Tutorial 2

Tuesday 17th January, 2017

Problem 1. Posterior Distribution of w with very imprecise prior:

Let $y = \mathbf{w}^T\phi(\mathbf{x}) + \varepsilon$ and let dataset $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_i, Y_i), \ldots, (\mathbf{X}_m, Y_m)\}$ was provided. Recall that the posterior distribution for $\mathbf{w}$ under a Gaussian prior was $\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m)$ where

$$\Sigma_m^{-1} = \lambda I + \Phi^T\Phi/\sigma^2$$

and

$$\mu_m = (\lambda\sigma^2 I + \Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

How would you model a very imprecise Gaussian prior on $\Pr(\mathbf{w})$? Explain what happens to the parameters of the posterior $\Pr(\mathbf{w} \mid \mathcal{D})$ as this precision on the prior $\Pr(\mathbf{w})$ tends to 0. What is the connection between this expression and the data likelihood expression?

Solution:

The key is to realize (from discussions in the class) that corresponding to the posterior distribution $\Pr(\mathbf{w} \mid \mathcal{D}) = \mathcal{N}(\mathbf{w} \mid \mu_m, \Sigma_m)$ **was the prior** $\Pr(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid 0, \frac{1}{\lambda}I)$. We discussed how λ (reciprocal of variance) corresponds to precision of the belief of 0 mean for each of the individual w_i's and therefore actually reflects precision of the prior. As $\lambda \to 0$, the prior will tend to have 0 precision or ∞ spread (variance), meaning that the prior is very imprecise. As $\lambda \to 0$ $\Pr(\mathbf{w} \mid \mathcal{D}) \to \mathcal{N}(\mathbf{w} \mid \mu_m^0, \Sigma_m^0)$ where

$$(\Sigma_m^0)^{-1} = \Phi^T\Phi/\sigma^2$$

and

$$\mu_m^0 = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$$

Problem 2. Case for non-IID dataset:

In the class, we discussed the case of Bayesian estimation for a univariate Gaussian from dataset $\mathcal{D}$ that consisted of IID (independent and identically distributed) observations.

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \ldots x_m$ be IID. Let σ^2 be known.
- $\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$
- The conjugate prior is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$, And the **posterior** is: $\Pr(\mu|x_1 \ldots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that

- $\mu_m = (\frac{\sigma^2}{m\sigma_0^2 + \sigma^2}\mu_0) + (\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2}\hat{\mu}_{ML})$ and $\frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$

Prove the above

Answer: We have already done this in the class: `https://www.cse.iitb.ac.in/~cs725/notes/lecture-slides/lecture-06-unannotated.pdf`).

Now suppose, the examples $x_1...x_m$ in the dataset $\mathcal{D}$ were not necessarily independent and whose possible dependence was expressed by known covariance matrix Ω but with a common unknown (to be estimated) mean $\mu \in \Re$. Let $\mathbf{u} = [1, 1, \ldots 1]$ a $m-$dimensional vector of 1's and $\mathbf{x} = [x_1...x_m]$ and

$$Pr(x_1...x_m; \mu, \Omega) = \frac{1}{(2\pi)^{\frac{m}{2}}|\Omega|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu\mathbf{u})^T\Omega^{-1}(\mathbf{x}-\mu\mathbf{u})}$$

Assume that $\Omega \in \Re^{m\times m}$ is positive-definite. Now answer the following questions

1. What would be the maximum likelihood estimate for μ?

   Answer: This would correspond to MLE estimate for a multivariate Gaussian but with a single data point. Additionally, the restriction is that the mean vector is of the form $\mu\mathbf{u}$:

   We have already seen that maximizing a monotonically increasing transformation of the objective should yield the same point of optimality (and proved the same in this tutorial). So taking logs of the likelihood gives us the log likelihood above:

   $$\mu_{MLE} = \underset{\mu}{\text{argmax}} \ -\frac{1}{2}(\mathbf{x}-\mu\mathbf{u})^T\Omega^{-1}(\mathbf{x}-\mu\mathbf{u})$$

   Setting the derivative with respect to μ to 0:

   $$\frac{d}{d\mu}\left(-\frac{1}{2}(\mathbf{x}^T\Omega^{-1}\mathbf{x} - 2\mu\mathbf{x}^T\Omega^{-1}\mathbf{u} + \mu^2\mathbf{u}^T\Omega^{-1}\mathbf{u})\right) = \left(\mathbf{x}^T\Omega^{-1}\mathbf{u} - \mu\mathbf{u}^T\Omega^{-1}\mathbf{u}\right) = 0$$

   $\Rightarrow \mu_{MLE} = \frac{\mathbf{x}^T\Omega^{-1}\mathbf{u}}{\mathbf{u}^T\Omega^{-1}\mathbf{u}}$

2. How would you go about doing Bayesian estimation for μ?

3. What will be an appropriate conjugate prior?

4. What will the posterior be? And what will be the MAP and Bayes estimates?

   Answers to 2, 3 and 4: As hinted in the class, we will expect the conjugate prior of mean μ of the (product of) Gaussian to be Gaussian. Let $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with a fixed and known σ_0^2.

   $\mathcal{N}(\mu_m, \sigma_m^2) = exp\left(\frac{-1}{2\sigma_m^2}(\mu - \mu_m)^2\right) = \Pr(\mu|\mathcal{D}) \propto \Pr(\mathcal{D}|\mu)\Pr(\mu) =$

   $\frac{1}{(2\pi)^{\frac{m}{2}}|\Omega|^{\frac{1}{2}}}\frac{1}{\sqrt{2\pi\sigma_0^2}} exp\left(-\frac{1}{2}(\mathbf{x}-\mu\mathbf{u})^T\Omega^{-1}(\mathbf{x}-\mu\mathbf{u}) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \propto exp\left(-\frac{1}{2}(\mathbf{x}-\mu\mathbf{u})^T\Omega^{-1}(\mathbf{x}-\mu\mathbf{u}) - \frac{(}{}\right.$

Our reference equality is:

$$exp\left(-\frac{1}{2}(\mathbf{x}^T\Omega^{-1}\mathbf{x} - 2\mu\mathbf{x}^T\Omega^{-1}\mathbf{u} + \mu^2\mathbf{u}^T\Omega^{-1}\mathbf{u}) - \frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) = exp\left(\frac{-1}{2\sigma_m^2}(\mu-\mu_m)^2\right)$$

Matching coefficients of μ^2, we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-1}{2}\mu^2\mathbf{u}^T\Omega^{-1}\mathbf{u} + \frac{-\mu^2}{2\sigma_0^2} \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \mathbf{u}^T\Omega^{-1}\mathbf{u}$$

Matching coefficients of μ, we get

$$\frac{2\mu\mu_m}{2\sigma_m^2} = \mu\left(\mathbf{x}^T\Omega^{-1}\mathbf{u} + \frac{2\mu_0}{2\sigma_0^2}\right) \Rightarrow \mu_m = \sigma_m^2\left(\mathbf{x}^T\Omega^{-1}\mathbf{u} + \frac{\mu_0}{\sigma_0^2}\right) \Rightarrow \frac{1}{1+\sigma_0^2\mathbf{u}^T\Omega^{-1}\mathbf{u}}\left(\sigma_0^2\mathbf{x}^T\Omega^{-1}\mathbf{u} + \mu_0\right)$$

μ_m will be the MAP estimate of μ.

HOMEWORK: What about the special cases of Ω being diagonal matrices with the same or different values along the diagonal?

Problem 3. We discussed atleast two settings where maximizing a monotonically increasing function of the objective is somewhat more intuitive than maximizing the original objective. Recall the two settings. Now prove that maximizing the monotonically increasing transformation of the objective gives the same optimality point as does maximizing the original objective.

Answer: We will prove by contradiction. Let $O(\theta)$ be the objective function being maximized. Let $\theta^* = \underset{\theta}{\text{argmax}}\ O(\theta)$. Let $f(\beta)$ be a monotonically increasing function. Let $\hat{\theta} = \underset{\theta}{\text{argmax}}\ f(O(\theta))$ such that $\hat{\theta} \neq \theta^*$ and $f(O(\hat{\theta})) > f(O(\theta^*))$. Since f is a monotonically increasing function of its arguments, it must be that $O(\hat{\theta}) > O(\theta^*)$. Which is a contradiction, since we had $\theta^* = \underset{\theta}{\text{argmax}}\ O(\theta)$. Thus either, it must be that $\hat{\theta} = \theta^*$ OR $f(O(\hat{\theta})) = f(O(\theta^*))$.