ECE595 / STAT598: Machine Learning I
Lecture 12.2: Bayesian Parameter Estimation - Choosing Priors

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE
UNIVERSITY

Outline

Generative Approaches

- Lecture 9 Bayesian Decision Rules
- Lecture 10 Evaluating Performance
- Lecture 11 Parameter Estimation
- Lecture 12 Bayesian Prior
- Lecture 13 Connecting Bayesian and Linear Regression

Today's Lecture

- Basic Principles
  - Posterior
  - 1D Illustration
  - Interpretations
- Choosing Priors
  - Prior for Mean
  - Prior for Variance
  - Conjugate Prior

Prior for μ

So far we have considered

$$p(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^{N}\frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}.$$

- Unknown μ, and known σ^2.
- The likelihood is Gaussian (by problem setup).
- The prior for μ is Gaussian (by our choice).
- Good, because posterior remains a Gaussian.
- What happens if σ^2 is unknown but μ is known?

Prior for σ^2

- Let us define the precision: $\lambda = \frac{1}{\sigma^2}$.
- The likelihood is

$$\begin{aligned} p(\mathcal{D}|\lambda) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^N \frac{(x_n-\mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{\lambda^{N/2}}{(\sqrt{2\pi})^N}\right) \exp\left\{-\sum_{n=1}^N \frac{\lambda}{2}(x_n-\mu)^2\right\} \\ &= \frac{1}{(2\pi)^{N/2}}\, \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^N (x_n-\mu)^2\right\}. \end{aligned}$$

- We want to choose $p(\lambda)$ in a similar form:

$$p(\lambda) = A\lambda^B \exp\{-C\lambda\}$$

so that the posterior $p(\lambda|\mathcal{D})$ is easy to compute.

Prior for σ^2

- We want to choose $p(\lambda)$ in a similar form:

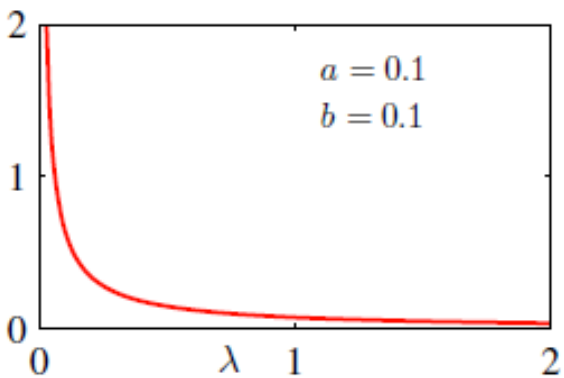
$$p(\lambda) = A\lambda^B \exp\{-C\lambda\}$$

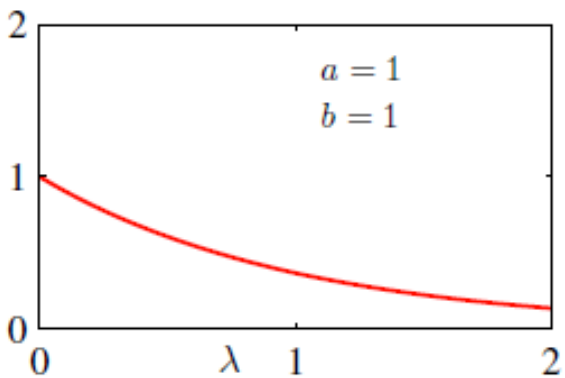
- The candidate is ...

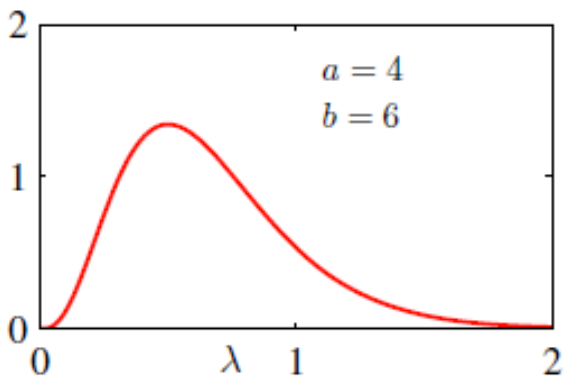
$$p(\lambda) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

- This distribution is called the Gamma distribution $\mathrm{Gam}(\lambda|a, b)$.
- We can show that

$$\mathbb{E}[\lambda] = \frac{a}{b}, \qquad \mathrm{Var}[\lambda] = \frac{a}{b^2}.$$







Prior for σ^2

- If we consider this pair of likelihood and prior

$$p(\mathcal{D}|\lambda) = \frac{1}{(2\pi)^{N/2}} \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

$$p(\lambda) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0-1} \exp(-b_0\lambda),$$

- then the posterior is

$$p(\lambda|\mathcal{D}) \propto \lambda^{(a_0+N/2)-1} \exp\left\{-\left(b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)\lambda\right\}$$

- Just another Gamma distribution.
- You can now do estimation on this Gamma by finding λ which maximizes the posterior. Details: See Appendix.

Prior for Both μ and σ^2

- Again, let $\lambda = \frac{1}{\sigma^2}$.
- The likelihood is

$$p(\mathcal{D}|\mu,\lambda) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \exp\left\{-\sum_{n=1}^N \frac{(x_n-\mu)^2}{2\sigma^2}\right\}$$

$$\propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^N \exp\left\{\lambda\mu\sum_{n=1}^N x_n - \frac{\lambda}{2}\sum_{n=1}^N x_n^2\right\}$$

- Candidate for the prior is

$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^\beta \exp\{c\lambda\mu - d\lambda\}$$

$$= \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}}_{\mathcal{N}(\mu|\mu_0,\sigma_0^2)} \underbrace{\lambda^{\beta/2}\exp\left\{-\left(d-\frac{c^2}{2\beta}\right)\lambda\right\}}_{\text{Gam}(\lambda|a,b)}$$

- The prior distribution is called the Normal-Gamma distribution.

Priors for High-dimension Gaussians

- Let $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$.
- The likelihood is

$$p(\boldsymbol{x}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{x}_m|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$$

- Prior for $\boldsymbol{\mu}$: Gaussian.

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}).$$

- Prior for $\boldsymbol{\Sigma}$: Wishart.

$$p(\boldsymbol{\Lambda}) = \mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu).$$

- Prior for both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$: Normal-Wishart.

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, (\beta\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|\boldsymbol{W}, \nu).$$

Conjugate Prior

- You have a likelihood $p_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}|\theta)$
- You want to choose a prior $p_{\boldsymbol{\Theta}}(\theta)$ so that ...
- the posterior $p_{\boldsymbol{\Theta}|\boldsymbol{X}}(\theta|\boldsymbol{x})$ takes the same form as the prior
- Such prior is called the **conjugate prior**
- Conjugate *with respect to* the likelihood
- Finding the conjugate prior may not be easy!
- Good news: Any likelihood belong to the **exponential family** will have a conjugate prior also in the exponential family.
- Exponential family: Gaussian, Exponential, Poisson, Bernoulli, etc
- For more discussions, see Bishop Chapter 2.4

Reading List

Bayesian Parameter Estimation

- Duda-Hart-Stork, Pattern Classification, Chapter 3.3 - 3.5
- Bishop, Pattern Recognition and Machine Learning, Chapter 2.4
- M. Jordan (Berkeley), https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter9.pdf
- CMU Note, http://www.cs.cmu.edu/~aarti/Class/10701_Spring14/slides/MLE_MAP_Part1.pdf
- A. Kak (Purdue), https://engineering.purdue.edu/kak/Tutorials/Trinity.pdf

Appendix

Prior for σ^2: Solution

- The posterior is

$$p(\lambda|\mathcal{D}) \propto \lambda^{(a_0+N/2)-1} \exp\left\{-\left(b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right)\lambda\right\}$$
$$\propto \lambda^{a_N-1}\exp\left\{-b_N\lambda\right\}.$$

- The maximum-a-posteriori estimate of λ is

$$\begin{aligned}\widehat{\lambda} &= \underset{\lambda}{\text{argmax}} \quad p(\lambda|\mathcal{D}) \\ &= \underset{\lambda}{\text{argmax}} \quad \lambda^{a_N-1}\exp\left\{-b_N\lambda\right\} \\ &= \underset{\lambda}{\text{argmax}} \quad (a_N-1)\log\lambda - b_N\lambda.\end{aligned}$$

- Taking derivative and setting to zero:

$$\frac{d}{d\lambda}\Big((a_N-1)\log\lambda - b_N\lambda\Big) = \frac{a_N-1}{\lambda} - b_N = 0.$$

Prior for σ^2: Solution

- Therefore,

$$\lambda = \frac{a_N - 1}{b_N}.$$

- where the parameters are

$$a_N = a_0 + \frac{N}{2},$$

$$b_N = b_0 + \frac{1}{2}\sum_{n=1}^{N}(x_n - \mu)^2 = b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}.$$

- Hence, the MAP estimate is

$$\lambda = \frac{a_0 + \frac{N}{2}}{b_0 + \frac{N}{2}\sigma^2_{\mathrm{ML}}}.$$

- As $N \to \infty$, $\lambda \to \frac{1}{\sigma^2_{\mathrm{ML}}}$.
- As $N \to 0$, $\lambda \to \frac{a_0}{b_0}$.

Prior for Both μ and σ^2: Detailed Derivation

- Again, let $\lambda = \frac{1}{\sigma^2}$.
- The likelihood is

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

Prior for Both μ and σ^2: Detailed Derivation

- The likelihood is

$$p(\mathcal{D}|\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^N \exp\left\{\lambda\mu\sum_{n=1}^N x_n - \frac{\lambda}{2}\sum_{n=1}^N x_n^2\right\}$$

- Candidate for the prior is

$$\begin{aligned} p(\mu,\lambda) &\propto \left[\lambda^{1/2}\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^\beta \exp\{c\lambda\mu - d\lambda\} \\ &= \left[\exp\left\{-\frac{\lambda\mu^2}{2}\right\}\right]^\beta \left[\lambda^{\beta/2}\exp\{c\lambda\mu - d\lambda\}\right] \\ &= \underbrace{\exp\left\{-\frac{\beta\lambda}{2}(\mu - c/\beta)^2\right\}}_{\mathcal{N}(\mu|\mu_0,\sigma_0^2)} \underbrace{\lambda^{\beta/2}\exp\left\{-\left(d - \frac{c^2}{2\beta}\right)\lambda\right\}}_{\text{Gam}(\lambda|a,b)} \end{aligned}$$

-

$$\mu_0 = c/\beta, \quad \sigma_0^2 = (\beta\lambda)^{-1}, \quad a = 1 + \beta/2, \quad b = d - c^2/2\beta$$

Prior for Both μ and σ^2: Detailed Derivation

- The prior distribution is

$$p(\mu, \lambda) = \mathcal{N}\left(\mu|\mu_0, (\beta\lambda)^{-1}\right) \mathrm{Gam}(\lambda|a, b)$$

- This is called the Normal-Gamma distribution
- Here is a 2D plot of $p(\mu, \lambda)$ when $\mu_0 = 0$, $\beta = 2$, $a = 5$, $b = 6$.