本篇文章从机器学习的宏观问题讲述，然后再着重于近似推断，读者读完后会对机器学习的整个问题有大致的理解，全程干货。

# 前言

## 频率角度 （本质为优化问题）

### 回归问题：

以最简单的逻辑回归为例：

$f(w) = w^\top x$, Dataset = $(x_i, y_i)$

Loss function:

$$\begin{cases} L(w) = \sum_{i=1}^{N} ||w^\top x_i - y_i||^2 \\ \hat{w} = \arg\min L(w) \end{cases}$$

解法：

1. 解析解（求导，令导数为0， concave）

$$\frac{\partial L(W)}{\partial W} = 0 \Rightarrow w^* = \left(X^\top X\right)^{-1} X^\top Y$$

2. 数值解： GD： Gradient Decent/SGD

### SVM（有约束优化）

$$f(w) = \text{sign}\left(w^\top x + b\right)$$

Loss function:

$$min \frac{1}{2} w^\top w$$

$$y_i \cdot \left(w^\top x_i + b\right) \geqslant 1$$
$$i = l, \cdots, N$$

求法： QP/Lagrange/对偶

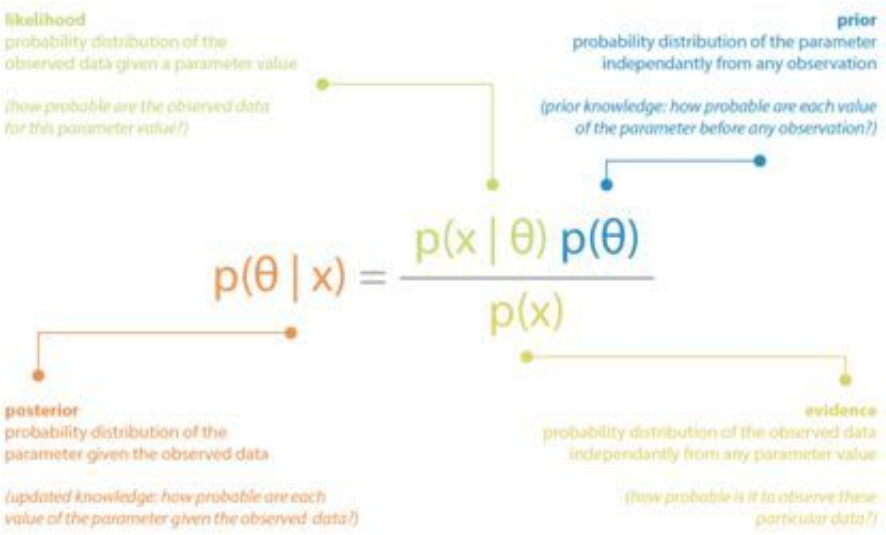
## EM

$$\hat{\theta} = \arg max \log P(x|\theta)$$

$$\theta^{(t+1)} = \arg\max_{\theta} \int_t \log P(x, z|\theta) \cdot P\left(z|x, \theta^{(t)}\right) dz$$

## 贝叶斯角度 （本质为积分问题）

### 贝叶斯推断




$$P(\theta|x) = \frac{P(x|\theta) \cdot P(\theta)}{P(x)}$$

$$P(x) = \int_\theta P(x|\theta) \cdot P(\theta) d\theta$$

Baysian Inference即贝叶斯推断，贝叶斯推断的过程主要是求后验概率(posterior)的过程。

1. 精确推断
2. 近似推断(Approximate Inference)
    1. 确定性推断- deterministic approximation
        1. 变分推断（Variational Inference）
    2. 随机推断-Stochastic Approximation
        1. MCMC
        2. MH
        3. Gibbs Sampling

## 贝叶斯决策

已知有个N个样本数据集 x，求新的样本$\hat{x}$, $P(\hat{x}|x)$ 的概率

$$
\begin{aligned}
P(\hat{x}|x) &= \int_{\theta} p(\tilde{x}, \theta|x) d\theta \\
&= \int_{\theta} p(\tilde{x}|\theta, \tilde{x}) \cdot p(\theta|x) d\theta \\
&= \int_{\theta} p(\tilde{x}|\theta) \cdot p(\theta|x) d\theta \\
&= E_{\theta|x}[P(\hat{x}|\theta)]
\end{aligned}
$$

$\tilde{x}, \theta$ 相互独立

因此可以看到，只要求得posterior就可以进行贝叶斯决策，因此求后验分布至关重要。

## 公式推导

X: Observed data; Z: latent variable parameter; (X, Z): complete data

$$
\begin{aligned}
P(X)P(Z|X) &= P(X,Z) \\
P(X) &= \frac{P(X,Z)}{P(Z|X)} \\
\log P(X) = \log \frac{P(X,Z)}{P(Z|X)} &= \log \frac{P(X,Z)}{q(Z)} \frac{q(Z)}{P(Z|X)} \\
&= \log \frac{P(X,Z)}{q(Z)} + \log \frac{q(Z)}{P(Z|X)} \\
&= \log \frac{P(X,Z)}{q(Z)} - \log \frac{P(Z|X)}{q(Z)} \\
&= \int \log \frac{P(X,Z)}{q(Z)} q(Z) dZ - \int \log \frac{P(Z|X)}{q(Z)} q(Z) dZ \\
&= E_q[\log \frac{P(X,Z)}{q(Z)}] + KL(q(Z) \| P(Z|X)) \\
&= E_q[\log P(X,Z) - \log q(Z)] + KL(q(Z) \| P(Z|X)) \\
&= E_q[\log P(X,Z)] + H[q] + KL(q||P) \\
&= ELBO : \mathcal{L}(X,Z,q) + KL(q||P)
\end{aligned}
$$

$$
\begin{aligned}
P(X)P(Z|X) &= P(X,Z) \\
P(X) &= \frac{P(X,Z)}{P(Z|X)} \\
\log P(X) &= \log P(X,Z) - \log P(Z|X) \\
&= \log \frac{P(X,Z)}{q(Z)} - \log \frac{P(Z|X)}{q(Z)} \\
left &= \int_z \log P(X) q(Z) dZ = \log P(X) \\
right &= \int_z q(Z) \cdot \log \frac{P(X,Z)}{q(Z)} dZ - \int_z q(Z) \log \frac{P(Z|X)}{q(Z)} dZ \\
&= ELBO(EvidenceLowerBound) \quad - \quad KL(q\|p) \\
&= \mathcal{L} + KL(q\|p) \\
&= variational + \geq 0
\end{aligned}
$$

The goal is to find $q(Z) \approx P(Z|X)$

$$\hat{q}(z) = \arg\max_{q(z)} \mathcal{L}(q) \Rightarrow \hat{q}(z) \approx p(z|x)$$

Mean field theory:

$$q(Z) = \prod_{i=1}^{M} q_i(z_i)$$

$$\begin{aligned}
\mathcal{L}(q) &= \int_z q(Z)\log p(X,Z)dZ - \int_z q(Z)\log q(Z)dZ \\
&= Part_1 \quad - Part_2 \\
&= \int_{z_j} q_j(z_j)\cdot \log \frac{\hat{P}(X,Z_j)}{q_j(z_j)} dz_j \\
Part_1 &= \int_z \prod_{i=1}^{M} q_i(z_i)\log P(X,Z)dz_1 dz_2\cdots dz_n \\
&= \int_{z_j} q_j(z_j)\left(\int_{z_1}\int_{z_2}\cdots_{z1,\cdots,z_n,\neq z_j}\int_{z_n}\prod_{i\neq j}^{m} q_i(z_i)\log P(X,Z)\cdot dz_1,\cdots,dz_n\right)dz_j \\
&= \int_{z_j} q_j(z_j)\cdot E_{\prod_{i\neq j}^{M} q_i(z_i)}[\log P(X,Z)]dz_j \\
&= \int_{z_j} q_j(z_j)\cdot \log \hat{P}(X,Z_j)dz_j \\
&= -KL(q_j\|\hat{P}(X,Z_j)) \leq 0 \\
Part_2 &= \int_z q(Z)\log q(Z)dZ \\
&= \int_{z_i}\prod_{i=1}^{M} q_i(z_i)\cdot\sum_{i=1}^{M}\log q_i(z_i)\,dz \\
&= \int_{z_i}\prod_{i=1}^{M} q_i(z_i)\left[\log q_1(z_1)+\log q_2(z_2)+\cdots+\log q_m(z_m)\right]dz \\
&= \sum_{i=1}^{M}\int_{z_i} q_i(z_i)\log q_i(z_i)\,dz_i
\end{aligned}$$

because we only care about j:

$$= \int_{z_j} q_j(z_j)\log q_j(z_j)\,dz_j + C$$

now we just consider the first one:

$$\begin{aligned}
\int_z \prod_{i=1}^{m} q_i(z_i)\cdot\log q_1(z_1)\,dz &= \int_z q_1 q_2\cdots q_m\cdot\log q_1 dz \\
&= \int_{z_1}\cdots\int_{z_m} q_1 q_2\cdots q_m\cdot\log q_1 dz_1 dz_2\cdots dz_m \\
&= \int_{Z_1} q_1\log q_1 dz_1\cdot\int_{z_2} q_2 dz_2\cdots\int_m q_m dz_m \\
&= \int_{Z_1} q_1\log q_1 dz_1
\end{aligned}$$

## Via Jensen's Inequality(ELBO)

$$P(X) = \int_Z P(X, Z) dZ$$

$$\log P(X) = \log \left[ \int_Z P(X, Z) dZ \right]$$

Introducing another distribution on Z, we can rewrite as;

$$\log P(X) = \log \left[ \int_Z P(X, Z) \frac{q(Z)}{q(Z)} dZ \right] = \log E_q \left[ \frac{P(X, Z)}{q(Z)} \right]$$

By Jensen's Inequality we know $\log E_q \left[ \frac{P(X, Z)}{q(Z)} \right] \geq E_q \left[ \log \frac{P(X, Z)}{q(Z)} \right]$

$$\log P(X) = \log E_q \left[ \frac{P(X, Z)}{q(Z)} \right] \geq E_q \left[ \log \frac{P(X, Z)}{q(Z)} \right] = \mathcal{L}$$

## Relative Entropy(KL divergence)

$$\begin{aligned}
\mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \mathrm{d}\mathbf{x} - \left( -\int p(\mathbf{x}) \ln p(\mathbf{x}) \mathrm{d}\mathbf{x} \right) \\
&= -\int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} \mathrm{d}\mathbf{x} \\
&= -E_p[\ln \frac{q(\mathbf{x})}{p(\mathbf{x})}] \geq -\ln E_p[\frac{q(\mathbf{x})}{p(\mathbf{x})}] \\
&= -\ln \int p(x) \cdot \frac{q(\mathbf{x})}{p(\mathbf{x})} dx \\
&= -\ln \int q(x) dx = -\ln 1 = 0
\end{aligned}$$

So $KL(p\|q)$ is nonnegative.

By Jensen's Inequality we know $\log E_q \left[ \frac{P(X, Z)}{q(Z)} \right] \geq E_q \left[ \log \frac{P(X, Z)}{q(Z)} \right]$

# 近似推断

## Big Picture

最大似然估计属于频率派统计 (frequentist statistics) 的方法，即对模型的参数 $\theta$ 进行点估计，然后基于该估计对新来的样本做预测。这种视角下：真实参数 $\theta$ 是未知的**定值**，而数据集是由该分布下产生的随机变量。最大似然方法即是一种对真实参数 $\theta$ 进行点估计的方法，做法即是求得参数在数据集上的似然值，取似然值最大的那组参数 $\hat{\theta}$ 来作为真实参数 $\theta$ 的估计。

VI 属于贝叶斯统计 (Bayesian statistics) 的范畴，这种视角下，概率反映的是知识状态的确定性程度。数据集不是随机变量，而是固定的观测。真实参数 $\theta$ 是未知和不确定的，被认为是随机变量。在这个范畴下，由先验概率 $\theta$ 的概念，我们用它来表示对参数$\theta$ 已经有的一些知识。先验 $P(\theta)$ 是关于参数 $\theta$ 的分布，一般将这个分布定义为简单的均匀分布或其他一些熵值较大的分布。

问题所在：

这个模型会含有确定性参数，譬如Z，这有可能存在在数据纬度比较大的情况下，隐藏变量的空间域太大而不能直接计算。在大部分连续型的变量中，需要的积分运算没有很好的解决方案。对于离散型变量，边缘概率涉及到所有因变量的求和，并且在实际运用过程中会有成倍的隐状态，导致计算十分昂贵。

- Posterior is intractable for large $n$, and we might want to add priors

$$p(\mu_{1:K}, z_{1:n} \mid x_{1:n}) = \frac{\prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})}{\int_{\mu_{1:K}} \sum_{z_{1:n}} \prod_{k=1}^{K} p(\mu_k) \prod_{i=1}^{n} p(z_i) p(x_i \mid z_i, \mu_{1:K})} \quad (2)$$

再比如

假设我们观测到了一组样本 $x^{(1)}, \cdots, x(m)$ ，此时，我们可以在先验的基础上不断修正 $\theta$ 的分布，使得熵值越来越低，写成贝叶斯规则就是：

$$p\left(\theta | x^{(1)}, \cdots, x^{(m)}\right) = \frac{p\left(x^{(1)}, \cdots, x^{(m)} | \theta\right) p(\theta)}{p\left(x^{(1)}, \cdots, x^{(m)}\right)}$$

其中， $p\left(x^{(1)}, \cdots, x^{(m)} | \theta\right)$ 是参数 $\theta$ 在目前观测下的似然。整个过程直观上理解就是：观测到的样本集合逐步地对参数 $\theta$ 施加影响，由原来熵值较高的先验分布得到了熵值较低的后验分布 (即$\theta$ 只在若干值得较大的概率值，大多数地方取值很小)。

对于新来的数据 $x^{m+1}$ ，最大似然估计使用参数 $\theta$ 的点估计值来对样本的概率值作出预测，而贝叶斯估计则使用参数 $\theta$ 的全分布来进行预测：

$$p\left(x^{(m+1)}|x^{(1)},\cdots,x^{(m)}\right)=\int p\left(x^{(m+1)}|\theta\right)p\left(\theta|x^{(1)},\cdots,x^{(m)}\right)d\theta$$

这里可以看出两点：

- 在贝叶斯估计中，估计的参数$\theta$的不确定性是直接包含在预测过程中的；对应的最大似然估计则是用过预测结果的方差来反映这种不确定性。
- 贝叶斯估计的计算量很大，涉及后验概率分布的计算以及参数$\theta$全分布的处理。

**所以：**

在这个情况，我们需要需要采用approximation schemes，这个scheme分为两大类，分别是stochastic和determini\thetain Monte Carlo，它需要很大的计算量，计算结果产生于有限的CPU时间。在实现运用，采样方法可能会对计算需求很大，因此经常会限制他们，仅用于小规模的问\theta度函数进行估计的方法，另一类方法是基于采样的方法 (Markov chain Monte Carlo, MCMC)，这两类方法可以有一个简单的对比，VI 更适合于大数据集场景下（VI 将推断视作优化问题)，而 MCMC 更适合于数据集高质量但是却很少的情况下。

在这个文章，我们首先讨论deterministic approximation。我们统一将approximation techniques称为variational inference或者variantinal bayes（变分推理或者变分贝叶斯）

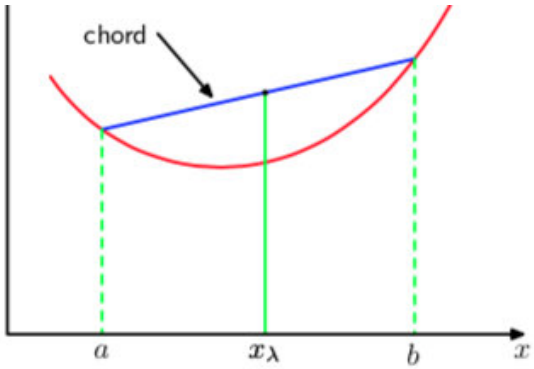
## Variational Inference

变分方法起源于18世纪Euler, Lagrange的研究。标准的微积分是关注于找到函数的导数，而这个导数就是体现了当我们对输入参数做了极小的改变的时候怎么影响到输出的参数。

同样地，我们可以定义一个函数entropy $H[p]$ ,输入一个概率分布，返回它的quantity。

$$\mathrm{H}[p] = \int p(x) \ln p(x) \, \mathrm{d}x$$

假设我们有一个完全贝叶斯模型，它的所有参数都用于先验分布。这里我们标记Z为隐藏变量，X为观察变量。




and the corresponding value of the function is $f(\lambda a + (1-\lambda)b)$. Convexity then implies

$$f(\lambda a + (1-\lambda)b) \leqslant \lambda f(a) + (1-\lambda) f(b). \tag{1.114}$$

This is equivalent to the requirement that the second derivative of the function be everywhere positive. Examples of convex functions are $x \ln x$ (for $x > 0$) and $x^2$. A function is called *strictly convex* if the equality is satisfied only for $\lambda = 0$ and $\lambda = 1$. If a function has the opposite property, namely that every chord lies on or below the function, it is called *concave*, with a corresponding definition for *strictly concave*. If a function $f(x)$ is convex, then $-f(x)$ will be concave.

Using the technique of proof by induction, we can show from (1.114) that a convex function $f(x)$ satisfies

$$f\left(\sum_{i=1}^{M} \lambda_i x_i\right) \leqslant \sum_{i=1}^{M} \lambda_i f(x_i) \tag{1.115}$$

where $\lambda_i \geqslant 0$ and $\sum_i \lambda_i = 1$, for any set of points $\{x_i\}$. The result (1.115) is known as *Jensen's inequality*. If we interpret the $\lambda_i$ as the probability distribution over a discrete variable $x$ taking the values $\{x_i\}$, then (1.115) can be written

$$f(\mathbb{E}[x]) \leqslant \mathbb{E}[f(x)] \tag{1.116}$$

where $\mathbb{E}[\cdot]$ denotes the expectation. For continuous variables, Jensen's inequality takes the form

$$f\left(\int \mathbf{x} p(\mathbf{x}) \, \mathrm{d}\mathbf{x}\right) \leqslant \int f(\mathbf{x}) p(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \tag{1.117}$$

We can apply Jensen's inequality in the form (1.117) to the Kullback-Leibler divergence (1.113) to give

$$\mathrm{KL}(p\|q) = -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x} \geqslant -\ln \int q(\mathbf{x}) \, \mathrm{d}\mathbf{x} = 0 \tag{1.118}$$

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

where we have defined

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z} \\ \mathrm{KL}(q\|p) &= -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}.\end{aligned}$$

我们的目标是限制$q(Z)$ 的分布为可以运算处理的分布，并且能提供一个很好的后验分布的预估。

用参数分布 $q(Z|\omega)$ 来限制the family of approximating distributions，然后我们可以用非线性优化来找出最优的参数值。譬如，高斯分布就是典型的一个变分分布。

- We measure the closeness of distributions using Kullback-Leibler Divergence

$$\mathrm{KL}(q \,||\, p) \equiv \mathbb{E}_q \left[ \log \frac{q(Z)}{p(Z \,|\, x)} \right] \tag{4}$$

- Characterizing KL divergence
  - If $q$ and $p$ are high, we're happy
  - If $q$ is high but $p$ isn't, we pay a price
  - If $q$ is low, we don't care
  - If $\mathrm{KL} = 0$, then distribution are equal

### Relation to KL Divergence

- Conditional probability definition

$$p(z \mid x) = \frac{p(z, x)}{p(x)} \tag{5}$$

- Plug into KL divergence

$$\begin{aligned}
\mathrm{KL}(q(z) \,||\, p(z \mid x)) =& \mathbb{E}_q \left[ \log \frac{q(z)}{p(z \mid x)} \right] \\
=& \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z \mid x)] \\
=& \mathbb{E}_q [\log q(z)] - \mathbb{E}_q [\log p(z, x)] + \log p(x) \\
=& - (\mathbb{E}_q [\log p(z, x)] - \mathbb{E}_q [\log q(z)]) + \log p(x)
\end{aligned}$$

- Negative of ELBO (plus constant); minimizing KL divergence is the same as maximizing ELBO

## Factorized distributions

$$\mu = E(X) \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\sigma^2 = E(X - E(X))^2 or E(X^2) - \mu^2$$

==> approximate by

`$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - [\frac{\sum_{i=1}^{n} x_i}{n}]^2$$

$`

> https://blog.csdn.net/KYJL888/article/details/85060883