Multiparameter Models (Contd.)

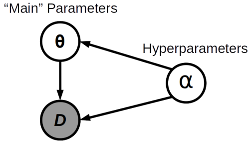
Piyush Rai

Topics in Probabilistic Modeling and Inference (CS698X)

Jan 23, 2018

Recap: Learning Hyperparameters via MLE-II

- Denoting all the "main" parameters by θ and all the hyperparameters by α



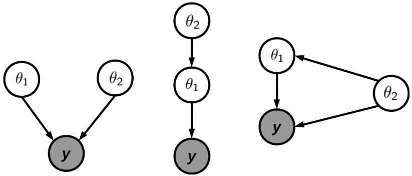
- MLE-II learns hyperparameters by maximizing the marginal likelihood

$$
\begin{aligned}
\hat{\alpha} = \arg\max_{\alpha} p(\mathcal{D}|\alpha) &= \arg\max_{\alpha} \int p(\mathcal{D}, \theta|\alpha) d\theta \\
&= \arg\max_{\alpha} \int p(\mathcal{D}|\theta, \alpha) p(\theta|\alpha) d\theta
\end{aligned}
$$

- Note: As with standard MLE, we usually maximize the log of the marginal likelihood $p(\mathcal{D}|\alpha)$
- The updates of θ and α are usually coupled with each other (as we saw in linear regression)

Multiparameter Models

- Multiparameter models consist of two or more unknowns, say θ_1 and θ_2
- Given data $\boldsymbol{y}$, some examples for the simple two parameter case



- Assume the likelihood model to be of the form $p(\boldsymbol{y}|\theta_1, \theta_2)$ (e.g., case 1 and 3 above)
- Assume a joint prior distribution $p(\theta_1, \theta_2)$
- The joint posterior $p(\theta_1, \theta_2|\boldsymbol{y}) \propto p(\boldsymbol{y}|\theta_1, \theta_2)p(\theta_1, \theta_2)$
  - Can be found easily if the joint prior is conjugate to the likelihood (will seen an example today)
  - Otherwise needs more work, e.g., MLE-II, MCMC, VB, etc. (already saw MLE-II, will see more later)
- Other quantities of interest: marginal post. (e.g., $p(\theta_1|\boldsymbol{y})$), conditional post. (e.g., $p(\theta_1|\theta_2, \boldsymbol{y})$), marginal likelihood ($p(\boldsymbol{y})$), posterior predictive distribution ($p(y_*|\boldsymbol{y})$), etc.

A Simple Multiparameter Model (with easy computations!)

- Gaussian with unknown scalar mean and unknown scalar precision (two parameters)
- Consider N i.i.d. observations $\mathbf{X} = \{x_1, \ldots, x_N\}$ drawn from a one-dim Gaussian $\mathcal{N}(x|\mu, \lambda^{-1})$
- Assume both mean μ and precision λ to be unknown. The likelihood will be

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

- If we want a conjugate joint prior $p(\mu, \lambda)$, it must have the same form as likelihood. Suppose

$$p(\mu,\lambda) \propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0} \exp\left[\lambda\mu c - \lambda d\right]$$

- What's this prior? A normal-gamma (Gaussian-gamma) distribution! (will see its form shortly)
    - Can be used in models where we wish to estimate an unknown mean and unknown precision
    - Note: Its multivariate version is the Normal-Wishart (for multivariate mean and precision matrix)

Normal-gamma (Gaussian-gamma) Distribution

- We saw that the conjugate prior needed to have the form

$$
\begin{aligned}
p(\mu,\lambda) &\propto \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^{\kappa_0}\exp\left[\lambda\mu c-\lambda d\right] \\
&= \underbrace{\exp\left[-\frac{\kappa_0\lambda}{2}(\mu-c/\kappa_0)^2\right]}_{\text{prop. to a Gaussian}}\underbrace{\lambda^{\kappa_0/2}\exp\left[-\left(d-\frac{c^2}{2\kappa_0}\right)\lambda\right]}_{\text{prop. to a gamma}} \qquad \text{(re-arranging terms)}
\end{aligned}
$$

- The above is product of a normal and a gamma distribution[1]

$$p(\mu,\lambda)=\mathcal{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1})\text{Gamma}(\lambda|\alpha_0,\beta_0)=\text{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$$

  where $\mu_0=c/\kappa_0$, $\alpha_0=1+\kappa_0/2$, $\beta_0=d-c^2/2\kappa_0$ are prior's hyperparameters

- $p(\mu,\lambda)=\text{NG}(\mu_0,\kappa_0,\alpha_0,\beta_0)$ is a conjugate for the mean-precision pair (μ,λ)
  - A useful prior in many problems involving Gaussians with unknown mean and precision

[1] shape-rate parametrization assumed for the gamma

Joint Posterior

- Due to conjugacy, the joint posterior $p(\mu, \lambda|\mathbf{X})$ will also be normal-gamma

$$p(\mu, \lambda|\mathbf{X}) \propto p(\mathbf{X}|\mu, \lambda)p(\mu, \lambda)$$

- Plugging in the expressions for $p(\mathbf{X}|\mu, \lambda)$ and $p(\mu, \lambda)$, we get

$$p(\mu, \lambda|\mathbf{X}) \quad = \quad \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\text{Gamma}(\lambda|\alpha_N, \beta_N)$$

where the updated posterior hyperparameters are given by[2]

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

[2] For full derivation, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)

Other Quantities of Interest[3]

- Already saw that joint post. $p(\mu, \lambda|\mathbf{X}) = \text{NG}(\mu_N, \kappa_N, \alpha_N, \beta_N) = \mathcal{N}(\mu|\mu_N, (\kappa_N\lambda)^{-1})\text{Gamma}(\lambda|\alpha_N, \beta_N)$
- Marginal posteriors for μ and λ

$$\begin{aligned} p(\lambda|\mathbf{X}) &= \int p(\mu, \lambda|\mathbf{X})d\mu = \text{Gamma}(\lambda|\alpha_N, \beta_N) \\ p(\mu|\mathbf{X}) &= \int p(\mu, \lambda|\mathbf{X})d\lambda = \int p(\mu|\lambda, \mathbf{X})p(\lambda|\mathbf{X})d\lambda = \underbrace{t_{2\alpha_N}(\mu|\mu_N, \beta_N/(\alpha_N\kappa_N))}_{\text{t distribution}} \end{aligned}$$

- Exercise: What will be the conditional posteriors $p(\mu|\lambda, \mathbf{X})$ and $p(\lambda|\mu, \mathbf{X})$?
- Marginal likelihood of the model

$$p(\mathbf{X}) = \frac{\Gamma(\alpha_N)}{\Gamma(\alpha_0)} \frac{\beta_0^{\alpha_0}}{\beta_N^{\alpha_N}} \left(\frac{\kappa_0}{\kappa_N}\right)^{\frac{1}{2}} (2\pi)^{-N/2}$$

- Posterior predictive distribution of a new observation x_*

$$p(x_*|\mathbf{X}) = \int \underbrace{p(x_*|\mu, \lambda)}_{\text{Gaussian}} \underbrace{p(\mu, \lambda|\mathbf{X})}_{\text{Normal-Gamma}} d\mu d\lambda = t_{2\alpha_N}\left(x_*|\mu_N, \frac{\beta_N(\kappa_N + 1)}{\alpha_N\kappa_N}\right)$$

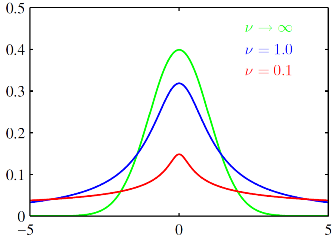
[3] For full derivations, refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007)

An Aside: general-t and Student-t distribution

- Equivalent to an infinite sum of Gaussian distributions, with same means but different precisions

$$
\begin{aligned}
p(x|\mu, a, b) &= \int \mathcal{N}(x|\mu, \lambda^{-1})\text{Gamma}(\lambda|a, b)d\lambda \\
&= t_{2a}(x|\mu, b/a) = t_\nu(x|\mu, \sigma^2) \quad \text{(general-t distribution)}
\end{aligned}
$$

- $\mu = 0, \sigma^2 = 1$ gives the Student-t distribution (t_ν). Note: If $x \sim t_\nu(\mu, \sigma^2)$ then $\frac{x-\mu}{\sigma} \sim t_\nu$
- An illustration of student-t



- t distribution has a "fatter" tail than a Gaussian and also sharper around the mean
  - Also a useful prior for sparse modeling

Inferring Parameters of Gaussian: Some Other Cases

- We only considered the simple 1-D Gaussian distribution
- The approach also extends to inferring parameters of a multivariate Gaussian
  - For the unknown mean and precision matrix, normal-Wishart distribution can be used as prior
- Posterior updates have forms similar to that in the 1-D case
- When working with mean-variance, we can use normal-inverse gamma as conjugate prior (or normal-inverse Wishart when working with mean-covariance matrix in case of multivariate Gaussian distribution)
- Other priors can also be used as well when inferring parameters of Gaussians, e.g.,
  - normal-Inverse χ^2 distribution is commonly used in Statistics community for scalar mean-variance
  - Uniform priors can also be used
  - Look at BDA Chapter 3 for such examples
- Also refer to "Conjugate Bayesian analysis of the Gaussian distribution" - Murphy (2007) for various examples and more detailed derivations

Multiparameter Models: Handling the non-easy cases

- What if we don't have a conjugate pair of likelihood and joint prior?
- Won't be able to get a closed form joint posterior $p(\theta_1, \theta_2 | \boldsymbol{y})$
- In such cases, can still (sometimes approximately) compute the joint posterior $p(\theta_1, \theta_2 | \boldsymbol{y})$
- One approach is to iteratively estimate the conditional posteriors, e.g., $p(\theta_1 | \theta_2, \boldsymbol{y})$ and $p(\theta_2 | \theta_1, \boldsymbol{y})$
  - These conditional posteriors, together, give the joint posterior
- Many inference algorithms are based on estimating the conditional posteriors
  - Gibbs sampling (an MCMC algorithm): Based on sampling from conditional posteriors
  - Variational inference: Based on iteratively approximating the conditional posteriors

Gibbs Sampling (Geman and Geman, 1982)

- A general sampling algorithm to simulate samples from multivariate distributions
- Samples one component at a time from its conditional, conditioned on all other components
  - Assumes that the conditional distributions are available in a closed form

Suppose

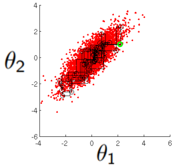
$$\theta \sim N_2(0, \Sigma) \qquad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

Then

$$\begin{aligned} \theta_1|\theta_2 &\sim N\left(\rho\theta_2, [1-\rho^2]\right) \\ \theta_2|\theta_1 &\sim N\left(\rho\theta_1, [1-\rho^2]\right) \end{aligned}$$

are the conditional distributions.

- The generated samples give a sample-based approximation of the multivariate distribution



Gibbs Sampling (Geman and Geman, 1982)

- Can be used to get a sampling-based approximation of a multiparameter posterior distribution
- Gibbs sampler iteratively draws random samples from conditional posteriors
- When run long enough, the sampler produces samples from the joint posterior
- For the simple two-parameter case $\theta = (\theta_1, \theta_2)$, the Gibb sampler looks like this
  - Initialize $\theta_2^{(0)}$
  - For $s = 1, \ldots, S$
    - Draw a random sample for θ_1 as $\theta_1^{(s)} \sim p(\theta_1 | \theta_2^{(s-1)}, \boldsymbol{y})$
    - Draw a random sample for θ_2 as $\theta_2^{(s)} \sim p(\theta_2 | \theta_1^{(s)}, \boldsymbol{y})$
- The set of S random samples $\{\theta_1^{(s)}, \theta_2^{(s)}\}_{s=1}^{S}$ represent the joint posterior distribution $p(\theta_1, \theta_2 | \boldsymbol{y})$
- More on Gibbs sampling when we discuss MCMC sampling algorithms (above is the high-level idea)