Mixture of Gaussians CAVI Update Proof

Luiz do Valle

July 2020

1 Bayesian Mixture of Gaussians Model

Consider a Bayesian mixture of unit-variance univariate Gaussians. There are K mixture components, corresponding to K Gaussian distributions with means $\boldsymbol{\mu} = \{\mu_1, \cdots, \mu_K\}$. The mean parameters are drawn independently from a common prior $p(\mu_k)$, which is assumed to be a Gaussian distribution $\mathcal{N}(0, \sigma^2)$, where σ^2 is a hyperparemeter. To generate an observation x_i from the model, we first choose a cluster assignment c_i, which indicates which latent cluster x_i comes from and is drawn from a categorical distribution over $\{1, \cdots, K\}$ (c_i is one-hot encoded, meaning it is a K dimensional vector with all zeros except for a 1 at the position corresponding to x_i's cluster). We draw x_i from the Gaussian $\mathcal{N}(c_i^T \boldsymbol{\mu}, 1)$

$$\begin{aligned} \mu_k &\sim \mathcal{N}(0, \sigma^2) && k = 1, \cdots, K \\ c_i &\sim \text{Categorical}\left(\tfrac{1}{K}, \cdots, \tfrac{1}{K}\right) && i = 1, \cdots, N \\ x_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}\left(c_i^T \boldsymbol{\mu}, 1\right) && i = 1, \cdots, N \end{aligned}$$

2 The Problem of Approximate Inference

Let $\boldsymbol{x} = x_{1:n}$ be the set of observed variables and $\boldsymbol{z} = z_{1:m}$ be the set of latent variables ($\boldsymbol{c}$ and $\boldsymbol{\mu}$ in the mixture of Gaussians case). The inference problem is to compute the conditional density of the latent variables given the observations, $p(\boldsymbol{z} \mid \boldsymbol{x})$. Using Bayes' rule, the conditional density can be written as follows:

$$p(\boldsymbol{z} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{z}, \boldsymbol{x})}{p(\boldsymbol{x})} \tag{1}$$

The denominator, called the evidence, is often intractable to calculate directly, so we would like to approximate it.

3 The Evidence Lower Bound

In variational inference, we specify a family Q of densities over the latent variables. Each $q(\boldsymbol{z}) \in Q$ is a candidate approximation to the exact conditional.

The goal is to find the one closest in KL divergence to the exact conditional. Inference is now reduced to the following optimization problem,

$$q^*(\boldsymbol{z}) = \arg\min_{q(\boldsymbol{z})\in Q} \mathrm{KL}\Big(q(\boldsymbol{z}) \;||\; p(\boldsymbol{z} \mid \boldsymbol{x})\Big) \tag{2}$$

The KL divergence has the following form,

$$\mathrm{KL}\Big(q(\boldsymbol{z}) \;||\; p(\boldsymbol{z} \mid \boldsymbol{x})\Big) = \mathbb{E}\Big[\log q(\boldsymbol{z})\Big] - \mathbb{E}\Big[\log p(\boldsymbol{z} \mid \boldsymbol{x})\Big] \tag{3}$$

where all expectations are taken with respect to $q(\boldsymbol{z})$. If we expand the conditional, we see that the KL divergence depends on $\log p(\boldsymbol{x})$,

$$\mathrm{KL}\Big(q(\boldsymbol{z}) \;||\; p(\boldsymbol{z} \mid \boldsymbol{x})\Big) = \mathbb{E}\Big[\log q(\boldsymbol{z})\Big] - \mathbb{E}\Big[\log p(\boldsymbol{z}, \boldsymbol{x})\Big] + \log p(\boldsymbol{x}) \tag{4}$$

Since $\log p(\boldsymbol{x})$ is a constant with respect to what we are trying to optimize, we can safely ignore it. The Evidence Lower Bound (ELBO) objective functions is, therefore, the negative of the KL divergence plus $\log p(\boldsymbol{x})$.

$$\mathrm{ELBO}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}\Big[\log p(\boldsymbol{z}, \boldsymbol{x})\Big] - \mathbb{E}\Big[\log q(\boldsymbol{z})\Big] \tag{5}$$

Maximizing the ELBO is equivalent to minimizing the KL divergence.

4 The Mean-Field Approximation

In this derivation, we focus on the mean-field variational family, where the latent variables are mutually independent and each governed by a distinct factor in the variational density. A generic member of the mean-field variational density is

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j) \tag{6}$$

In the case of the Bayesian mixture of Gaussians, we have,

$$q(\boldsymbol{z}) = q(\boldsymbol{\mu}, \boldsymbol{c}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i) \tag{7}$$

5 Evidence Lower Bound (ELBO) for Mixture of Gaussians

The general form for the ELBO can be written as shown in Equation (8); all expectations from now on are taken with respect to $q(\boldsymbol{z})$.

$$\mathrm{ELBO}(\boldsymbol{x}, \boldsymbol{z}) = \mathbb{E}\Big[\log p(\boldsymbol{z})\Big] + \mathbb{E}\Big[\log p(\boldsymbol{x} \mid \boldsymbol{z})\Big] - \mathbb{E}\Big[\log q(\boldsymbol{z})\Big] \tag{8}$$

By combining the joint and mean-field family, we obtain the ELBO for the mixture of Gaussians,

$$\begin{aligned}\mathrm{ELBO}(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\varphi}) = &\sum_{k=1}^{K} \mathbb{E}\Big[\log p(\mu_k); m_k, s_k^2\Big] \\ &+ \sum_{i=1}^{n} \left(\mathbb{E}\Big[\log p(c_i); \varphi_i\Big] + \mathbb{E}\Big[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \boldsymbol{m}, \boldsymbol{s}^2\Big]\right) \\ &- \sum_{i=1}^{n} \mathbb{E}\Big[\log q(c_i; \varphi_i)\Big] - \sum_{k=1}^{K} \mathbb{E}\Big[\log q(\mu_k; m_k, s_k^2)\Big]\end{aligned} \tag{9}$$

We will now derive each of the expectations in (9). All the values that do not depend on the variables we are optimizing with respect to (i.e. m_k, s_k, and φ_i) are absorbed in a constant term. First, the log-prior of μ_k is obtained as shown in (10).

$$\begin{aligned}\log p(\mu_k) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu_k^2}{2\sigma^2}\right)\right) \\ &= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu_k^2}{2\sigma^2} \\ &= -\frac{\mu_k^2}{2\sigma^2} + \text{constant}\end{aligned} \tag{10}$$

The only random variable in (10) is μ_k itself, so the expectation of (10) is as shown in (11).

$$\begin{aligned}\mathbb{E}\Big[\log p(\mu_k)\Big] &= \mathbb{E}\left[-\frac{\mu_k^2}{2\sigma^2}\right] + \text{constant} \\ &= -\frac{s_k^2 + m_k^2}{2\sigma^2} + \text{constant}\end{aligned} \tag{11}$$

The log-prior of c_i is simply a constant, hence its expectation is also a constant.

$$\begin{aligned}\log p(c_i) &= \log\left(\prod_{k=1}^{K}\left(\frac{1}{K}\right)^{c_{ik}}\right) \\ &= \log\left(\frac{1}{K}\right) \\ &= -\log(K) \\ &= \text{constant}\end{aligned} \tag{12}$$

$$\mathbb{E}\Big[\log p(c_i)\Big] = \text{constant} \tag{13}$$

We obtain the log-likelihood of the data as shown in (14), where the dot-product between c_i^T and $\boldsymbol{\mu}$ is represented in its summation form.

$$\begin{aligned}\log p(x_i \mid c_i, \boldsymbol{\mu}) &= \log\left(\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_i - c_i^T\boldsymbol{\mu})^2}{2}\right)\right) \\ &= -\frac{(x_i - \sum_{k=1}^{K} c_{ik}\mu_k)^2}{2} + \text{constant}\end{aligned} \tag{14}$$

Using the mean-field approximation, the expectation of (14) is obtained as shown below. Here, the only two random variables are c_{ik} and μ_k. μ_k and c_i are independent by the mean-field approximation, so the expectation of their product is the product of their expectations.

$$\begin{aligned}\mathbb{E}\Big[\log p(x_i \mid c_i, \boldsymbol{\mu})\Big] &= \mathbb{E}\left[-\frac{(x_i - \sum_{k=1}^{K} c_{ik}\mu_k)^2}{2}\right] + \text{constant} \\ &= -\frac{1}{2}\mathbb{E}\left[x_i^2 - 2x_i\sum_{k=1}^{K} c_{ik}\mu_k + \left(\sum_{k=1}^{K} c_{ik}\mu_k\right)^2\right] + \text{constant} \\ &= -\frac{1}{2}\left(x_i^2 - 2x_i\sum_{k=1}^{K}\mathbb{E}[c_{ik}\mu_k] + \mathbb{E}\left[\left(\sum_{k=1}^{K} c_{ik}\mu_k\right)^2\right]\right) + \text{constant} \\ &= -\frac{1}{2}\left(x_i^2 - 2x_i\sum_{k=1}^{K}\varphi_{ik}m_k + \sum_{k=1}^{K}\varphi_{ik}(s_k^2 + m_k^2)\right) + \text{constant} \\ &= -\frac{1}{2}\left(-2x_i\sum_{k=1}^{K}\varphi_{ik}m_k + \sum_{k=1}^{K}\varphi_{ik}(s_k^2 + m_k^2)\right) + \text{constant}\end{aligned} \tag{15}$$

Similar to its prior, the log-posterior of c_i is simply the log of the posterior-probability of x_i being in that cluster. φ_i is a constant with respect to $q(\boldsymbol{z})$, so the expectation of $\log(\varphi_i)$ is simply $\log(\varphi_i)$ itself.

$$\log q(c_i; \varphi_i) = \sum_{k=1}^{K} c_{ik}\log(\varphi_{ik}) \tag{16}$$

$$\mathbb{E}[\log q(c_i; \varphi_i)] = \sum_{k=1}^{K}\varphi_{ik}\log(\varphi_{ik}) \tag{17}$$

The log of the posterior of μ_k is as shown in (18). Again, the only random variable here is μ_k itself.

$$\begin{aligned}\log q(\mu_k; m_k, s_k^2) &= \log\left(\frac{1}{\sqrt{2\pi s_k^2}}\exp\left(-\frac{(\mu_k - m_k)^2}{2s_k^2}\right)\right) \\ &= -\frac{1}{2}\log(s_k^2) - \frac{(\mu_k - m_k)^2}{2s_k^2} + \text{constant}\end{aligned} \tag{18}$$

$$
\begin{aligned}
\mathbb{E}\Big[\log q(\mu_k; m_k, s_k^2)\Big] &= -\frac{1}{2}\left(\log(s_k^2) + \frac{1}{s_k^2}\Big(\mathbb{E}[\mu_k^2] - 2m_k\mathbb{E}[\mu_k] + m_k^2\Big)\right) + \text{constant} \\
&= -\frac{1}{2}\left(\log(s_k^2) + \frac{1}{s_k^2}\Big((s_k^2 + m_k^2) - 2m_k^2 + m_k^2\Big)\right) + \text{constant} \\
&= -\frac{1}{2}\left(\log(s_k^2) + \frac{1}{s_k^2}s_k^2\right) + \text{constant} \\
&= -\frac{1}{2}\log(s_k^2) + \text{constant}
\end{aligned} \tag{19}
$$

Plugging these expectations back into (9), we obtain (20).

$$
\begin{aligned}
\text{ELBO}(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\varphi}) = &-\frac{1}{2\sigma^2}\sum_{k=1}^{K}(s_k^2 + m_k^2) \\
&+ \frac{1}{2}\sum_{i=1}^{n}\left(2x_i\sum_{k=1}^{K}\varphi_{ik}m_k - \sum_{k=1}^{K}\varphi_{ik}(s_k^2 + m_k^2)\right) \\
&- \sum_{i=1}^{n}\sum_{k=1}^{K}\varphi_{ik}\log(\varphi_{ik}) + \frac{1}{2}\sum_{k=1}^{K}\log(s_k^2) + \text{constant}
\end{aligned} \tag{20}
$$

6 m_j Update

The update for m_j is found by taking the derivative of Equation (20) with respect to m_j and solving for m_j when the derivative is equal to zero. The result is shown in (22).

$$
\frac{\partial}{\partial m_j}\text{ELBO}(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\varphi}) = -\frac{m_j}{\sigma^2} + \sum_{i=1}^{n}(x_i\varphi_{ij} - \varphi_{ij}m_j) \tag{21}
$$

$$
m_j = \frac{\sum_{i=1}^{n}\varphi_{ij}x_i}{1/\sigma^2 + \sum_{i=1}^{n}\varphi_{ij}} \tag{22}
$$

7 s_j^2 Update

The update for s_j^2 is found by taking the derivative of Equation (20) with respect to s_j^2 and solving for s_j^2 when the derivative is equal to zero. The result is shown in (24).

$$
\frac{\partial}{\partial s_j^2}\text{ELBO}(\boldsymbol{m}, \boldsymbol{s}^2, \boldsymbol{\varphi}) = -\frac{1}{2\sigma^2} - \frac{1}{2}\sum_{i=1}^{n}\varphi_{ij} + \frac{1}{2s_j^2} \tag{23}
$$

$$
s_j^2 = \frac{1}{1/\sigma^2 + \sum_{i=1}^{n}\varphi_{ij}} \tag{24}
$$

8 φ_{ik} Update

Optimizing φ_{ik} is not as simple as solving for φ_{ik} when $\frac{\partial}{\partial \varphi_{ik}}\text{ELBO} = 0$, the way we did above for m_j and s_j^2. That is because we have an additional constraint,

$$\sum_{k=1}^{K} \varphi_{ik} = 1 \tag{25}$$

Therefore, in order to find the value for φ_{ik} that maximizes the ELBO while also respecting (25), we will use the Lagrange multiplier technique. Accordingly, the ELBO function will be maximized with respect to the constrained φ_{ij} when the gradient of the ELBO and the gradient of $\sum_{k=1}^{K} \varphi_{ik}$ are scalar multiples of each other, where the scalar is the Lagrange multiplier. This situation is shown below, where ∇_{φ_i} represents the gradient with respect to φ_i only (treating m_k ad s_k^2 as constants) and λ is the Lagrange multiplier.

$$\nabla_{\varphi_i}\text{ELBO}(\boldsymbol{m}, \boldsymbol{s^2}, \boldsymbol{\varphi}) = \lambda \nabla \sum_{k=1}^{K} \varphi_{ik} \tag{26}$$

This gives us a system of $K+1$ equations, which is good since we have $K+1$ unknowns: K φ_{ik} variables and λ. The partial derivative of the ELBO with respect to a single φ_{ij} is shown below along with the partial derivative with respect to φ_{ij} of $\sum_{k=1}^{K} \varphi_{ik}$.

$$\frac{\partial}{\partial \varphi_{ij}}\text{ELBO}(\boldsymbol{m}, \boldsymbol{s^2}, \boldsymbol{\varphi}) = x_i m_j - \frac{s_j^2 + m_j^2}{2} - \log(\varphi_{ij}) - 1 \tag{27}$$

$$\frac{\partial}{\partial \varphi_{ij}} \sum_{k=1}^{K} \varphi_{ik} = 1 \tag{28}$$

The system of equations generated from (26), (27), and (28) is as shown below,

$$\begin{gathered} x_i m_1 - \frac{s_1^2 + m_1^2}{2} - \log(\varphi_{i1}) - 1 = \lambda \\ \vdots \\ x_i m_K - \frac{s_K^2 + m_K^2}{2} - \log(\varphi_{iK}) - 1 = \lambda \\ \sum_{k=1}^{K} \varphi_{ik} = 1 \end{gathered}$$

Solving for any of the φ_{ik} from the first K equations, we get

$$\varphi_{ij} = \exp\left(x_i m_j - \frac{s_j^2 + m_j^2}{2} - \lambda - 1 \right) \tag{29}$$

Plugging (29) into (25), we obtain an expression for λ in terms of x_i, m_j, and s_j^2.

$$\lambda = \log\left(\sum_{k=1}^{K} \exp\left(x_i m_j - \frac{s_j^2 + m_j^2}{2}\right)\right) - 1 \tag{30}$$

By plugging (30) back into (29), we obtain the update for φ_{ij}

$$\varphi_{ij} = \frac{\exp\left(x_i m_j - \frac{s_j^2+m_j^2}{2}\right)}{\sum_{k=1}^{K} \exp\left(x_i m_k - \frac{s_k^2+m_k^2}{2}\right)} \tag{31}$$