Chapter 10 Variational Inference

1. **Goal** Evaluate posterior $P(\mathbf{Z}|\mathbf{X})$ of the latent variables $\mathbf{Z}$ given observed variables $\mathbf{X}$.
2. **Use Case** In EM, need to evaluate expectation of $\log(p(\mathbf{X}, \mathbf{Z}))$ w.r.t. $p(\mathbf{Z}|\mathbf{X})$
3. **Problem** Latent space may have very high dimension or posterior distribution too complex to be analytically tractable.
4. **Approaches**
   (a) **Stochastic** Markov chain Monte Carlo and other sampling methods
   (b) **Deterministic** variational inference, laplace approximation

10.1 Variational Inference

1. **Functional derivative** How value of a functional changes in response to infinitesimal changes to the input function
2. **Bayesian Model** Given joint distribution $p(\mathbf{X}, \mathbf{Z})$ goal is try to find an approximation for the posterior $p(\mathbf{Z}|\mathbf{X})$ and model evidence $p(\mathbf{X})$. We can decompose log marginal probability

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathbf{KL}\left(q||p\right)$$

where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left(\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right) d\mathbf{Z} \qquad \mathbf{KL}\left(q||p\right) = -\int q(\mathbf{Z}) \ln \left(\frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right) d\mathbf{Z}$$

Note $\mathbf{\Theta}$ are stochastic and are absorbed into $\mathbf{Z}$. We can maximize lower bound $\mathcal{L}(q)$ w.r.t. the distribution $q(\mathbf{Z})$, which is equivalent to minimizing KL divergence. The maximum of the lower bound occurs when KL divergence vanishes. when $p(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X})$. Now we choose a restricted family of distributions $q(\mathbf{Z})$ and then seek the member of this family for which the KL divergence is minimized. $q(\mathbf{Z})$ should be

(a) tractable
(b) sufficiently rich and flexible, such that they are capable of good approximations to the true posterior
(c) Can be as flexible as possible without worrying about overfitting. A complex/flexible distribution simply means we can get closer to the true posterior distribution

One such family is a Gaussian, we can optimize $q(\mathbf{Z}|w)$ w.r.t. mean/variance w

3. **Factorized distributions** A restriction of distribution for $q(\mathbf{Z})$. Assumption is we can parittion elements of $\mathbf{Z}$ into disjiont groups $\mathbf{Z}_i$ such that the q distributions factorizes

$$q(\mathbf{Z}) = \prod_{i}^{M} q_i(\mathbf{Z}_i)$$

This corresponds to **mean field theory**.

4. **Variational Optimization** Goal is to find a distribution $q(\mathbf{Z})$ for which the lower bound $\mathcal{L}(q)$ is largest. Idea is to optimize $\mathcal{L}(q)$ with respect to all distribution $q_i(\mathbf{Z}_i)$, which we do by optimizing with respect to each of the factors in turn

Variational Mixture of Gaussians

1. **MOdel** Given observed data $\mathbf{X}$ and latent variables $\mathbf{Z}$, mixing coefficients $\boldsymbol{\pi}$, we can write conditional distribution of the latent variable

$$p(\mathbf{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{z_{nk}}$$

we can write the conditional distribution of the observed varialbe

$$p(\mathbf{X}|\mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mathcal{N}\left(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{nk}}$$

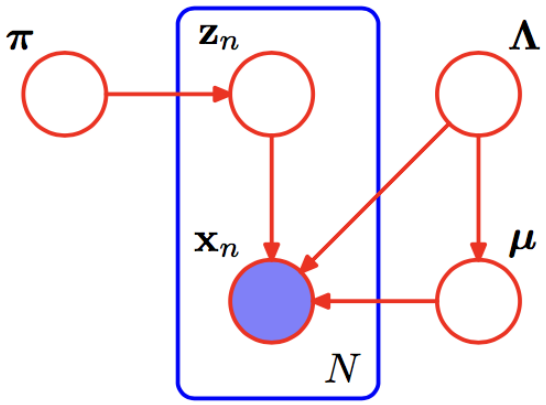
we use dirichlet distribution for the mixing coefficients $\boldsymbol{\pi}$

$$p(\boldsymbol{\pi}) = Dir(\boldsymbol{\pi}|\alpha_0) = C(\alpha_0)\prod_{k=1}^{K} \pi_k^{\alpha_0 - 1}$$

for the same parameter α_0. We use Gaussian-Wishart prior for mean/variance of each Gaussian

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{\mu}|\boldsymbol{\Lambda})p(\boldsymbol{\Lambda})$$

We have the bayesian mixture of Gaussian model



2. **Variational Distribution** Joint distribution given by

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

We consider a variational distribution that factorizes between the latent variable and the parameters

$$q(\mathbf{Z}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\mathbf{Z})q(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

Variational Inference Review

1. **Idea** Posit a family of approximate densities Q and find the member of that family q^*, parameterized by its variational parameters, that minimizes the Kullback-Leibler (KL) divergence to the exact posterior

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z}) \in Q} \mathbf{KL}\left(q(\mathbf{z})||p(\mathbf{z}|\mathbf{x})\right)$$

2.