{\sum_{n=1}^{N}x_n}{\sigma^2}\mu+\frac{\mu\mu_0}{\sigma_0^2}\\&=(\frac{\sum_{n=1}^{N}x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\mu\end{aligned} \tag{7}$$

And the linear part is

$$\mu^T\Sigma^{-1}\mu_N \tag{8}$$

It turns out that

$$\begin{aligned}\mu_N&=(\frac{\sum_{n=1}^{N}x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\sigma_N^2\\&=(\frac{\sum_{n=1}^{N}x_n}{\sigma^2}+\frac{\mu_0}{\sigma_0^2})\frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}\\&=\frac{N\sigma_)^2}{N\sigma_0^2+\sigma^2}\mu_{ML}+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0\end{aligned} \tag{9}$$

where μ_{ML} is the maximum likelihood solution for μ given by the sample mean

$$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n \tag{10}$$

Now let's spend a moment studying the form of the posterior mean and variance. Here are several property of the posterior mean and variance.

- If the observed data points $N=0$, then μ_N will reduce to the prior mean μ and σ_N reduce to the prior variance σ.

- If $N \to \infty$, the posterior mean μ_N is given by μ_{ML} and the posterior variance $\sigma_N \to 0$, which means that the posterior distribution becomes infinitely peaked around the maximum likelihood solution.
- Note that for finite N, if we set $\sigma_0^2 \to \infty$, which means that the prior distribution has the infinite variance, then the posterior mean μ_N reduces to μ_{ML} while the the posterior variance is given by $\sigma_N^2 = \sigma^2/N$

1.2 Case 2: Mean is Known and Precision is Unknown

Assume that

- μ: known
- σ^2: unknown

Let $\lambda = \frac{1}{\sigma^2}$. Then we have the likelihood:

$$p(D \mid \lambda) = \prod_{n=1}^{N} N(x_n \mid \mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\{-\frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2\} \tag{11}$$

The corresponding conjugate prior should be proportional to the product of a power of λ and the exponential of a linear function of λ. Let's define the gamma distribution

$$\mathrm{Gam}(\lambda \mid a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda) \tag{12}$$

where $\Gamma(a)$ is defined by

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du \tag{13}$$

and it ensures that $\mathrm{Gam}(\lambda \mid a, b)$ is normalized. The mean and variance of the gamma distribution are given by

$$\mathbb{E}[\lambda] = \frac{a}{b} \tag{14}$$

$$\mathrm{VAR}[\lambda] = \frac{a}{b^2} \tag{15}$$

Now consider a prior distribution $\mathrm{Gam}(\lambda \mid a_0, b_0)$. We multiply by the likelihood function, we can attain the posterior distribution

$$p(\lambda \mid \mathbf{D}) \propto \lambda^{a_0 - 1} \lambda^{N/2} \exp[-b_0 \lambda - \frac{\lambda}{2} \sum_{n=1}^{N} (x_n - \mu)^2] \tag{16}$$

which is a gamma distribution of the form $\mathrm{Gam}(\lambda \mid a_N, b_N)$ where

$$a_N = a_0 + \frac{N}{2} \tag{17}$$

$$b_N = b_0 + \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^2 = b_0 + \frac{N}{2} \sigma_{ML}^2 \tag{18}$$

where σ_{ML}^2 is the maximum likelihood estimator of the variance.

1.3 Case 3: Both Mean and Precision Are Unknown

Assume that

- μ: unknown
- σ^2: unknown

To find the conjugate prior, we consider the depence of the likelihood function on μ and λ

$$\begin{aligned} p(\mathbf{D} \mid \mu, \lambda) &= \prod_{n=1}^{N} (\frac{\lambda}{2\pi})^{1/2} \exp\{-\frac{\lambda}{2}(x_n - \mu)^2\} \\ &\propto (\lambda^{1/2} \exp\{-\frac{\lambda\mu^2}{2}\})^N \exp\{\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2}\sum_{n=1}^{N} x_n^2\} \end{aligned} \tag{19}$$

We wish to identify a prior distribution $p(\mu, \lambda)$ that has the same functional dependence on μ and λ as the likelihood function and it takes the form

$$\begin{aligned} p(\mu, \lambda) &\propto (\lambda^{1/2} \exp\{-\frac{\lambda\mu^2}{2}\})^{\beta} \exp\{c\lambda\mu - d\lambda\} \\ &= \exp\{-\frac{\beta\lambda}{2}(\mu - \frac{c}{\beta})^2\}\lambda^{\beta/2} \exp\{-(d - \frac{c^2}{2\beta})\lambda\} \end{aligned} \tag{20}$$

where c, d, β are constants. Consider the prior distribution $p(\mu, \lambda)$. It turns out that

$$p(\mu, \lambda) = p(\mu \mid \lambda) p(\lambda) \tag{21}$$

where $p(\mu \mid \lambda)$ is a Gaussian whose precision is a linear function of λ and $p(\lambda)$ is a gamma distribution. By normalization, we obtain the form

$$p(\mu, \lambda) = \mathscr{N}(\mu \mid \mu_0, (\beta\lambda)^{-1}) \mathrm{Gam}(\lambda \mid a, b) \tag{22}$$

where $\mu_0 = c/\beta$, $a = 1 + \beta/2$, $b = d - c^2/2\beta$. Then we can combine them to obtain t he posterior distribution.

$$\begin{aligned} &p(\mu, \lambda \mid \mathbf{D}) \\ &\propto p(\mathbf{D} \mid \mu, \lambda) p(\mu, \lambda) \\ &\propto (\lambda^{1/2} \exp\{-\frac{\lambda\mu^2}{2}\})^N \exp\{\lambda\mu \sum_{n=1}^{N} x_n - \frac{\lambda}{2}\sum_{n=1}^{N} x_n^2\} (\lambda^{1/2} \exp\{-\frac{\lambda\mu^2}{2}\})^{\beta} \exp\{c\lambda\mu - d\lambda\} \\ &= (\lambda^{1/2} \exp\{-\frac{\lambda\mu^2}{2}\})^{N+\beta} \exp\{(c + \sum_{n=1}^{N} x_n)\lambda\mu - (d + \sum_{n=1}^{N} \frac{x_n^2}{2})\lambda\} \end{aligned} \tag{23}$$

It turns out that it has a *Gaussian-gamma* distribution form

$$p(\mu, \lambda \mid \mathbf{D}) = \mathscr{N}(\mu \mid \mu_0^{'}, (\beta^{'}\lambda)^{-1}) \mathrm{Gam}(\lambda \mid a^{'}, b^{'}) \tag{24}$$

where

$$\beta^{'} = N + \beta \tag{25}$$

$$c^{'} = c + \sum_{n=1}^{N} x_n \tag{26}$$

$$d^{'} = d + \sum_{n=1}^{N} \frac{x_n^2}{2} \tag{27}$$

$$\mu_0^{'} = \frac{c^{'}}{\beta'} = \frac{c + \sum_{n=1}^{N} x_n}{N + \beta} \tag{28}$$

$$a^{'} = 1 + \frac{\beta^{'}}{2} = 1 + \frac{N + \beta}{2} \tag{29}$$

$$b^{'} = d^{'} - \frac{c^{'2}}{2\beta'} = (d + \sum_{n=1}^{N} \frac{x_n^2}{2} - \frac{(c + \sum_{n=1}^{N} x_n)^2}{2(\beta + N)}) \tag{30}$$