

Variational Inference

Fall 2019

Instructor: Shandian Zhe

zhe@cs.utah.edu

School of Computing



Outline

- Gaussian Mixture Model and EM algorithm
- Variational Inference
 - Variational evidence lower bound
 - Mean-field variational inference
- Local variational inference
 - Convex conjugate
 - Logistic regression
 - Variational EM
- Variational message passing

Outline

- Gaussian Mixture Model and EM algorithm
- Global variational Inference
 - Mean-field variational inference
- Local variational inference
 - Logistic regression
- Variational message passing

Gaussian mixture model (GMM)

- A probabilistic version of the k-means clustering algorithm
- Given a set of data points and a cluster number K , how do you group the data points into K clusters?
- Clustering is a fundamental data mining and pattern recognition task

K-means application

$K = 2$



$K = 3$



$K = 10$



Original image



How do we use probabilistic modeling to represent the clustering procedure?

- Given $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and cluster number K
- For each data point n
 - Sample the cluster membership $\mathbf{z}_n$: $K \times 1$ one-hot vector,
 $z_{nk}=1$ means $\mathbf{x}_n$ belongs to cluster k

e.g., $[0, 0, 1, \dots, 0]^\top$

$$p(\mathbf{z}_n) = \prod_{k=1}^K \pi_k^{z_{nk}}$$

$$\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^\top$$
$$\mathbf{z}_n = [z_{n1}, \dots, z_{nK}]^\top$$

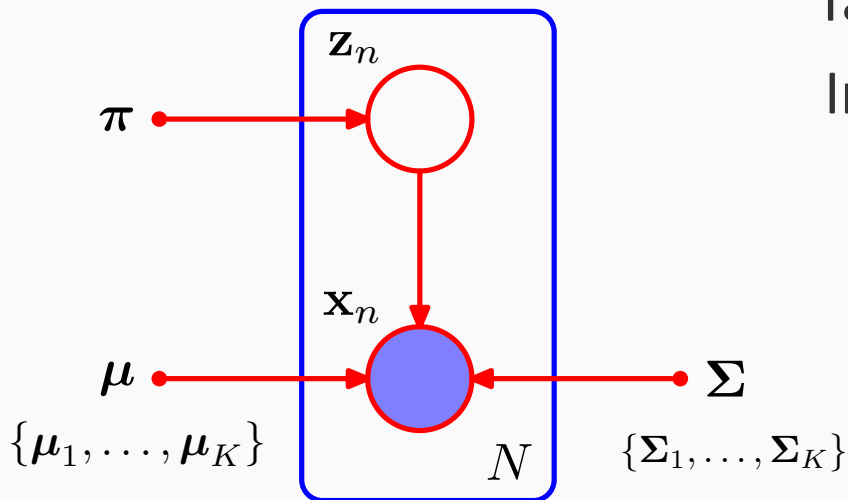
- Given $\mathbf{z}_n$, sample $\mathbf{x}_n$ from the cluster-specific Gaussian

$$p(\mathbf{x}_n | \mathbf{z}_n) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

cluster center

soft width

Graphical model representation



Task: Given $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and K

Infer: π

$\{\mu_1, \dots, \mu_K\}$ } parameters

$\{\Sigma_1, \dots, \Sigma_K\}$

$p(\mathbf{z}_1, \dots, \mathbf{z}_N | \mathbf{X})$ } posterior

How to learn GMMs?

- Marginalize out $\mathbf{z}$ and do MLE

$$p(\mathbf{x}_n) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



$$\ln p(\mathbf{X} | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

$$\begin{aligned} & 0 \leq \pi_k \leq 1 \\ \text{s.t.} \quad & \sum_{k=1}^K \pi_k = 1 \end{aligned}$$

How to learn GMMs?

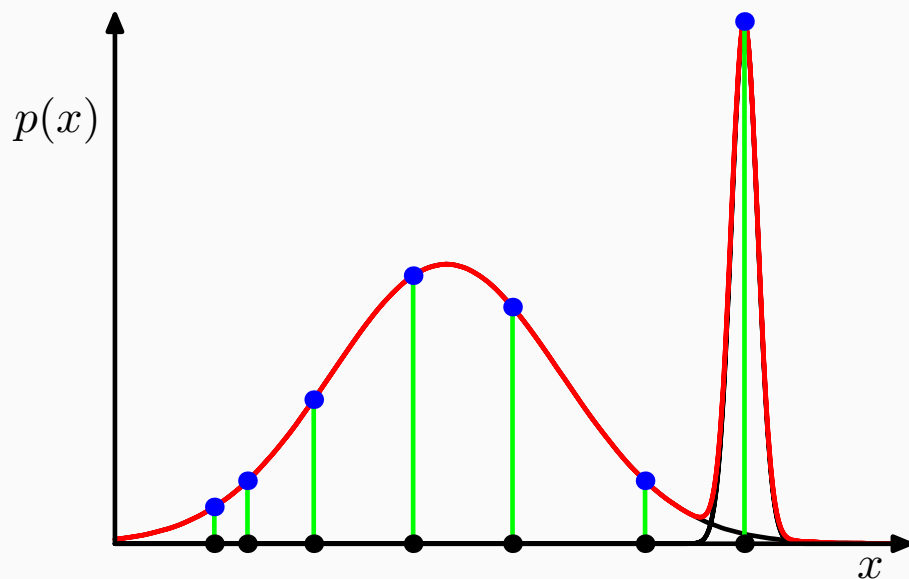
- Given the parameters, we calculate the posterior of the cluster membership

$$p(\mathbf{z}_n | \mathbf{X}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Why ? Leave it as your exercise

How to learn GMMs?

- Singularity issues



$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_j^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\sigma_j}$$

$$\sigma_j \rightarrow 0$$

Suppose we use diagonal covariance, when the cluster center is close to a data point, it tends to collapse onto the point to increase the likelihood

EM algorithm to learn GMM

- Can we get rid of the singularity issue?
- Can we jointly estimate the parameters and the posterior?

$$p(\mathbf{z}_1, \dots, \mathbf{z}_N | \mathbf{X})$$


 π

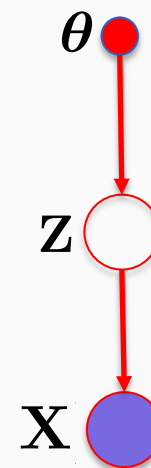
$$\begin{aligned} &\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K\} \\ &\{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\} \end{aligned}$$


Let us look at a general case

Suppose we have a model governed by parameters θ

$$p(\mathbf{X}|\theta) = \int p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z}$$

Observations Latent random variables



Question: what are θ and $\mathbf{Z}$ for GMMs?

EM algorithm: how to estimate θ

$$\log(p(\mathbf{X}|\theta)) = \log\left(\int p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z}\right)$$

$$= \log\left(\int \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} q(\mathbf{Z}) d\mathbf{Z}\right)$$

$$\geq \int q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}\right) d\mathbf{Z}$$

Jensen's inequality

$$\begin{aligned} \log(p(\mathbf{X}|\theta)) &= \underbrace{\int q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}\right) d\mathbf{Z}}_{L(\theta, q(\mathbf{Z}))} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} d\mathbf{Z}}_{\text{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta))} \\ &\geq 0 \end{aligned}$$

EM Algorithm

$$\log(p(\mathbf{X}|\boldsymbol{\theta})) = \underbrace{\int q(\mathbf{Z}) \log\left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})}\right) d\mathbf{Z}}_{L(\boldsymbol{\theta}, q(\mathbf{Z}))} + \underbrace{\int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})} d\mathbf{Z}}_{\text{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}))} \geq 0$$

$$\log(p(\mathbf{X}|\boldsymbol{\theta})) = L(\boldsymbol{\theta}, q^*(\mathbf{Z})) \text{ when } q^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$$

Now fix $q^*(\mathbf{Z})$

$$\boldsymbol{\theta}^{\text{new}} = \underset{\hat{\boldsymbol{\theta}}}{\operatorname{argmax}} L(\hat{\boldsymbol{\theta}}, q^*(\mathbf{Z}))$$

$$\log(p(\mathbf{X}|\boldsymbol{\theta}^{\text{new}})) \geq L(\boldsymbol{\theta}^{\text{new}}, q^*(\mathbf{Z})) \geq L(\boldsymbol{\theta}, q^*(\mathbf{Z})) = \log(p(\mathbf{X}|\boldsymbol{\theta}))$$

Like a bridge to improve the parameters!

EM Algorithm

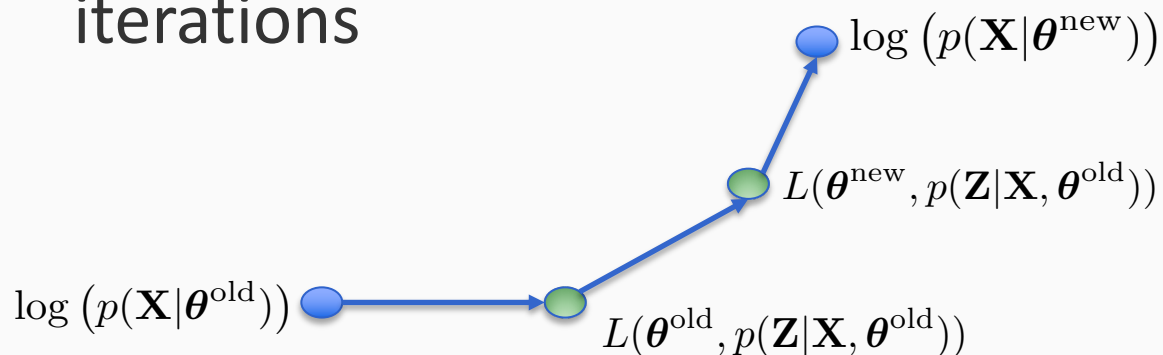
- Choose an initial setting θ^{new}
- Repeat

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}}$$

Evaluate $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ E step

Fix $q(\mathbf{Z})$, $\theta^{\text{new}} = \underset{\theta}{\operatorname{argmax}} L(\theta, q(\mathbf{Z}))$ M step

- Until $\|\theta^{\text{old}} - \theta^{\text{new}}\| \leq \epsilon$ or reach the maximum # of iterations



GMM revisited

- Given $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ and cluster number K
- For each data point n
 - Sample the cluster membership $\mathbf{z}_n$: $K \times 1$ one-hot vector,
 $z_{nk}=1$ means $\mathbf{x}_n$ belongs to cluster k

e.g., $[0, 0, 1, \dots, 0]^\top$

$$p(\mathbf{z}_n) = \prod_{k=1}^K \pi_k^{z_{nk}}$$

$$\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]^\top$$
$$\mathbf{z}_n = [z_{n1}, \dots, z_{nK}]^\top$$

- Given $\mathbf{z}_n$, sample $\mathbf{x}_n$ from the cluster-specific Gaussian

$$p(\mathbf{x}_n | \mathbf{z}_n) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

cluster center

soft width

EM algorithm for GMM

$$p(\mathbf{X}, \mathbf{Z} | \underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}_{\boldsymbol{\theta}}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

E step: $q(\mathbf{Z}) = \prod_{n=1}^N p(\mathbf{z}_n | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = \prod_{n=1}^N \prod_{k=1}^K \gamma_{nk}^{z_{nk}}$

$$\gamma_{nk} \equiv p(\mathbf{z}_{nk} = 1 | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})}{\sum_{j=1}^K \pi_j^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}})}$$

EM algorithm for GMM

$$p(\mathbf{X}, \mathbf{Z} | \underbrace{\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}}_{\boldsymbol{\theta}}) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

M step: $L(\boldsymbol{\theta}, q(\mathbf{Z})) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{q(\mathbf{Z})} \right)$

$$\begin{aligned} &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \log (p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})) + \text{const} \\ &= \mathbb{E}_{q(\mathbf{Z})} \log (p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})) + \text{const} \\ &\quad \mathbb{E}_{q(\mathbf{Z})} \left[\sum_{n=1}^N \sum_{k=1}^K z_{nk} \log \pi_k + z_{nk} \log (\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \right] \end{aligned}$$

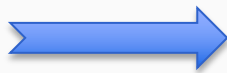
$$\sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \pi_k + \gamma_{nk} \log (\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

EM algorithm for GMM

- M step:

$$\max \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \pi_k + \gamma_{nk} \log (\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$$

$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^N \gamma_{nk}}{\sum_{k=1}^K \sum_{n=1}^N \gamma_{nk}}$$



$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} \mathbf{x}_n$$

$$N_k = \sum_{n=1}^N \gamma_{nk}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top$$

EM algorithm for GMMs

- E step $\gamma_{nk} \equiv p(\mathbf{z}_{nk} = 1 | \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) = \frac{\pi_k^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})}{\sum_{j=1}^K \pi_j^{\text{old}} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j^{\text{old}}, \boldsymbol{\Sigma}_j^{\text{old}})}$

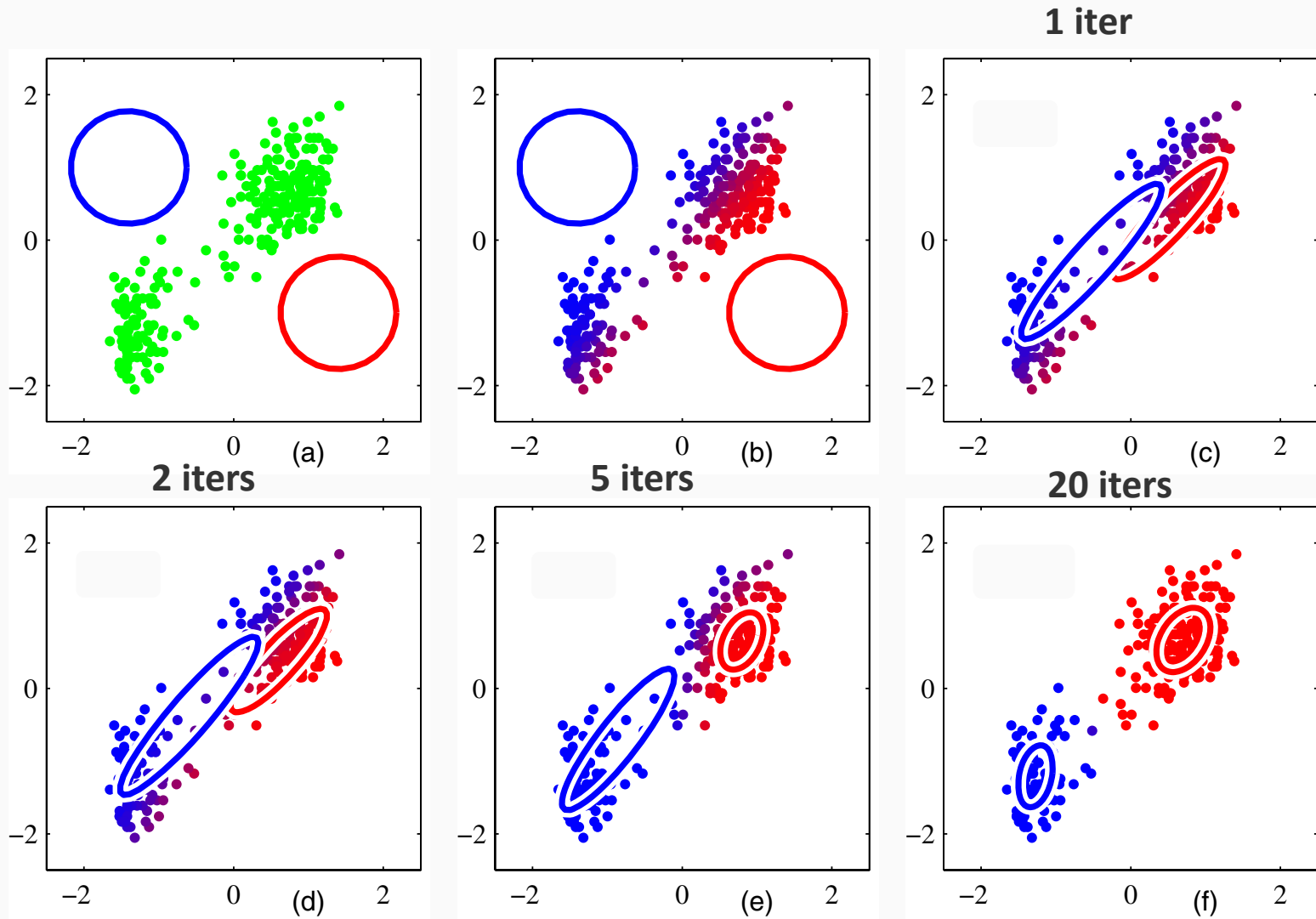
- M step
$$\pi_k^{\text{new}} = \frac{\sum_{n=1}^N \gamma_{nk}}{\sum_{k=1}^K \sum_{n=1}^N \gamma_{nk}}$$

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^\top$$

We do not have any singularity issues!

EM algorithm for GMMs



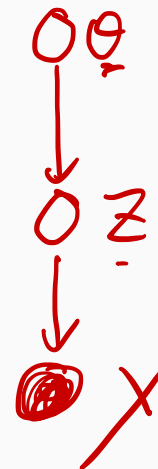
Practice

- Derive EM algorithm for mixture of Bernoulli distributions
- Derive EM algorithm for Bayesian linear regression

Outline

$$\log p(x, z | \theta)$$

$$\log p(x | \theta)$$



- Gaussian Mixture Model and EM algorithm
- Global variational Inference
- Local variational inference
- Variational message passing

$$\log p(x | \theta) = \int q(z) \log \frac{p(x, z | \theta)}{q(z)} dz + \underbrace{KL(q(z) \| p(z | x, \theta))}_{\geq 0}$$

$$\mathbb{E}_{q(z)} \log \frac{p(x, z | \theta)}{q(z)} = \mathcal{L}(\theta, q(z))$$

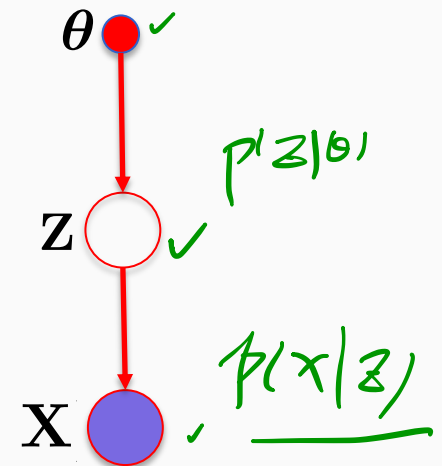
$$q^*(z) = \underline{p(z | x, \theta)}$$

$$\log p(x | \theta) = \mathcal{L}(\theta, q^*(z)) \xrightarrow{\mathcal{L}(\theta^{\text{new}}, q^*(z))} \frac{\log p(x | \theta^{\text{new}})}{p(z | x, \theta^{\text{new}})}$$

Global variational inference

- Consider a general model

$$\underbrace{p(\mathbf{X}|\boldsymbol{\theta})}_{\text{Observations}} = \int \underbrace{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}_{\text{Latent random variables}} d\mathbf{Z}$$



Put aside the parameters first (either we use full Bayesian treatment to absorb $\boldsymbol{\theta}$ into $\mathbf{Z}$ or consider $\boldsymbol{\theta}$ as fixed constant)

Global variational inference

Question: how to compute the posterior $p(\mathbf{Z}|\mathbf{X})$

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{Z})}{\int p(\mathbf{X}, \mathbf{Z}) d\mathbf{Z}}$$

Usually infeasible!!

GMMs, Bayesian linear regression are a few exceptions...

In most cases, you cannot get an analytical result

e.g.,

$$p(z) = \mathcal{N}(z|0, 1)$$
$$p(x|z) = \sigma(z)^x (1 - \sigma(x))^{1-z}$$

Global variational inference

Question: how to compute the posterior $p(\mathbf{Z}|\mathbf{X})$

- ✓ Idea: Now that the true posterior is complicated and tricky to compute, can we find a simple form of distribution (e.g., Gaussian) that approximates the true posterior? In other words, can we designate a family of simple distributions, from which we find the best member that is closest to the true posterior?

$$\underline{p(z)} = \mathcal{N}(z|0, 1) \quad \underline{p(x|z)} = \sigma(z)^x (1 - \sigma(z))^{1-x}$$

$p(z|x)$

Let us use a Gaussian $q(z) = N(z|\mu, \sigma^2)$ to approximate the true posterior $p(z|x)$

The problem is how to determine the best μ, σ^2

Intuitive thoughts

- Suppose we assume the family (form) of approximate posterior $q(\mathbf{Z}|\alpha)$

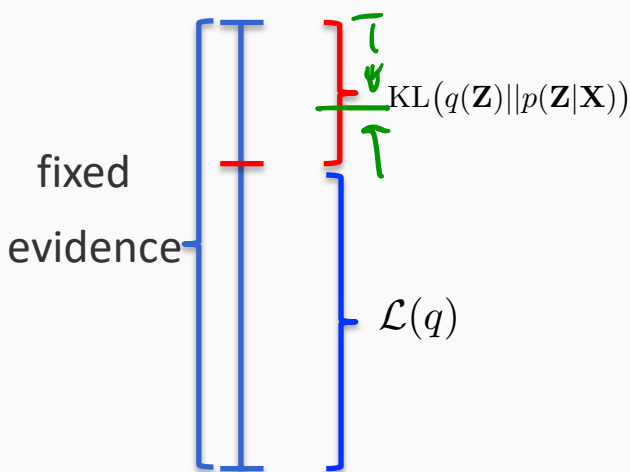
$$\alpha^* = \operatorname{argmin}_{\alpha} \operatorname{KL}(\underbrace{q(\mathbf{Z}|\alpha)}_{\text{red underline}} \parallel \underbrace{p(\mathbf{Z}|\mathbf{X})}_{\text{green circle}})$$


KL divergence is commonly used to measure the difference between distributions

What is the issue?

Detour: go back to what we have derived before

$$\underbrace{\log(p(\mathbf{X}))}_{\text{Evidence}} = \underbrace{\int q(\mathbf{Z}) \log\left\{\frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})}\right\} d\mathbf{Z}}_{\substack{\mathcal{L}(q) \\ \text{Variational Lower Bound}}} + \underbrace{\int q(\mathbf{Z}) \log\left\{\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})}\right\} d\mathbf{Z}}_{\text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}))} \geq 0$$



$$\max \mathcal{L}(q) \iff \min \text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}))$$

Key: Maximize the variational lower bound is equivalent to minimizing the KL divergence!

Global variational inference

- Given a family $\mathcal{S}$ of the approximate posterior $q(\mathbf{Z})$,

$$\underline{q^*(\mathbf{Z})} = \operatorname{argmax}_{q \in \mathcal{S}} \underline{\mathcal{L}(q)} = \int q(\mathbf{Z}) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} d\mathbf{Z}$$
$$\mathbb{E}_q \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\}$$

Usually there is a trade-off: The larger the family $\mathcal{S}$, the better the approximation quality, but the harder the optimization

Mean-field variational inference

$$\mathbf{z} = \begin{pmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 \\ \vdots \\ \mathbf{z}_n \end{pmatrix}$$

- Assume the approximate posterior is factorized:

$$q(\mathbf{Z}) = \prod_i q(\mathbf{Z}_i)$$

$$\mathbf{Z} = \{\mathbf{Z}_1, \mathbf{Z}_2, \dots\} \text{ nonoverlapping}$$

Each $q(\mathbf{Z}_i)$ is a free form distribution

$$\max_q \mathcal{L}(q) = \int \prod_i q(\mathbf{Z}_i) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{\prod_i q(\mathbf{Z}_i)} \right\} d\mathbf{Z}$$

$$\int \frac{q(\mathbf{z})}{q(\mathbf{z})} \log \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z})} d\mathbf{z}$$

Solve this optimization by alternative updating

Mean-field variational inference

$$p(\mathbf{z}_1, \dots, \mathbf{z}_3) = p(\mathbf{z}_1) p(\mathbf{z}_2) p(\mathbf{z}_3)$$

$\checkmark \quad \checkmark \quad \checkmark$
 $\checkmark \quad \checkmark \quad \checkmark$

$$\max \mathcal{L}(q) = \int \left(\prod_i q(\mathbf{z}_i) \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{\prod_i q(\mathbf{z}_i)} \right\} \right) d\mathbf{Z}$$

Update $q(\mathbf{z}_j)$ giving $\{q(\mathbf{z}_i)\}_{i \neq j}$ fixed

$$p(\mathbf{z}_{-j})$$

$$\log p(\mathbf{x}, \mathbf{z}) = \log \pi^\top \mathbf{u}(\mathbf{z}_i)$$

$$\mathcal{L}(q(\mathbf{z}_j)) = \int q(\mathbf{z}_j) \prod_{i \neq j} q(\mathbf{z}_i) \log(p(\mathbf{X}, \mathbf{Z})) d\mathbf{Z} - \int q(\mathbf{z}_j) \log(q(\mathbf{z}_j)) d\mathbf{z}_j + \text{const}$$

$\downarrow$

$$\mathbb{E}_{q(\mathbf{z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z}))$$

$$g(\mathbf{z}_j)$$

$$q(\mathbf{z}_{-j}) = \prod_{i \neq j} q(\mathbf{z}_i)$$

$$q(\mathbf{z}_i)$$

Solve this

$$q(\mathbf{z}_j) \propto \exp \left\{ \mathbb{E}_{q(\mathbf{z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z})) \right\}$$

$$q(\mathbf{z}_j) = \frac{\exp \left\{ \mathbb{E}_{q(\mathbf{z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z})) \right\}}{\int \exp \left\{ \mathbb{E}_{q(\mathbf{z}_{-j})} \log(p(\mathbf{X}, \mathbf{Z})) \right\} d\mathbf{z}_j}$$

$$\int q(\mathbf{z}_1) \left\{ q(\mathbf{z}_2) p(\mathbf{z}_2) \log p(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) d\mathbf{z}_2 d\mathbf{z}_3 \right\} d\mathbf{z}_1$$

$$\mathbb{E}_{q(\mathbf{z}_2, \mathbf{z}_3)} \log p(\mathbf{x}, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3)$$

$$\begin{aligned}
\checkmark L(q(z_j)) &= \int q(z_j) E_{q(z_j)} \log p(x, z) dz_j & \left. \begin{aligned} &= \int q \log \frac{q}{p} dx \\ &= \int q \log \frac{q}{p} dx \end{aligned} \right\} K_L(q||p) \\
&\quad - \int q(z_j) \log q(z_j) dz_j \\
&= \int q(z_j) \left[\underbrace{E_{q(z_j)} (\log p(x, z))}_{\log [\exp [E_{q(z_j)} (\log p(x, z))]]} - \log q(z_j) \right] dz_j \\
&= \int q(z_j) \log \left\{ \frac{\exp (E_{q(z_j)} \log p(x, z))}{q(z_j)} \right\} dz_j & q(z_j) \\
&= - \int q(z_j) \log \left(\frac{q(z_j)}{\exp (-)} \right) dz_j & T = \int \exp (E_{q(z_j)} \log p(x, z)) dz_j \\
&= - \int q(z_j) \log \left[\frac{q(z_j)}{\frac{\exp (-)}{T}} \right] dz_j \\
&= - \int q(z_j) \log \frac{q(z_j)}{\frac{\exp (-)}{T}} dz_j - \int q(z_j) \log \frac{1}{T} dz_j \\
&= - K_L(q(z_j) || \frac{\exp (-)}{T})
\end{aligned}$$

Mean-field variational Inference: algorithmic framework

- Choose a factorized posterior form $q(\mathbf{Z}) = \prod_i q(\mathbf{Z}_i)$
- Repeat
 - For each j
 - Fixed $\{q(\mathbf{Z}_i)\}_{i \neq j}$ and update $q(\mathbf{Z}_j) \propto \exp \{ \mathbb{E}_{q(\mathbf{Z}_{-j})} \log (p(\mathbf{X}, \mathbf{Z})) \}$
 - End for
- Until each $q(\mathbf{Z}_j)$ changes little or reach maximum # of iterations

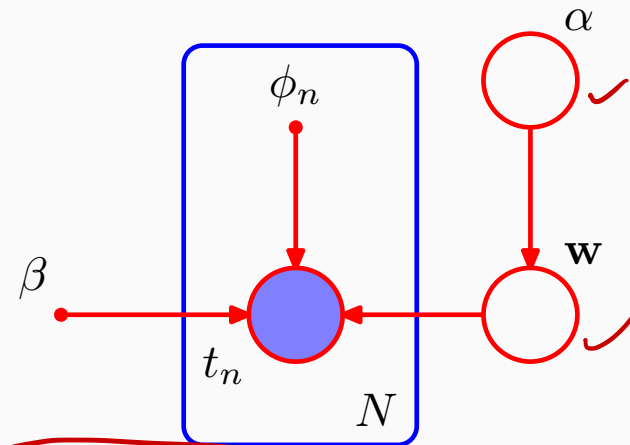
Variational Linear regression

$$p(\underline{\alpha}) = \text{Gam}(\underline{\alpha} | a_0, b_0)$$

$$p(\mathbf{w} | \alpha) = \mathcal{N}(\mathbf{w} | \mathbf{0}, \alpha^{-1} \mathbf{I})$$

$$p(\underline{\mathbf{t}} | \mathbf{w}) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi_n, \beta^{-1})$$

$$p(\underline{\mathbf{t}}, \mathbf{w}, \alpha) = p(\underline{\mathbf{t}} | \mathbf{w}) p(\mathbf{w} | \alpha) p(\alpha)$$



$$\mathcal{N}(\underline{\mathbf{t}} | \underline{\Phi} \mathbf{w}, \beta^{-1} \mathbf{I})$$

Observed data $\underline{\mathbf{t}} = [t_1, \dots, t_N]^T$

Inference task: $p(\mathbf{w}, \alpha | \underline{\mathbf{t}})$

Variational Linear regression

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

$$\underline{q(\mathbf{w}, \alpha)} = \underline{q(\mathbf{w})} \underline{q(\alpha)}$$

Alternative updating

$$\underline{q(\alpha)} \propto \exp\{\underline{\mathbb{E}_{q(\mathbf{w})} \log p(\mathbf{t}, \mathbf{w}, \alpha)}\}$$



$$q(\alpha) = \text{Gam}(\alpha | a_N, b_N)$$

$$\underline{a_N} = a_0 + \frac{d}{2}$$

$$\underline{b_N} = b_0 + \frac{1}{2} \mathbb{E}[\mathbf{w}^\top \mathbf{w}]$$

$$\begin{aligned} \log p(\mathbf{t}, \mathbf{w}, \alpha) &= \log p(\alpha | a_0, b_0) p(\mathbf{w} | \alpha) p(\mathbf{t} | \mathbf{w}) \\ &= \log p(\alpha | a_0, b_0) + \frac{\log p(\mathbf{w} | \alpha)}{\alpha^{a_0-1} e^{-b_0 \alpha}} \log \left(\frac{1}{2\pi\alpha} \right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\alpha} \mathbf{w}^\top \mathbf{w}\right) \\ &= (a_0 - 1) \log \alpha - b_0 \alpha + \frac{d}{2} \log \alpha - \frac{1}{2} \mathbf{w}^\top \mathbf{w} \alpha \end{aligned}$$

d: dimension of $\mathbf{w}$

Variational Linear regression

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

$$q(\mathbf{w}, \alpha) = \underline{q(\mathbf{w})} q(\alpha)$$

Alternative updating

$$\underline{q(\mathbf{w})} \propto \exp\{\mathbb{E}_{q(\alpha)} \log p(\mathbf{t}, \mathbf{w}, \alpha)\}$$



$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t}$$

$$\mathbf{S}_N = (\mathbb{E}[\alpha] \mathbf{I} + \beta \Phi^T \Phi)^{-1}$$

$$\begin{aligned} \log p(\mathbf{t}, \mathbf{w}, \alpha) &= \log p(\alpha | a_0, b_0) p(\mathbf{w} | \alpha) p(\mathbf{t} | \mathbf{w}) \\ &= \log p(\mathbf{w} | \alpha) + \log p(\mathbf{t} | \mathbf{w}) \\ &= -\frac{1}{2} \alpha \mathbf{w}^T \mathbf{w} - \frac{1}{2} \beta (\mathbf{t} - \Phi \mathbf{w})^T (\mathbf{t} - \Phi \mathbf{w}) \end{aligned}$$

$$-\frac{1}{2} \mathbf{w}^T \mathbf{A} \mathbf{w} + \mathbf{w}^T \mathbf{b}$$

$$f(x_1, x_2)$$

Variational Linear regression

Obviously, the posterior distribution is intractable, we introduce factorized approximation:

$$\underline{q(\mathbf{w}, \alpha)} = q(\mathbf{w}) \underline{q(\alpha)}$$

The required moments

$$\begin{aligned} \underline{\mathbb{E}[\alpha]} &= \underline{a_N / b_N} \\ \mathbb{E}[\underline{\mathbf{w}\mathbf{w}^T}] &= \underline{\mathbf{m}_N \mathbf{m}_N^T + \mathbf{S}_N} \end{aligned}$$

Predictive distribution

$$\begin{aligned} \underline{p(t|\mathbf{x}, \mathbf{t})} &= \int \underline{p(t|\mathbf{x}, \mathbf{w})} \underline{p(\mathbf{w}|\mathbf{t})} d\mathbf{w} \simeq \int \underline{p(t|\mathbf{x}, \mathbf{w})} q(\mathbf{w}) d\mathbf{w} \\ &= \mathcal{N}(t | \mathbf{m}_N^T \phi(\mathbf{x}), \sigma^2(\mathbf{x})) \end{aligned}$$

Exponential family

$$p(\boldsymbol{\eta}|\nu_0, \mathbf{v}_0) = f(\nu_0, \boldsymbol{\chi}_0)g(\boldsymbol{\eta})^{\nu_0} \exp \{ \nu_0 \boldsymbol{\eta}^T \boldsymbol{\chi}_0 \}$$

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\eta}) = \prod_{n=1}^N h(\mathbf{x}_n, \mathbf{z}_n)g(\boldsymbol{\eta}) \exp \{ \boldsymbol{\eta}^T \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \}$$



Task: $p(\boldsymbol{\eta}, \mathbf{Z}|\mathbf{X})$

Assume: $q(\mathbf{Z}, \boldsymbol{\eta})$ = $q(\mathbf{Z})q(\boldsymbol{\eta})$

Exponential family

The updates are analytical

$$q^*(\mathbf{z}_n) = h(\mathbf{x}_n, \mathbf{z}_n) g(\mathbb{E}[\boldsymbol{\eta}]) \exp \left\{ \mathbb{E}[\boldsymbol{\eta}^T] \mathbf{u}(\mathbf{x}_n, \mathbf{z}_n) \right\}$$

$$q^*(\boldsymbol{\eta}) = f(\nu_N, \boldsymbol{\chi}_N) g(\boldsymbol{\eta})^{\nu_N} \exp \left\{ \boldsymbol{\eta}^T \boldsymbol{\chi}_N \right\}$$

$$\begin{aligned} \nu_N &= \nu_0 + N \quad \checkmark \\ \boldsymbol{\chi}_N &= \boldsymbol{\chi}_0 + \sum_{n=1}^N \mathbb{E}_{\mathbf{z}_n} [\mathbf{u}(\mathbf{x}_n, \mathbf{z}_n)] \quad \checkmark \end{aligned}$$

Outline

- Gaussian Mixture Model and EM algorithm
- Global variational Inference
- **Local variational inference**
- Variational message passing
- Stochastic variational inference

Local Variational Inference

- Seeks an bound for a factor function of individual variables or a subset of variables
- Convex conjugate

$$f(x) = \max_{\lambda} \lambda x - g(\lambda)$$

$$g(\lambda) = \max_{\mathbf{x}} \lambda x - f(x)$$

- Key idea: if a factor is convex, use the convex conjugate obtain an bound (easier form)

Local Variational Inference

In general

$$\log(p(\mathbf{X}, \boldsymbol{\theta})) = \log(p(\boldsymbol{\theta})) + \sum_n \underbrace{\log(p(\mathbf{x}_n | \boldsymbol{\theta}))}$$


If it is convex to $\boldsymbol{\theta}$, $f(\boldsymbol{\theta}) \geq \boxed{\boldsymbol{\lambda}^\top \boldsymbol{\theta} - g(\boldsymbol{\lambda})}$

Variational logistic regression

- Let us consider the sigmoid function in the likelihood

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\log(\sigma(x)) = -\log(1 + e^{-x}) \text{ is concave}$$

Let's verify it

Variational logistic regression

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\log(\sigma(x)) = -\log(1 + e^{-x})$$

$$= -\log \{e^{-x/2}(e^{x/2} + e^{-x/2})\}$$

$$= x/2 - \log \{e^{x/2} + e^{-x/2}\}$$



$$f(x)$$

Variational logistic regression

$$f(x) = -\ln(e^{x/2} + e^{-x/2}) \quad \text{Symmetric}$$

$$f(x) = f(|x|) = f(\sqrt{x^2})$$

We can show that f is convex to x^2 , so we can use convex conjugate

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f\left(\sqrt{x^2}\right) \right\}$$

Variational logistic regression

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f\left(\sqrt{x^2}\right) \right\}$$



$$0 = \lambda - \frac{dx}{dx^2} \frac{d}{dx} f(x) = \lambda + \frac{1}{4x} \tanh\left(\frac{x}{2}\right)$$



ξ is the optimal x corresponding to λ

$$\lambda(\xi) = -\frac{1}{4\xi} \tanh\left(\frac{\xi}{2}\right) = -\frac{1}{2\xi} \left[\sigma(\xi) - \frac{1}{2} \right]$$

Note: $\lambda(\xi) = \lambda(-\xi)$

Variational logistic regression

$$g(\lambda) = \max_{x^2} \left\{ \lambda x^2 - f\left(\sqrt{x^2}\right) \right\}$$

$$g(\lambda) = \lambda(\xi)\xi^2 - f(\xi) = \lambda(\xi)\xi^2 + \ln(e^{\xi/2} + e^{-\xi/2})$$



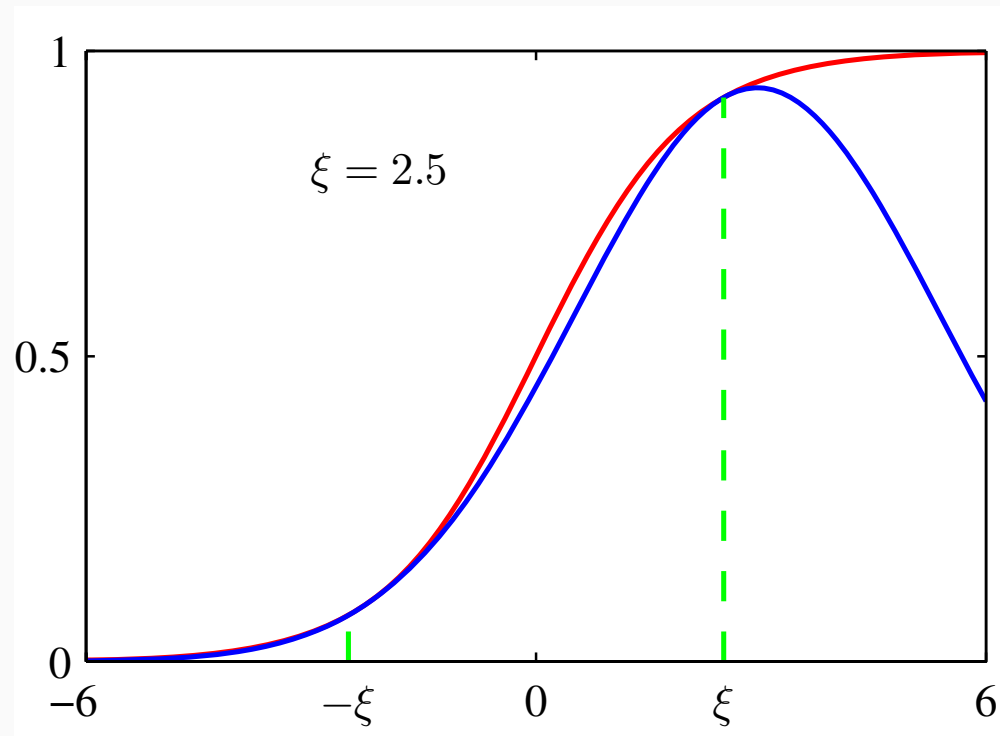
$$f(x) \geq \lambda x^2 - g(\lambda) = \lambda x^2 - \lambda \xi^2 - \ln(e^{\xi/2} + e^{-\xi/2})$$



$$\log(\sigma(x)) = x/2 + f(x)$$

$$\sigma(x) \geq \sigma(\xi) \exp \left\{ (x - \xi)/2 + \lambda(\xi)(x^2 - \xi^2) \right\}$$

Lower-bound of $\sigma(x)$



Variational logistic regression

- Given an arbitrary feature vector ϕ , the binary response t is sampled from

$$p(t|\mathbf{w}) = \sigma(a)^t \{1 - \sigma(a)\}^{1-t} \quad \boxed{a = \mathbf{w}^T \phi}$$

$$= \left(\frac{1}{1 + e^{-a}} \right)^t \left(1 - \frac{1}{1 + e^{-a}} \right)^{1-t}$$

$$= e^{at} \frac{e^{-a}}{1 + e^{-a}} = \boxed{e^{at} \sigma(-a)}$$

Variational logistic regression

From the previous result

$$\sigma(z) \geq \sigma(\xi) \exp \left\{ (z - \xi)/2 - \lambda(\xi)(z^2 - \xi^2) \right\}$$

$$\text{where } \lambda(\xi) = \frac{1}{2\xi} \left[\sigma(\xi) - \frac{1}{2} \right]$$

Note: We omit $-$ in the previous symbol



$$p(t|\mathbf{w}) = e^{at} \sigma(-a) \geq e^{at} \sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\}$$

Variational logistic regression

$$p(t|\mathbf{w}) = e^{at} \sigma(-a) \geq e^{at} \sigma(\xi) \exp \left\{ -(a + \xi)/2 - \lambda(\xi)(a^2 - \xi^2) \right\}$$

Given the design matrix (features after appropriate (nonlinear) transformations)

$$\Phi = [\phi_1, \dots, \phi_N]^\top$$

Binary output $\mathbf{t} = [t_1, \dots, t_N]^\top$

Each $t_n \in \{0, 1\}$

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) d\mathbf{w} = \int \left[\prod_{n=1}^N p(t_n|\mathbf{w}) \right] p(\mathbf{w}) d\mathbf{w}$$



$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}) \geq h(\mathbf{w}, \boldsymbol{\xi})p(\mathbf{w})$$

$$h(\mathbf{w}, \boldsymbol{\xi}) = \prod_{n=1}^N \sigma(\xi_n) \exp \left\{ \mathbf{w}^\top \phi_n t_n - (\mathbf{w}^\top \phi_n + \xi_n)/2 - \lambda(\xi_n)([\mathbf{w}^\top \phi_n]^2 - \xi_n^2) \right\}.$$

Variational logistic regression

- Consider approximate posterior $q(\mathbf{w})$

$$\max \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{t}, \mathbf{w})}{q(\mathbf{w})} \right\} \quad \longleftrightarrow \quad \min \text{KL}(q(\mathbf{w}) \| p(\mathbf{w} | \mathbf{t}))$$

Infeasible!

Also, if you directly optimize w.r.t a free from, you obtain the true posterior

Solution: We maximize its variational lower bound!

$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{t} | \mathbf{w}) p(\mathbf{w}) \geq h(\mathbf{w}, \boldsymbol{\xi}) p(\mathbf{w})$$

Why called "variational LB"?

It's possible to take equality

$$\max \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w}) h(\mathbf{w}, \boldsymbol{\xi})}{q(\mathbf{w})} \right\}$$

Variational logistic regression

$$\max \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w}) h(\mathbf{w}, \boldsymbol{\xi})}{q(\mathbf{w})} \right\}$$



The same as Mean-Field

$$q(\mathbf{w}) \propto \exp \left(\underbrace{\log \{ p(\mathbf{w}) h(\mathbf{w}, \boldsymbol{\xi}) \}}_{-\frac{1}{2}(\mathbf{w} - \mathbf{m}_0)^T \mathbf{S}_0^{-1} (\mathbf{w} - \mathbf{m}_0)} \right) \quad p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0)$$

$$+ \sum_{n=1}^N \left\{ \mathbf{w}^T \boldsymbol{\phi}_n (t_n - 1/2) - \lambda(\xi_n) \mathbf{w}^T (\boldsymbol{\phi}_n \boldsymbol{\phi}_n^T) \mathbf{w} \right\} + \text{const}$$



Complete squares

$$q(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N)$$

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1} \mathbf{m}_0 + \sum_{n=1}^N (t_n - 1/2) \boldsymbol{\phi}_n \right)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + 2 \sum_{n=1}^N \lambda(\xi_n) \boldsymbol{\phi}_n \boldsymbol{\phi}_n^T.$$

Are we done?

- No, we haven't identified the variational parameters ξ

$$\max_{\xi} \quad \mathcal{L}(q, \xi) = \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w}) h(\mathbf{w}, \xi)}{q(\mathbf{w})} \right\}$$



$$\sum_{n=1}^N \left\{ \ln \sigma(\xi_n) - \xi_n/2 - \lambda(\xi_n) (\phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n - \xi_n^2) \right\} + \text{const}$$



$$0 = \lambda'(\xi_n) (\phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n - \xi_n^2)$$

$$(\xi_n^{\text{new}})^2 = \phi_n^T \mathbb{E}[\mathbf{w}\mathbf{w}^T] \phi_n = \phi_n^T (\mathbf{S}_N + \mathbf{m}_N \mathbf{m}_N^T) \phi_n$$

Variational logistic regression

- We conduct an EM procedure

$$\max \mathcal{L}(q, \xi) = \mathbb{E}_{q(\mathbf{w})} \log \left\{ \frac{p(\mathbf{w}) h(\mathbf{w}, \xi)}{q(\mathbf{w})} \right\}$$

E step: update $q(\mathbf{w})$

M step: update ξ

Alternatively maximize the variational lower bound

Why is it called **variational** bound

- The variational bound is variational transformation, it means, if you **do NOT restrict the range of the variational parameters**, they always have settings that reach **equality**

$$\log(p(\mathbf{t})) \geq \mathbb{E}_{q(\mathbf{w})} \log\left\{\frac{p(\mathbf{t}, \mathbf{w})}{q(\mathbf{w})}\right\}$$

$$\mathbb{E}_{q(\mathbf{w})} \log\left\{\frac{p(\mathbf{t}, \mathbf{w})}{q(\mathbf{w})}\right\} \geq \mathcal{L}(q, \xi) = \mathbb{E}_{q(\mathbf{w})} \log\left\{\frac{p(\mathbf{w})h(\mathbf{w}, \xi)}{q(\mathbf{w})}\right\}$$

In practice, we often restrict the family/range of the variational parameters to gain the computational easiness

Variational EM algorithm

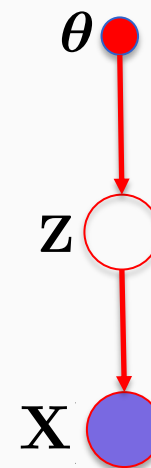
- In general, if we also need to estimate hyper-parameters.

$$p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})$$

$$\max_{q(\mathbf{Z}), \boldsymbol{\theta}} \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{Z})} \log \left\{ \frac{p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

E step: $q(\mathbf{Z}) \leftarrow \operatorname{argmax}_{q \in \mathcal{S}} \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta})$ fix $\boldsymbol{\theta}$

M step: $\boldsymbol{\theta} \leftarrow \operatorname{argmax}_{\boldsymbol{\theta}} \mathcal{L}(q(\mathbf{Z}), \boldsymbol{\theta})$ fix $q(\mathbf{Z})$



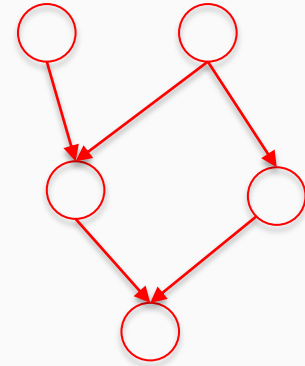
Variational Message Passing

- Consider a Bayesian network

$$p(\mathbf{x}) = \prod_i p(\mathbf{x}_i | \text{pa}(\mathbf{x}_i))$$

- Assume a factorized posterior over the nodes

$$q(\mathbf{x}) = \prod_i q_i(\mathbf{x}_i)$$



Variational Message Passing

- Consider the update on each node

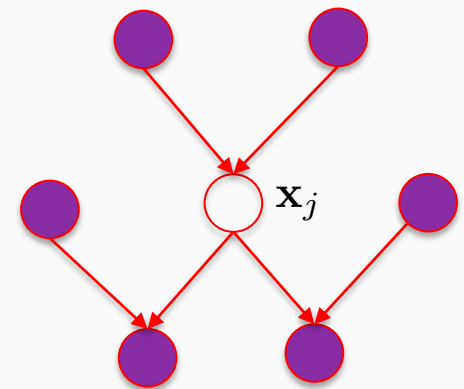
$$q(\mathbf{x}_j) \propto \exp\{\mathbb{E}_{q(\mathbf{x}_{\neg j})}[\log p(\mathbf{x})]\}$$

Questions: which factors involve $\mathbf{x}_j$?

The conditional probabilities where $\mathbf{x}_j$ is a parent/child

$$q(\mathbf{x}_j) \propto \exp\{\mathbb{E}[\log p(\mathbf{x}_j | \text{pa}(\mathbf{x}_j))]\} \\ + \sum_{\mathbf{x}_t \in \text{pa}(\mathbf{x}_t)} \mathbb{E}[\log p(\mathbf{x}_t | \mathbf{x}_j, \text{pa}(\mathbf{x}_t) \setminus \{\mathbf{x}_j\})]\}$$

Markov blanket



Variational Message Passing

- Given a graphical model, the mean-field variational update only requires a local computation on the graph

What you need to know

- What is EM algorithm
- $\text{Log}(\text{Evidence}) = \text{Variational Lower Bound} + \text{KL}$
- What is EM algorithm
- Global variational inference, mean-field
- General update in exponential family
- Local variational inference, convex conjugate
- Variational message passing
- Being able to derive and implement variational inference