Mean-field Variational Inference for Conjugate-Exponential Models: A Tutorial

Revisiting Chapter 21: Variational Inference of *Machine Learning: A Probabilistic Perspective*

BY ZHIHAN YANG

Department of Mathematics and Statistics

Carleton College

Abstract

All models encounterd in Murphy begs a question: why?

First show that the model actually belongs to the family

Using this fact, we pick the approximate posterior that allows for tractable VI

Then we derive the coordinate descent VI algorithm

1 Conjugate-exponential models

1.1 Exponential family

2 Variational inference in general

3 Examples

3.1 Univariate Gaussian

3.2 Linear regression

3.3 Linear regression with ARD

3.4 Mixture of Gaussians

Complete-data likelihood:

$$\prod_{i=1}^{n} p(\boldsymbol{x}_i, \boldsymbol{z}_i | \, \boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{k=1}^{c} \pi_k^{z_{ik}} \mathcal{N}(\boldsymbol{x}_i | \, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\mu}_{1:k}, \boldsymbol{\Sigma}_{1:k})$.

Prior for $\boldsymbol{\pi}$:

Dirichlet

$$\mathrm{Dir}(\boldsymbol{\pi} | \, \boldsymbol{\alpha}_0) \propto \prod_{k=1}^{c} \pi_k^{(\alpha_0)_k - 1}$$

Relevant part of the complete data log likelihood

$$\prod_{i=1}^{n}\prod_{k=1}^{c}\pi_k^{z_{ik}}$$

Proof that they are conjugate (this is not actually useful)

$$\left(\prod_{k=1}^{c}\pi_k^{(\alpha_0)_k-1}\right)\left(\prod_{i=1}^{n}\prod_{k=1}^{c}\pi_k^{z_{ik}}\right)=\left(\prod_{k=1}^{c}\pi_k^{(\alpha_0)_k-1}\right)\left(\prod_{k=1}^{c}\pi_k^{\sum_{i=1}^{n}z_{ik}}\right)=\prod_{k=1}^{c}\pi_k^{(\alpha_0)_k+\sum_{i=1}^{n}z_{ik}-1}$$

The joint prior for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$:

Gaussian wishart

$$p(\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k|\,\boldsymbol{m}_0,\beta_0,v_0)=\mathcal{N}(\boldsymbol{\mu}_k|\,\boldsymbol{m}_0,(\beta_0\boldsymbol{\Lambda}_k)^{-1})\,\mathrm{Wi}(\boldsymbol{\Lambda}_k|\,\boldsymbol{L}_0,v_0)$$

Relevant part of the complete data log likelihood

$$\prod_{i=1}^{n}\mathcal{N}(\boldsymbol{x}_i|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)^{z_{ik}}=\prod_{i\text{ s.t. }z_{i,k=1}}\mathcal{N}(\boldsymbol{x}_i|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$$

Proof that they are conjugate to each other

Omitted, see section 4.6.3.3

Instantiate the CAVI algorithm

$$\begin{aligned}
\log q(\boldsymbol{Z}) &= \mathbb{E}_{q(\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\log p(\boldsymbol{X},\boldsymbol{Z},\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})]+\text{const}\\
&= \mathbb{E}_{q(\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\log p(\boldsymbol{Z}|\,\boldsymbol{\pi})+\log p(\boldsymbol{X}|\,\boldsymbol{Z},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})+\log p(\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})]+\text{const}\\
&= \mathbb{E}_{q(\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\log p(\boldsymbol{Z}|\,\boldsymbol{\pi})+\log p(\boldsymbol{X}|\,\boldsymbol{Z},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})]+\text{const}\\
&= \mathbb{E}_{q(\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\log p(\boldsymbol{Z}|\,\boldsymbol{\pi})]+\mathbb{E}_{q(\boldsymbol{\pi},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\log p(\boldsymbol{X}|\,\boldsymbol{Z},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})]+\text{const}\\
&= \mathbb{E}_{q(\boldsymbol{\pi})}[\log p(\boldsymbol{Z}|\,\boldsymbol{\pi})]+\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\log p(\boldsymbol{X}|\,\boldsymbol{Z},\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})]+\text{const}\\
&= \mathbb{E}_{q(\boldsymbol{\pi})}\left[\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\log\pi_k\right]+\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}\left[\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\log\mathcal{N}(\boldsymbol{x}_i|\,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right]+\text{const}\\
&= \sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\,\mathbb{E}_{q(\boldsymbol{\pi})}[\log\pi_k]+\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\,\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}\Big[\frac{1}{2}\ln|\boldsymbol{\Lambda}_k|-\frac{D}{2}\ln 2\pi-\frac{1}{2}(\boldsymbol{x}_i-\\
&\quad \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\Big]+\text{const}\\
&= \sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\,\mathbb{E}_{q(\boldsymbol{\pi})}[\log\,\pi_k]+\sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\left[\frac{1}{2}\,\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\ln|\boldsymbol{\Lambda}_k|]-\frac{D}{2}\ln\,2\pi-\right.\\
&\quad \left.\frac{1}{2}\,\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]\right]+\text{const}\\
&= \sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\Big(\mathbb{E}_{q(\boldsymbol{\pi})}[\log\pi_k]+\frac{1}{2}\,\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[\ln|\boldsymbol{\Lambda}_k|]-\frac{D}{2}\ln 2\pi-\frac{1}{2}\,\mathbb{E}_{q(\{\boldsymbol{\mu}_k\},\{\boldsymbol{\Sigma}_k\})}[(\boldsymbol{x}_i-\\
&\quad \boldsymbol{\mu}_k)^T\boldsymbol{\Lambda}_k(\boldsymbol{x}_i-\boldsymbol{\mu}_k)]\Big)\\
&\triangleq \sum_{i=1}^{n}\sum_{k=1}^{c}z_{ik}\log\rho_{ik}\\
q(\boldsymbol{Z}) &\propto \prod_{i=1}^{n}\prod_{k=1}^{c}\rho_{ik}^{z_{ik}}
\end{aligned}$$

Requring that the distributuon is normalized on each row:

$$q(\boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{c} r_{ik}^{z_{ik}}$$

where $r_{i,k}$ is the rho divided by row sum. But we need values of rho to compute this distribution.

3.5 Logistic regression*

3.6 Logistic regression with ARD*