Variational Bayesian Inference

Chris Mattioli

September 6, 2018

Posterior Predictive Distribution

$$p(\boldsymbol{x_{N+1}}|\boldsymbol{X}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x_{N+1}}|\boldsymbol{X}, \theta)p(\theta|\boldsymbol{X})d\boldsymbol{\theta}$$

Distribution over my *next* observation given my previous observations.

Why not just use the MAP?

$$p(\boldsymbol{x_{N+1}}|\boldsymbol{X}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x_{N+1}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}$$

vs.

$$p(\boldsymbol{x_{N+1}}|\boldsymbol{X}) = p(\boldsymbol{x_{N+1}}|\hat{\boldsymbol{\theta}}_{\boldsymbol{MAP}})$$

Recall $\hat{\boldsymbol{\theta}}_{\boldsymbol{MAP}} = \operatorname{argmax}_{\boldsymbol{\theta}}\, p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

Example1: 1D Gaussian with Prior over Mean, MAP

- Let $\boldsymbol{x}$ represent N i.i.d samples from a Gaussian so that $x_i \sim \mathcal{N}(\mu, \sigma^2)$
- Assume $\mu \sim \mathcal{N}(m, s^2)$ (the conjugate prior).

This implies that the posterior is Gaussian as follows:

$$p(\mu|\boldsymbol{x}) = \mathcal{N}\left(\hat{\mu}_{MAP}, \frac{\sigma^2 s^2}{\sigma^2 + Ns^2}\right)$$

where

$$\hat{\mu}_{MAP} = \frac{m\sigma^2 + s^2 \sum_i x_i}{\sigma^2 + Ns^2}$$

Example1: 1D Gaussian with Prior over Mean, PPD

$$
\begin{aligned}
p(x_{N+1}|\boldsymbol{x}) &= \int_{\mu} p(x_{N+1}|\mu)p(\mu|\boldsymbol{x})d\mu \\
&= \int_{\mu} \mathcal{N}(\mu, \sigma^2)\mathcal{N}\left(\hat{\mu}_{MAP}, \frac{\sigma^2 s^2}{\sigma^2 + Ns^2}\right) d\mu \\
&= \mathcal{N}\left(\hat{\mu}_{MAP}, \sigma^2 + \frac{\sigma^2 s^2}{\sigma^2 + Ns^2}\right)
\end{aligned}
$$

Why not just use the MAP?

$$p(x_{N+1}|\boldsymbol{x}) = \mathcal{N}\left(\hat{\mu}_{MAP},\ \sigma^2 + \frac{\sigma^2 s^2}{\sigma^2 + Ns^2}\right)$$

vs.

$$p(x_{N+1}|\boldsymbol{x}) = \mathcal{N}(\hat{\mu}_{MAP},\ \sigma^2)$$

Posterior Predictive Distribution

$$p(\boldsymbol{x_{N+1}}|\boldsymbol{X}) = \int_{\boldsymbol{\theta}} p(\boldsymbol{x_{N+1}}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})d\boldsymbol{\theta}$$

- The PPD incorporates our uncertainty about $\boldsymbol{\theta}$
- Using a point estimate such the MAP could result in a distribution that is too narrow

Using the PPD for inference in practice

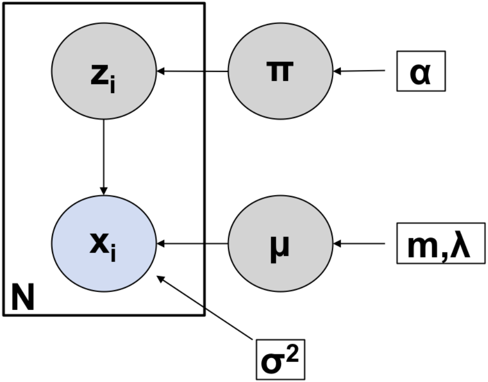
The crux the PPD is the posterior distribution, $p(\theta|\boldsymbol{X})$. In order to use the PPD, we require that the posterior be

- Computationally Tractable
- AnalyticallyTractable

This is not always the case for more complicated problems.

Example2: Bayesian Mixture of 1D Gaussians

We will refer to the (Bayesian) mixture of one dimensional, constant variance Gaussians model.



We want to make inferences under this model using the PPD, so for that we'll need to find $p(\boldsymbol{Z}, \pi, \mu|\boldsymbol{X})$

Example2: Model Set Up

- $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(\mu_k, \sigma^2)^{z_{ik}}$. $\boldsymbol{X}$ is an N length vector
- $p(\boldsymbol{Z}|\boldsymbol{\pi}) = \prod_{i=1}^{N} p(\boldsymbol{z_i}|\boldsymbol{\pi}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{ik}}$. $\boldsymbol{Z}$ is an $N \times K$ matrix where the ith row is $\boldsymbol{z_i}$. $\boldsymbol{z_i}$ is a Categorical RV represented by a "one-hot" encoding, e.g., $\boldsymbol{z_i} = [0, 1, 0]$
- $\boldsymbol{\mu} \sim \prod_{k=1}^{K} \mathcal{N}(m_k, \lambda_k)$, a K length vector representing the mean of the kth Gaussian. Think of these as the cluster centers.
- $\boldsymbol{\pi} \sim Dir(\boldsymbol{\alpha})$, a K length vector describing the probability of being under Gaussian k.

Example2: Bayesian Mixture of 1D Gaussians

In order to makes inferences using the PPD we need to find the posterior distribution.

$$\begin{aligned} p(\boldsymbol{Z}, \pi, \boldsymbol{\mu} | \boldsymbol{X}) &= \frac{1}{p(\boldsymbol{X})} p(\boldsymbol{X}, \boldsymbol{Z}, \pi, \boldsymbol{\mu}) \\ &= \frac{1}{p(\boldsymbol{X})} p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}) p(\boldsymbol{Z} | \pi) p(\pi) p(\boldsymbol{\mu}) \\ &= \frac{p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}) p(\boldsymbol{Z} | \pi) p(\pi) p(\boldsymbol{\mu})}{\sum_{\boldsymbol{z} \in Supp(\boldsymbol{z})} \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\mu}} p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}) p(\boldsymbol{Z} | \pi) p(\pi) p(\boldsymbol{\mu}) d\boldsymbol{\mu} d\boldsymbol{\pi}} \end{aligned}$$

Example2: A Closer Look at the Numerator

Can we infer the form of the posterior from the numerator?

$$\begin{aligned} p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu} | \boldsymbol{X}) &\propto p(\boldsymbol{X} | \boldsymbol{Z}, \boldsymbol{\mu}) p(\boldsymbol{Z} | \boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}) \\ &= Dir(\hat{\boldsymbol{a}}) \prod_{k=1}^{K} \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2) \end{aligned}$$

where

$$\hat{a}_k = \alpha_k + \sum_{i=1}^{N} z_{ik}, \quad \hat{\mu}_k = \frac{m_k \sigma^2 + \lambda_k \sum_i x_{ik}}{\lambda_k \sum_i z_{ik} + \sigma^2}, \quad \hat{\sigma}_k^2 = \frac{\sigma^2 \lambda^2}{\lambda_k \sum_i z_{ik} + \sigma^2}$$

Note that the parameters of the posterior are functions of observed $\boldsymbol{X}$ and *unobserved* Z.

Example2: Closer Look at the Denominator

$$p(\boldsymbol{X}) = \sum_{\boldsymbol{z} \in Supp(\boldsymbol{z})} \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\mu}} p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\mu}) d\boldsymbol{\mu} d\boldsymbol{\pi}$$

$$= \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\mu}} p(\boldsymbol{\pi}) p(\boldsymbol{\mu}) \prod_{i=1}^{N} \sum_{\boldsymbol{z} \in Supp(\boldsymbol{z})} \prod_{k=1}^{K} \left[\pi_k^{z_{ik}} \mathcal{N}(\mu_k, \sigma^2)^{z_{ik}} \right] d\boldsymbol{\mu} d\boldsymbol{\pi}$$

In order to fully compute the marginal, $p(\boldsymbol{X})$, we're required to consider K^N possible assignments of the data. This is absolutely ridiculous!

Example2: PPD of Bayesian Mixture of 1D Gaussians

$$p(\boldsymbol{Z}, \pi, \boldsymbol{\mu} | \boldsymbol{X}) = ?$$

$$p(x_{\boldsymbol{N+1}} | \boldsymbol{X}) = \sum_{\boldsymbol{z} \in Supp(\boldsymbol{z})} \int_{\boldsymbol{\pi}} \int_{\boldsymbol{\mu}} p(x_{\boldsymbol{N+1}} | \boldsymbol{Z}, \pi, \boldsymbol{\mu}) p(\boldsymbol{Z}, \pi, \boldsymbol{\mu} | \boldsymbol{X}) d\boldsymbol{\mu} d\boldsymbol{\pi} = ?$$

Quick Review

- We want to perform inference using the PPD. This encapsulates more information about the randomness of our parameters/latent variables
- In order to use the PPD, we need the posterior distribution of our parameters/latent variables given our observations
- The posterior needs to "look" nice: needs to be computable and/or needs to be of a common distribution (or product of distributions)

Can we approximate the posterior?

$$q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta}|\boldsymbol{X})$$

Pros

- If we can get a good approximate posterior, then we can use it to perform “full” Bayesian inference

Cons

- Approximations aren’t exact

KL Divergence: Similarity Between Two Distributions

The Kullback-Leibler Divergence is a “measure” of similarity between two distributions:

$$KL[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{X})] = \int_{\boldsymbol{\theta}} q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\boldsymbol{X})} d\boldsymbol{\theta}$$

It has the important properties:

$$KL[q||p] \geq 0, \quad \forall q, p$$

$$KL[p||p] = 0, \quad \forall p$$

$$KL(tq_1 + (1-t)q_2||tp_1 + (1-t)p_2) \leq tKL(q_1||p_1) + (1-t)KL(q_2||p_2)$$

for $0 \leq t \leq 1$ (KL Divergence is convex).

Variational Bayesian Inference: KL Divergence Rewritten

An equivalent definition of KL divergence is as follows:

$$\log p(\boldsymbol{X}) = \mathcal{L}(q) + KL[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{X})]$$

where

$$\mathcal{L}(q) = -KL[q(\boldsymbol{\theta})||p(\boldsymbol{X}, \boldsymbol{\theta})]$$

is called the variational lower bound

Variational Bayesian Inference: Maximize $\mathcal{L}(q)$

$$\log p(\boldsymbol{X}) = \mathcal{L}(q) + KL[q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{X})]$$
$$\mathcal{L}(q) = -KL[q(\boldsymbol{\theta})||p(\boldsymbol{X}, \boldsymbol{\theta})]$$

Note that:

$$\mathcal{L}(q) = \log p(\boldsymbol{X}) \quad \text{when } q = p$$
$$\mathcal{L}(q) < \log p(\boldsymbol{X}) \quad \text{when } q \neq p$$

So by maximizing $\mathcal{L}(q)$ over the possible choices of q, we effectively minimize the divergence between q and p.

Variational Bayesian Inference: Constraints

If our objective is the following:

$$q^*(\boldsymbol{\theta}) = \arg\max_q \mathcal{L}(q)$$

It may be prudent to constrain in some way the search for q. Specifically, we want q to be computationally tractable but also as flexible as possible.

Variational Bayes: Mean Field Approximation

Let $\boldsymbol{\theta} = \{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_n}\}$. Approximate $p(\boldsymbol{\theta}|\boldsymbol{X})$ with the following:

$$q(\boldsymbol{\theta}) = \prod_i q(\boldsymbol{\theta^{(i)}}), \quad \boldsymbol{\theta^{(i)}} \subseteq \boldsymbol{\theta},\ \boldsymbol{\theta^{(i)}} \cap \boldsymbol{\theta^{(j)}} = \emptyset\ \forall i \neq j.$$

This is the same as saying q factorizes over $\boldsymbol{\theta}$

Example: $p(\mu, \sigma^2|x) \approx q(\mu, \sigma^2) = q(\mu)q(\sigma^2)$

Mean Field Optimal Solution

Keeping $i \neq j$ fixed, the optimal solution is given by the following:

$$q^*(\boldsymbol{\theta}^{(j)}) = \arg\max_q \mathcal{L}(q)$$

where

$$\log q^*(\boldsymbol{\theta}^{(j)}) = \mathbb{E}_{(\boldsymbol{\theta} \setminus \boldsymbol{\theta}^{(j)}) \sim q(\boldsymbol{\theta} \setminus \boldsymbol{\theta}^{(j)})}[\log p(\boldsymbol{X}, \boldsymbol{\theta})] + const$$

Note the expectation is taken with respect to all the parameters *accept* the parameters in set $\boldsymbol{\theta}^{(j)}$ drawn from their *approximate* distribution.

Example: $\log q^*(\mu) = \mathbb{E}_{\sigma^2 \sim q(\sigma^2)}[\log p(x, \mu, \sigma^2)] + const$

Example3: Variational Bayesian Mixture of 1D Gaussians

$$p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu} | \boldsymbol{X}) \approx q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}) = q(\boldsymbol{Z}) q(\boldsymbol{\pi}) q(\boldsymbol{\mu})$$

Example3: VBGM, $q^*(\boldsymbol{Z})$

$$\begin{aligned}
\log q^*(\boldsymbol{Z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu}}[\log p(\boldsymbol{X},\boldsymbol{Z},\pi,\boldsymbol{\mu})] + c \\
&= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\mu}}[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu}) + \log p(\boldsymbol{Z}|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu})] + c \\
&= \mathbb{E}_{\boldsymbol{\mu}}[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu})] + \mathbb{E}_{\boldsymbol{\pi}}[\log p(\boldsymbol{Z}|\boldsymbol{\pi})] + c \\
&= \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\log\rho_{ik} + c \\
&\Rightarrow
\end{aligned}$$

$$q^*(\boldsymbol{Z}) = \prod_{i=1}^{N}\prod_{k=1}^{K} r_{ik}^{z_{ik}}$$

where

$$\log\rho_{ik} = \left[\mathbb{E}_{\boldsymbol{\pi}}[\log\pi_k] - \frac{1}{2\sigma^2}\mathbb{E}_{\boldsymbol{\mu}}[(x_i-\mu_k)^2] - \frac{\log 2\pi\sigma^2}{2}\right], \quad r_{ik} = \frac{\rho_{ik}}{\sum_{j=1}^{K}\rho_{ij}}$$

Example3: VBGM, $q^*(\boldsymbol{\pi})$

$$\begin{aligned}\log q^*(\boldsymbol{\pi}) &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{\mu}}[\log p(\boldsymbol{X},\boldsymbol{Z},\pi,\boldsymbol{\mu})] + c \\ &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{\mu}}[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu}) + \log p(\boldsymbol{Z}|\pi) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu})] + c \\ &= \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{Z}|\pi)] + \log p(\boldsymbol{\pi}) + c \\ &\Rightarrow \\ & q^*(\boldsymbol{\pi}) = Dir\left(\hat{\boldsymbol{\alpha}}\right)\end{aligned}$$

$$\hat{\alpha_k} = \alpha_k + \sum_{i=1}^{N} r_{ik}$$

Example3: VBGM, $q^*(\boldsymbol{\mu})$

$$\begin{aligned}\log q^*(\boldsymbol{\mu}) &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{\pi}}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu})] + c \\ &= \mathbb{E}_{\boldsymbol{Z},\boldsymbol{\pi}}[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}) + \log p(\boldsymbol{Z}|\boldsymbol{\pi}) + \log p(\boldsymbol{\pi}) + \log p(\boldsymbol{\mu})] + c \\ &= \mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu})] + \log p(\boldsymbol{\mu}) + c \\ &\Rightarrow\end{aligned}$$

$$q^*(\boldsymbol{\mu}) = \prod_{k=1}^{K} \mathcal{N}\left(\hat{m}_k, \hat{\lambda}_k\right)$$

$$\hat{m}_k = \frac{\sigma^2 m_k + \lambda_k \sum_{i=1}^N x_i r_{ik}}{\sigma^2 + \lambda_k \sum_{i=1}^N r_{ik}}, \quad \hat{\lambda}_k = \frac{\lambda_k \sigma^2}{\sigma^2 + \lambda_k \sum_{i=1}^N r_{ik}}$$

Example3: VBGM, the PPD

$$p(x_{N+1}|\boldsymbol{X}) \approx \frac{1}{\sum_j \hat{\alpha_j}} \sum_{k=1}^{K} \hat{\alpha_k} \mathcal{N}\left(\hat{m}_k, \sigma^2 + \hat{\lambda_k}\right)$$

$$\hat{\alpha_k} = \alpha_k + \sum_{i=1}^{N} r_{ik}, \quad \hat{m}_k = \frac{\sigma^2 m_k + \lambda_k \sum_{i=1}^{N} x_i r_{ik}}{\sigma^2 + \lambda_k \sum_{i=1}^{N} r_{ik}}, \quad \hat{\lambda_k} = \frac{\lambda_k \sigma^2}{\sigma^2 + \lambda_k \sum_{i=1}^{N} r_{ik}}$$

Example3: The True Posterior Predictive Distribution

$$p(x_{N+1}|\boldsymbol{X}) = \frac{1}{\sum_j \hat{a}_j} \sum_{k=1}^{K} \hat{a}_k \mathcal{N}\left(\hat{\mu_k}, \sigma^2 + \hat{\sigma}_k^2\right)$$

$$\hat{a}_k = \alpha_k + \sum_{i=1}^{N} z_{ik}, \quad \hat{\mu}_k = \frac{\sigma^2 m_k + \lambda_k \sum_i x_{ik}}{\sigma^2 + \lambda_k \sum_i z_{ik}}, \quad \hat{\sigma}_k^2 = \frac{\lambda_k \sigma^2}{\sigma^2 + \lambda_k \sum_i z_{ik}}$$

Posterior Comparison

$$\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu} | \boldsymbol{X} \sim Dir(\hat{\boldsymbol{a}}) \prod_{k=1}^{K} \mathcal{N}(\hat{\mu}_k, \hat{\sigma}_k^2)$$

$$\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\mu} \sim \prod_{i=1}^{N} \prod_{k=1}^{K} r_{ik}^{z_{ik}} Dir\left(\hat{\alpha}\right) \prod_{k=1}^{K} \mathcal{N}\left(\hat{m_k}, \hat{\lambda_k}\right)$$

chris.mattioli@ll.mit.edu