Notes of Machine Learning
Approximate Methods

Xiangli Chen

Computer Science Department
University of Illinois at Chicago

July 23, 2017

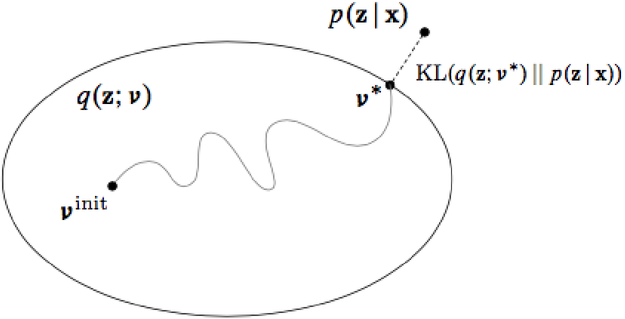
Approximate Methods

Motivation: in many prediction problems, it is infeasible to evaluate $p(z|x)$ or expectation with respect to $p(z|x)$.
Methods:

- **Deterministic approximate**: analytical approximation to $p(z|x)$
  e.g. factorization, specific parametric form
  variational inference (a family of approximaion techniques)
  - never generate exact result
- **Stochastic approximate**: sampling methods to $p(z|x)$
  e.g. Markov Chain Monte Carlo (MCMC)
  - computationally demanding
  - difficult to know if samples are generated independently

Outline

Variational Inference



Variational inference turns inference into optimization

- restrict the variational family $q(z; v)$ to be sufficiently tractable
- allow $q(z; v)$ to be sufficiently rich and flexible providing a good approximate

Outline

Mean Field Theory

Ideally

$$\hat{f}(x) = \arg\min_{f(z)} KL(f(z)||f(z|x)) = f(z|x)$$

Unfortunately, in many problems, $f(z|x)$ is not tractable.

Mean field

Assume $f(z)$ can be factorized as

$$f(z) = \prod_{i=1}^{M} f(z_i)$$

Recall EM algorithm in general

$$\log f(x) = L(f(z)) + KL(f(z)||f(z|x))$$

$$\arg\max_{f(z_i)'s} L(f(z)) = \arg\min_{f(z_i)'s} KL(f(z)||f(z|x))$$

Optimizing with respect to each of $f(z_i)$ in turn.

Mean Field Solution

$$
\begin{aligned}
L(f(z)) &= \int \prod_i f(z_i)(\log f(x,z) - \sum_i \log f(z_i))dz \\
&= \int f(z_j) \left(\int \log f(x,z) \prod_{i \neq j} f(z_j) dz_i \right) dz_j - \int f(z_j) \log f(z_j) dz_j + \text{const} \\
&= \underbrace{\int f(z_j) \log f(z_j|x) dz_j - \int f(z_j) \log f(z_j) dz_j}_{-KL(f(z_j|x)||f(z_j))} + \text{const}
\end{aligned}
$$

Let

$$
\mathbb{E}_{i \neq j}[\log f(x,Z)] = \int \log f(x,z) \prod_{i \neq j} f(z_j) dz_i
$$

then

$$
\log f(z_j|x) = \mathbb{E}_{i \neq j}[\log f(x,Z)] + \text{const}
$$

where the const gurantees that $\int f(z_j|x)dz_j = 1$.

Mean Field Algorithm

So we have (maximize negative KL divergence)

$$\forall j \qquad \hat{f}(z_j) = \arg\max_{f(z_j)} L(f(z)) = f(z_j|x)$$

that

$$\log \hat{f}(z_j) = \mathbb{E}_{i \neq j}\left[\log f(x, Z)\right] + \text{const.}$$

$\hat{f}(z_j)$ doesn't represent an explicit solution (depends on $f(z_i)$ for $i \neq j$).
Iterative learning algorithm

- Initialize $f(z_i)$ for $\forall i$
- do
$$\hat{f}^{m+1}(z_i) = \arg\max_{f(z_i)} L(\hat{f}(z))$$
where $\hat{f}(z) : \hat{f}^{m+1}(z_1), \cdots, \hat{f}^{m+1}(z_{i-1}), \hat{f}^{m}(z_i), \cdots, \hat{f}^{m}(z_M)$

Converge is guranteed ($L(f(z))$ is convex with respect to each $f(z_i)$).

Others

- Properties of factorized approximations
- Examples
  - The univariate Gaussian
  - Variational mixture of Gaussian
  - Variational linear regression
  - Variational logistic regression
- Local variational methods
- Expectation propagation

To be continued

Outline

Sampling

Monte Carlo techniques: approximate inference methods based on numerical sampling.

- Sampling from high-dimensional, complicatd distributions
- Bayesian inference and learning
  - Marginalization: $f(x) = \int_z f(x, z)dz$
  - Normalization: $f(x|y) = \frac{f(y|x)f(x)}{\int_X f(y|x)f(x)dx}$
  - Expectation: $\mathbb{E}_f[g(X)] = \int_X g(x)f(x)dx$
- Global optimization $\arg\max_x f(x)$