Bayesian Inference Examples

Rohit Budhiraja

Machine Learning for Wireless Communications (EE798L)

Feb 28, 2024

Recap of last lecture and today's agenda

- Recap of last class
  - Derive Marginal likelihood for Olympic data model - Chap-4 of FCML
  - Show its application for 5G wireless systems - sparse Bayesian learning
- Today's agenda
- Perform Bayesian learning by taking examples of Gaussian random variables
  - Ref: Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)

Bayesian Inference for mean of a univariate Gaussian

- Given N i.i.d observations $\mathbf{x} = \{x_1, x_2, ..., x_N\}$ which are assumed to be drawn from $\mathcal{N}(x|\mu, \sigma^2)$
- Likelihood of each observation

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x_n|\mu, \sigma^2) \propto \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

- Joint likelihood of N observations

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

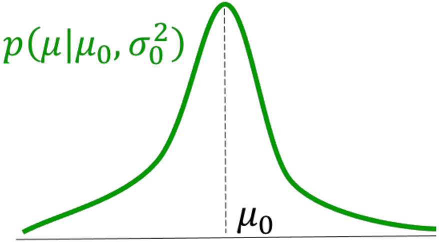
- Easy to see that, each x_n drawn from $\mathcal{N}(x|\mu, \sigma^2)$ is equivalent to the following:

$$x_n = \mu + \epsilon_n, \text{where } \epsilon_n \sim \mathcal{N}(x|0, \sigma^2)$$

- x_n is like a noisy version of μ with zero mean Gaussian noise added to it
- Let's estimate μ given $\mathbf{x}$ using fully Bayesian inference (not point estimation)

A prior distribution for the mean μ

- For Bayesian inference, we need a prior over μ
- We choose a Gaussian prior $p(\mu|\mu_0, \sigma_0^2) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$



- Prior says that a priori we believe μ is close to μ_0
- Prior's variance σ_0^2 denotes how certain we are about our belief
- We will assume that the prior's hyperparameters $(\mu_0, \sigma_0^2)$are known
- Since σ^2 in the likelihood $\mathcal{N}(x|0, \sigma^2)$ is known
  - Gaussian prior $\mathcal{N}(\mu|\mu_0, \sigma_0^2)$ on μ is also conjugate to the likelihood
  - Posterior distribution of unknown mean parameter μ will also be Gaussian

Posterior distribution for the mean (1)

- Posterior distribution of the unknown mean parameter μ

$$\begin{aligned}
p(\mu|\mathbf{x}) &= \frac{p(\mathbf{x}|\mu)p(\mu)}{p(\mathbf{x})} \propto \prod_{n=1}^{N} \exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right] \times \exp\left[-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right] \times \exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right] \\
&= \exp\left[-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n^2+\mu^2-2\mu x_n) - \frac{1}{2\sigma_0^2}(\mu^2+\mu_0^2-2\mu\mu_0)\right] \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\left(\mu^2 N - 2\mu\sum_{n=1}^{N} x_n\right) - \frac{1}{2\sigma_0^2}\left(\mu^2-2\mu\mu_0\right)\right] \qquad (1)
\end{aligned}$$

- Let us denote the posterior in compact form as

$$p(\mu|\mathbf{x}) \propto \exp\left[-\frac{(\mu-\mu_N)^2}{2\sigma_N^2}\right] = \exp\left[-\frac{1}{2\sigma_N^2}(\mu^2+\mu_N^2-2\mu\mu_N)\right] \qquad (2)$$

- We compare quadratic and linear part of μ in (1) and (2)

Comparing quadratic part of μ

- Posterior distribution of the unknown mean parameter μ

$$p(\mu|\mathbf{x}) \quad \propto \quad \exp\left[-\frac{1}{2\sigma^2}\left(\mu^2 N - 2\mu\sum_{n=1}^{N} x_n\right) - \frac{1}{2\sigma_0^2}\left(\mu^2 - 2\mu\mu_0\right)\right] \tag{3}$$

- Posterior in compact form as

$$p(\mu|\mathbf{x}) \quad \propto \quad \exp\left[-\frac{1}{2\sigma_N^2}(\mu^2 + \mu_N^2 - 2\mu\mu_N)\right] \tag{4}$$

- Comparing quadratic part of μ in (3) and (4), we have

$$\begin{aligned} -\frac{1}{2\sigma_N^2} &= -\frac{1}{2\sigma^2}N - \frac{1}{2\sigma_0^2} \\ \frac{1}{\sigma_N^2} &= \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \end{aligned}$$

- Posterior's precision is sum of prior's precision and sum of noise precisions of all observations

Comparing linear part of μ

- Posterior distribution of the unknown mean parameter μ

$$p(\mu|\mathbf{x}) \quad \propto \quad \exp\left[-\frac{1}{2\sigma^2}\left(\mu^2 N - 2\mu\sum_{n=1}^{N} x_n\right) - \frac{1}{2\sigma_0^2}\left(\mu^2 - 2\mu\mu_0\right)\right] \tag{5}$$

- Posterior in compact form as

$$p(\mu|\mathbf{x}) \quad \propto \quad \exp\left[-\frac{1}{2\sigma_N^2}(\mu^2 + \mu_N^2 - 2\mu\mu_N)\right] \tag{6}$$

- Comparing linear part of μ (5) and (6), we have

$$\begin{aligned}
\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n + \frac{\mu_0}{\sigma_0^2} &= \frac{\mu_N}{\sigma_N^2} \\
\frac{1}{\sigma^2}\sum_{n=1}^{N} x_n + \frac{\mu_0}{\sigma_0^2} &= \left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu_N \\
\mu_N &\overset{(a)}{=} \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\bar{x}
\end{aligned}$$

- Equality (a) is obtained by setting $\bar{x} = \frac{\sum_{n=1}^{N} x_n}{N}$.
- First term in posterior mean is contribution from prior, second is from data

Posterior distribution for large number of observations N

- Posterior variance from last slide

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

- Posterior mean from last slide

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\bar{x}$$

- What happens to the posterior as N (number of observations) grows very large?
  - Data (likelihood part) overwhelms the prior
  - Posterior's variance σ_N^2 will approximately be σ^2/N (and goes to 0 as $N \rightarrow \infty$)
  - Posterior's mean μ_N approaches $\bar{x}$,which is also the maximal likelihood solution

Bayesian inference for variance of a Gaussian

- Given N i.i.d observations which $\mathbf{x} = \{x_1, x_2, ..., x_N\}$, assumed to be drawn from $\mathcal{N}(x|\mu, \sigma^2)$
- Joint likelihood of N joint observations

$$p(x_n|\mu, \sigma^2) = \mathcal{N}(x|\mu, \sigma^2) \text{ and } p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^{N} p(x_n|\mu, \sigma^2)$$

- We want to estimate the variance σ^2. Assume μ to be known.
- If we want a conjugate prior $p(\sigma^2)$, its functional form must be same as likelihood

$$p(x_n|\mu, \sigma^2) \propto (\sigma^2)^{-1/2} \exp\left[-\frac{(x_n - \mu)^2}{2\sigma^2}\right]$$

- An inverse-gamma dist $IG(\alpha, \beta)$ has this form

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} \exp\left[-\frac{\beta}{\sigma^2}\right]$$

- Due to conjugacy, posterior will also be $IG(\alpha_N, \beta_N)$ with expression

$$p(\sigma^2|\mathbf{x}) \propto (\sigma^2)^{-(\alpha_N+1)} \exp\left(-\frac{\beta_N}{\sigma^2}\right)$$

Posterior distribution of the variance σ^2

- Posterior distribution for the unknown variance parameter σ^2

$$
\begin{aligned}
p(\sigma^2|\mathbf{x}) &= \frac{p(\mathbf{x}|\sigma^2)p(\sigma^2)}{p(\mathbf{x})} \\
&\propto (\sigma^2)^{-(\alpha+1)}\exp\left[-\frac{\beta}{\sigma^2}\right] \times \prod_{n=1}^{N}\left((\sigma^2)^{-1/2}\exp\left[-\frac{(x_n-\mu)^2}{2\sigma^2}\right]\right) \\
&= (\sigma^2)^{-(\alpha+1)}(\sigma^2)^{\left(-\frac{N}{2}\right)}\exp\left[-\frac{\beta}{\sigma^2} - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2\right] \\
&\overset{(a)}{=} (\sigma^2)^{-(\alpha_N+1)}\exp\left(-\frac{\beta_N}{\sigma^2}\right)
\end{aligned}
$$

- Equality (a) is obtained by denoting $\alpha_N = \alpha + \frac{N}{2}$, and $\beta_N = \beta + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2$
- Posterior is now

$$p(\sigma^2|\mathbf{x}) = IG(\alpha_N, \beta_N)$$

Working with Gaussians: Variance vs Precision

- Often, it is easier to work with the precision (=1/variance) rather than variance
- Likelihood is

$$p(x_n|\mu, \lambda^{-1}) = \mathcal{N}(x|\mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n - \mu)^2\right]$$

- Joint likelihood is

$$p(\mathbf{x}|\sigma^2) = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n - \mu)^2\right]$$

- If mean is known, for precision, Gamma(α, β) is a conjugate prior to Gaussian likelihood

$$p(\lambda) \propto (\lambda)^{(\alpha-1)} \exp[-\beta\lambda]$$

- Due to conjugacy, posterior will also be Gamma(α_N, β_N) with expression

$$p(\lambda|\mathbf{x}) \propto (\lambda)^{(\alpha_N-1)} \exp[-\beta_N\lambda]$$

Posterior distribution for the unknown precision λ

- Posterior distribution for the unknown precision λ

$$
\begin{aligned}
p(\lambda|\mathbf{x}) &= \frac{p(\mathbf{x}|\sigma^2)p(\lambda)}{p(\mathbf{x})} \\
&\propto \left(\prod_{n=1}^{N}\sqrt{\frac{\lambda}{2\pi}}\exp\left[-\frac{\lambda}{2}(x_n-\mu)^2\right]\right)\times\left(\lambda^{(\alpha-1)}\exp[-\beta\lambda]\right) \\
&= \left(\left(\frac{\lambda}{2\pi}\right)^{N/2}\exp\left[-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right]\right)\times\left(\lambda^{(\alpha-1)}\exp[-\beta\lambda]\right) \\
&= \lambda^{(\alpha-1+N/2)}\exp\left[-\left(\beta+\frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2}\right)\lambda\right] \\
&\overset{(a)}{=} (\lambda)^{(\alpha_N-1)}\exp[-\beta_N\lambda]
\end{aligned}
$$

- Equality (a) is obtained by denoting $\alpha_N=\alpha+\frac{N}{2}$, $\beta_N=\beta+\frac{\sum_{n=1}^{N}(x_n-\mu)^2}{2}$
- Posterior is now

$$p(\lambda|\mathbf{x})=\mathrm{Gamma}(\alpha_N,\beta_N)$$

Bayesian Inference for both parameters of a Gaussian

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Given N i.i.d observations which $\mathbf{x} = \{x_1, x_2, ..., x_N\}$, assumed to be drawn from $\mathcal{N}(x|\mu, \lambda)$
- Assume both mean μ and precision λ to be unknown. Likelihood can be written as

$$
\begin{aligned}
p(\mathbf{x}|\mu,\lambda) &= \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n-\mu)^2\right] = \prod_{n=1}^{N} \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{\lambda}{2}(x_n^2+\mu^2-2x_n\mu)\right] \\
&\propto \left[\lambda^{1/2} \exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left[\lambda\mu\sum_{n=1}^{N} x_n - \frac{\lambda}{2}\sum_{n=1}^{N} x_n^2\right]
\end{aligned}
$$

- Would like a jointly conjugate prior distribution $p(\mu, \lambda)$ - must have same form as above likelihood
- Normal-gamma (NG) distribution
  - Since it can be written as a product of a normal and a gamma (next slide)

Bayesian Inference for Both Parameters of a Gaussian

- Normal Gamma Distribution is defined as-

$$\begin{aligned}
\mathrm{NG}(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0) &= \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1}) \times \mathrm{Gamma}(\lambda|\alpha_0, \beta_0) \\
&= \sqrt{\frac{\kappa_0\lambda}{2\pi}}\exp\left(-\frac{1}{2}\kappa_0\lambda(\mu-\mu_0)^2\right) \times \frac{{\beta_0}^{\alpha_0}}{\Gamma(\alpha_0)}\lambda^{\alpha_0-1}\exp(-\beta_0\lambda) \\
&\propto \lambda^{1/2}\exp\left(-\frac{1}{2}\kappa_0\lambda(\mu-\mu_0)^2\right)\lambda^{\alpha_0-1}\exp(-\beta_0\lambda)
\end{aligned}$$

- NG also has a vector version
  - Normal-Wishart distribution to jointly model a real-valued vector and a PSD matrix
- Posterior is given as

$$\begin{aligned}
p(\mu, \lambda|\mathbf{x}) &\propto p(\mathbf{x}|\mu, \lambda)p(\mu, \lambda) \\
&= p(\mathbf{x}|\mu, \lambda) \times \mathrm{NG}(\mu, \lambda|\mu_0, \kappa_0, \alpha_0, \beta_0)
\end{aligned}$$

- Posterior can be shown as a product of normal and Gamma function (Tutorial problem)

$$p(\mu, \lambda|\mathbf{x}) = \mathrm{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N)$$

- Here $\mu_N = \frac{\kappa_0\mu_0+N\bar{x}}{\kappa_0+N}$, $\kappa_N = \kappa_0 + N$, $\alpha_N = \alpha_0 + N/2$, $\beta_N = \beta_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \bar{x})^2 + \frac{\kappa_0 N(\bar{x}-\mu_0)^2}{2(\kappa_0+N)}$