法律声明

- 本课件包括演示文稿、示例、代码、题库、视频和声音等内容，小象学院和主讲老师拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意及内容，我们保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象
  - 新浪微博：ChinaHadoop

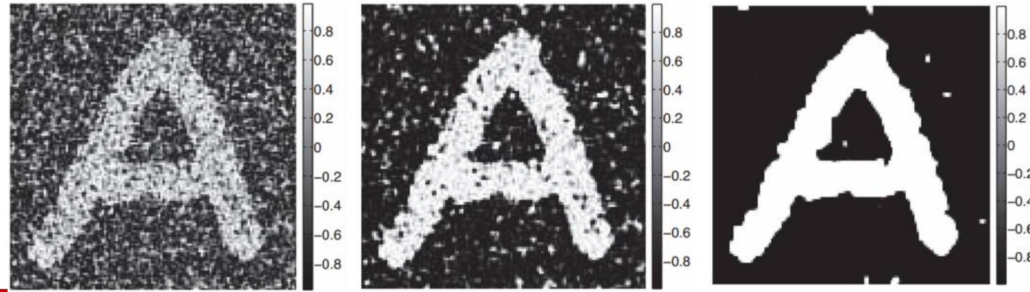
变分

小象学院 ChinaHadoop.cn 邹博

主要内容

- 理解变分的算法框架
  - 和采样算法的区别和联系
- 使用变分做隐变量的估计
  - 将去噪后的图像当做“隐变量”
- 使用变分做未知参数的估计
  - 取系统的参数为某概率分布

变分的核心



$$\log q_j(x_j) = E_{-q_j}\left[\log \tilde{p}(x)\right] + const$$

- 当更新q_j时，仅需要计算与x_j有公共边的那些变量即可——j的Markov毯包含的那些结点。
- Gibbs采样和变分：
  - Gibbs采样：使用邻居结点(相同文档的词)的主题采样值
  - 变分：采用相邻结点的期望。
  - 这使得变分往往比采样算法更高效：用一次期望计算代替了大量的采样。直观上，均值的信息是高密(dense)的，而采样值的信息是稀疏(sparse)的。

思考：MC-EM的启示

- 采样算法改造EM算法：Monte Carlo EM

$$Q(\theta,\bar{\theta})=\int p(Z\mid X,\bar{\theta})\ln p(Z,X\mid\theta)dZ$$

$$Q(\theta,\bar{\theta})\approx\frac{1}{L}\sum_{i=1}^{L}\ln p\left(Z^{(l)},X\mid\theta\right)$$

- 可否用计算统计量的方式改造采样？

$$p(z_i=k\mid\vec{z}_{\neg i},\vec{w})\propto\frac{n_{k,\neg i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,\neg i}^{(t)}+\beta_t}\left(n_{m,\neg i}^{(k)}+\alpha_k\right)$$

参数估计总结

- 给定样本$X_1, X_2 \dots X_n$，求系统参数θ
  - 最大似然估计：Maximum Likelihood Estimate

$$P(\theta \mid X) = \prod_i P(\theta \mid x_i)$$

  - 最大后验估计：Maximum A Posteriori

$$P(\theta \mid X) = \frac{P(X \mid \theta) P(\theta)}{P(X)} \propto P(X \mid \theta) P(\theta) = \prod_i P(x_i \mid \theta) P(\theta)$$

- 若存在隐变量：
  - EM算法——衍生品：随机EM、MAP-EM、IP算法
    - GMM、 pLSA、 HMM、CRF
  - 采样：MCMC、Gibbs

复习：线性回归的惩罚因子

- 线性回归的目标函数为：

$$J(\vec{\theta})=\frac{1}{2}\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right)-y^{(i)}\right)^{2}$$

- 将目标函数增加平方和损失：

$$J(\vec{\theta})=\frac{1}{2}\sum_{i=1}^{m}\left(h_{\vec{\theta}}\left(x^{(i)}\right)-y^{(i)}\right)^{2}+\lambda\sum_{j=1}^{n}\theta_j^2$$

- 本质即为假定参数θ服从高斯分布。

近似估计

- 变分推导(variational inference)是一般的确定性的近似推导算法。
- 基本思想：选择一个容易计算的近似分布q(x)，它能够尽可能的接近真正的后验分布p(x|D)。
  - 通过降低约束条件，在精度和速度上折中。
- 问题：如何定义两个分布的相似度？

变分的提法

- 假定p*(x)是真实(难解的)分布，q(x)是某个近似的(容易的)分布——如多元高斯分布或者多个简单分布的乘积。
- 假定q(x)有若干自由参数需要估计，我们需要优化这些未知参数使得q近似于p*。
- 一个显然的损失函数是最小化KL散度

$$KL\left(p^{*}\parallel q\right)=\sum_{x}p^{*}(x)\log\frac{p^{*}(x)}{q(x)}=E_{p^{*}(x)}\left(\log\frac{p^{*}(x)}{q(x)}\right)$$

变分法目标函数分析

$$KL(p^* \| q) = \sum_x p^*(x) \log \frac{p^*(x)}{q(x)} = E_{p^*(x)}\left(\log \frac{p^*(x)}{q(x)}\right)$$

- 上式关于后验概率p*的期望是不容易计算的，作为替代，将上述KL散度变成“逆KL散度”(reverse KL divergence)

$$KL(q \| p^*) = \sum_x q(x) \log \frac{q(x)}{p^*(x)} = E_{q(x)}\left(\log \frac{q(x)}{p^*(x)}\right)$$

- 第二个式子转换为计算关于q的期望(而q是关于未知参数的简单分布)，进一步，由于p(D)是归一化因子

$$p^*(x) = p(x \mid D) = \frac{p(x,D)}{p(D)} \overset{\Delta}{=\!=} \frac{\tilde{p}(x)}{Z} \Rightarrow \tilde{p}(x) = Z \cdot p^*(x)$$

- 上式变成：$J(q) = KL(q \| \tilde{p}) = \sum_x q(x) \log \frac{q(x)}{\tilde{p}(x)}$
- 使用逆KL散度的另一个好处是可以确保得到局部极值。

两个KL散度的区别

- KL(q||p)，又称为I-投影，信息投影(information projection)

$$KL(q\|p)=\sum_x q(x)\log\frac{q(x)}{p(x)}=E_{q(x)}\left(\log\frac{q(x)}{p(x)}\right)$$

  - 如果p(x)=0，q(x)>0，则KL为无穷大。因此，当p(x)=0时必须保证q(x)=0。即：该公式是对待求分布q"0强制"(zero forcing)的。从而，q往往被低估。

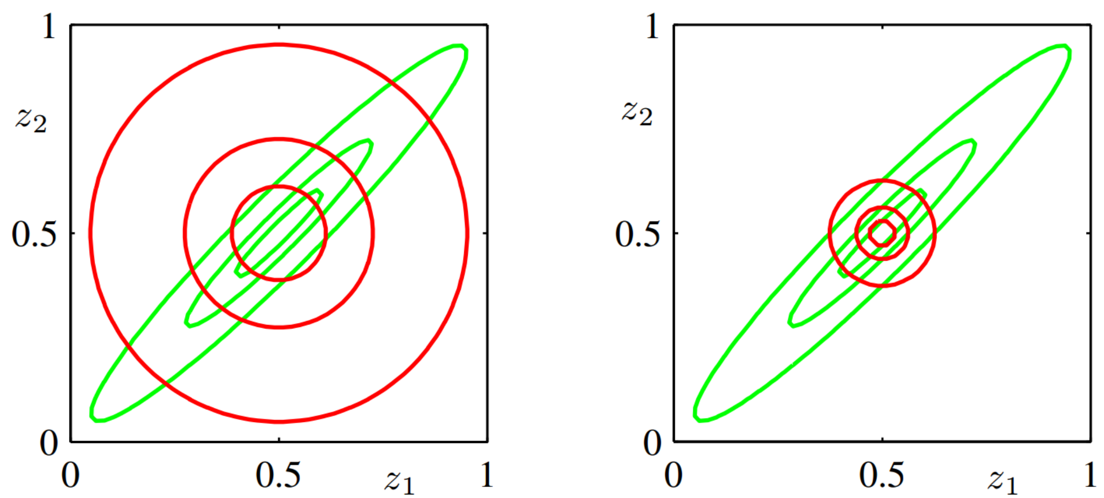
- KL(p||q)，又称为M-投影，矩投影(moment projection)

$$KL(p\|q)=\sum_x p(x)\log\frac{p(x)}{q(x)}=E_{p(x)}\left(\log\frac{p(x)}{q(x)}\right)$$

  - 如果p(x)>0，q(x)=0，则KL为无穷大。因此，当p(x)>0时必须保证q(x)>0。即：该公式是对待求分布q"0避免"(zero avoiding)的。从而，q往往被高估。

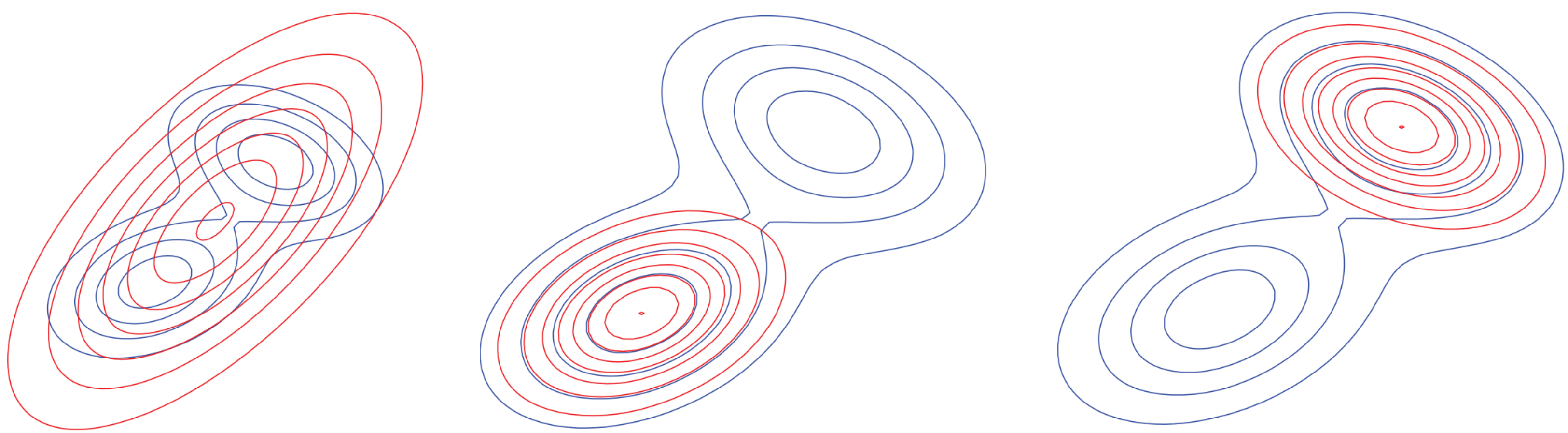
两个KL散度的区别

- 绿色曲线是真实分布p的等高线；红色曲线是使用近似$p(z_1,z_2)=p(z_1)p(z_2)$得到的等高线
  - 左：KL(p||q)：zero avoiding
  - 右：KL(q||p)：zero forcing



两个KL散度的区别

- 蓝色曲线是真实分布p的等高线；红色曲线是单模型近似分布q的等高线。
  - 左：KL(p||q)：q趋向于覆盖p
  - 中、右：KL(q||p)：q能够锁定某一个峰值



两个KL散度之间的联系

- 给定分布p和q的距离定义

$$D_\alpha(p\|q)=\frac{2}{1-\alpha^2}\left(1-\int p(x)^{\frac{1+\alpha}{2}}q(x)^{\frac{1-\alpha}{2}}dx\right)$$

- p和q的KL散度

$$KL(p\|q)=\int p(x)\log\frac{p(x)}{q(x)}dx=-\int p(x)\log\frac{q(x)}{p(x)}dx$$

- 变换：

$$-\int p(x)^{\frac{1+\alpha}{2}}q(x)^{\frac{1-\alpha}{2}}dx=-\int p(x)^{1+\frac{\alpha-1}{2}}q(x)^{\frac{1-\alpha}{2}}dx$$

$$=-\int p(x)p(x)^{\frac{\alpha-1}{2}}q(x)^{\frac{1-\alpha}{2}}dx=-\int p(x)\left(\frac{q(x)}{p(x)}\right)^{\frac{1-\alpha}{2}}dx$$

两个KL散度之间的联系

$$u=\frac{q(x)}{p(x)} \Rightarrow \begin{cases} f(u)=u^{\frac{1-\alpha}{2}} \\ g(u)=\log u \end{cases} \Rightarrow \begin{cases} f'(u)=\frac{1-\alpha}{2}u^{-\frac{1+\alpha}{2}} \\ g'(u)=u^{-1} \end{cases} \Rightarrow \begin{cases} \frac{1-\alpha}{2}=1 \\ -\frac{1+\alpha}{2}=-1 \end{cases} \Rightarrow \begin{cases} \alpha=-1 \\ \alpha=1 \end{cases}$$

- 当α=1时退化为KL(q||p)
- 当α=-1时退化为KL(q||p)
- 当α=0时？

Hellinger distance

$$D_\alpha(p\|q)=\frac{2}{1-\alpha^2}\left(1-\int p(x)^{\frac{1+\alpha}{2}}q(x)^{\frac{1-\alpha}{2}}dx\right)$$

$$\Rightarrow D_H(p\|q)=2\left(1-\int\sqrt{p(x)q(x)}dx\right)=2-2\int\sqrt{p(x)q(x)}dx$$

$$=\int p(x)dx+\int q(x)dx-\int 2\sqrt{p(x)q(x)}dx$$

$$=\int\left(p(x)-2\sqrt{p(x)q(x)}+q(x)\right)dx$$

$$=\int\left(\sqrt{p(x)}-\sqrt{q(x)}\right)^2dx$$

- 该距离满足三角不等式，是对称、非负距离

新目标函数的可行性 $J(q)=KL(q\parallel\tilde{p})$

$$
\begin{aligned}
J(q)=KL(q\parallel\tilde{p})&=\sum_x q(x)log\frac{q(x)}{\tilde{p}(x)}\\
&=\sum_x q(x)log\frac{q(x)}{Z\cdot p^*(x)}\\
&=\sum_x q(x)log\frac{q(x)}{p^*(x)}+\sum_x q(x)log\frac{1}{Z}\\
&=\sum_x q(x)log\frac{q(x)}{p^*(x)}-logZ\\
&=KL(q\parallel p^*(x))-logZ
\end{aligned}
$$

- 由于Z是常数，通过最小化J(q)，能够使得q接近p*。

变分和EM的联系

- 因为KL散度总是非负的，J(p)是NLL的上界
  - negative log likelihood

$$J(q) = KL(q \| p^*) - \log Z \geq -\log Z = -\log p(D)$$

  - 进一步：

$$L(q) \stackrel{\Delta}{=\!=} -J(q) = -KL(q \| p^*) + \log Z \leq \log Z = \log p(D)$$

- 因此，L(q)是似然函数的下界，当q=p*时取等号。
  - 可取等号，说明下界是紧的(tight)
- EM和变分
  - EM算法：计算关于隐变量后验概率的期望，得到下界；
  - 变分：计算KL散度，得到下界；
  - 相同的思维：不断迭代，得到更好的下界。
  - 不断上升。

思考：目标函数的物理含义

- 定义能量 $E(x) = -\log \tilde{p}(x)$
- 目标函数是能量的期望减去系统的熵。J(q)被叫做"变分自由能"或"Helmholtz free energy"。

$$J(q) = KL(q \| \tilde{p}) = \sum_x q(x) \log \frac{q(x)}{\tilde{p}(x)}$$

$$= E_q\left(\log \frac{q(x)}{\tilde{p}(x)}\right) = E_q(\log q(x) - \log \tilde{p}(x))$$

$$= E_q(\log q(x)) + E_q(-\log \tilde{p}(x))$$

$$\stackrel{\triangle}{==} -H(X) + E_q(E(x))$$

思考：似然函数期望与目标函数

- 负似然函数NLL的期望，加上一个惩罚项——近似分布与先验分布的KL距离。

$$J(q)=KL(q\|\tilde{p})=\sum_x q(x)\log\frac{q(x)}{\tilde{p}(x)}=E_q\left(\log\frac{q(x)}{\tilde{p}(x)}\right)$$

$$=E_q\left(\log\frac{q(x)}{p(x,D)}\right)=E_q\left(\log\frac{q(x)}{p(x)p(D|x)}\right)$$

$$=E_q\left(\log\frac{1}{p(D|x)}+\log\frac{q(x)}{p(x)}\right)=E_q\left(\log\frac{1}{p(D|x)}\right)+E_q\left(\log\frac{q(x)}{p(x)}\right)$$

$$=E_q\left(-\log p(D|x)\right)+KL(q\|p)$$

平均场方法(Mean field method)

- 最流行的变分方法之一是平均场近似。在这种方法中，假定后验概率能够近似分解为若干因子的乘积。
  - 思考：无向图中的“最大团” Hammersley-Clifford定理

$$q(x)=\prod_i q_i(x_i)$$

- 我们的目标是解决最优化问题：$\min_{q_1,\cdots q_D} KL(q\parallel p)$
- 平均场方法使得可以在若干边界分布q_i上进行(依次)优化。事实上，很快将得知，有如下近似等式：

$$\log q_j(x_j)=E_{-q_j}[\log \tilde{p}(x)]+const$$

- 其中，未正则化的后验概率 $\tilde{p}(x)=p(x,D)$
- 关于除了x_j的所有其他变量的f(x)的期望 $E_{-q_j}[f(x)]$

平均场方法(Mean field method)

$$\log q_j(x_j) = E_{-q_j}\left[\log \tilde{p}(x)\right] + const$$

- 当更新q_j时，仅需要计算与x_j有公共边的那些变量即可——j的Markov毯包含的那些结点。因为该方法使用相邻结点的期望(均值)，所以称作平均场。
- Gibbs采样和变分：
  - Gibbs采样：使用邻居结点(相同文档的词)的主题采样值
  - 变分：采用相邻结点的期望。
  - 这使得变分往往比采样算法更高效：用一次期望计算代替了大量的采样。直观上，均值的信息是高密(dense)的，而采样值的信息是稀疏(sparse)的。

变分推导/似然下界L $J(q)=KL(q\parallel\tilde{p})=\sum_x q(x)log\frac{q(x)}{\tilde{p}(x)}$

$$
\begin{aligned}
L(q_j) &\stackrel{\Delta}{=\!=} -J(q_j)=-\sum_x q(x)log\frac{q(x)}{\tilde{p}(x)}\\
&=\sum_x q(x)[\log\tilde{p}(x)-\log q(x)]\\
&=\sum_x\prod_i q_i(x_i)\left[\log\tilde{p}(x)-\log\prod_i q_i(x_i)\right]\\
&=\sum_{x_j}\sum_{x_{-j}}q_j(x_j)\prod_{i\neq j}q_i(x_i)\left[\log\tilde{p}(x)-\sum_k\log q_k(x_k)\right]\\
&=\sum_{x_j}q_j(x_j)\sum_{x_{-j}}\prod_{i\neq j}q_i(x_i)\left[\log\tilde{p}(x)-\left(\log q_j(x_j)+\sum_{k\neq j}\log q_k(x_k)\right)\right]\\
&=\left(\sum_{x_j}q_j(x_j)\sum_{x_{-j}}\prod_{i\neq j}q_i(x_i)\log\tilde{p}(x)\right)-\left(\sum_{x_j}q_j(x_j)\log q_j(x_j)\right)+const\\
&=\left(\sum_{x_j}q_j(x_j)\log f_j(x_j)\right)-\left(\sum_{x_j}q_j(x_j)\log q_j(x_j)\right)+const=-KL(q_j\parallel f_j)
\end{aligned}
$$

$$\boxed{\log f_j(x_j)\stackrel{\triangle}{=\!=}\sum_{x_{-j}}\prod_{i\neq j}q_i(x_i)\log\tilde{p}(x)=E_{-q_j}[\log\tilde{p}(x)]}$$

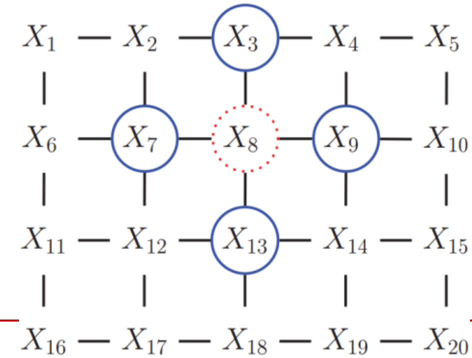
变分推导最终结论

- 下界 $L(q_j) = -KL(q_j \| f_j)$ 取极大，则 $KL(q_j \| f_j)$ 取极小，此刻，要求二者分布相同。
- 由于 $\log f_j(x_j) \stackrel{\triangle}{=} \sum_{x_{-j}} \prod_{i \neq j} q_i(x_i) \log \tilde{p}(x) = E_{-q_j}[\log \tilde{p}(x)]$
- 所以

$$q_j(x_j) = f_j(x_j) = \frac{1}{Z_j} \exp\left(E_{-q_j}[\log \tilde{p}(x)]\right)$$

- 忽略归一化因子

$$\log q_j(x_j) = E_{-q_j}[\log \tilde{p}(x)] + const$$



Ising model

- Ising模型是统计物理提出的马尔科夫随机场MRF，它最初是用来对磁化行为建模。令 $y_s \in \{-1,+1\}$ 表示原子的自旋，它的旋转角速度方向要么朝上，要么朝下。在某些环境下，表现为铁磁现象(ferro-magnets)：相邻结点的自旋方向趋近于同向；而其他环境中表现为反铁磁现象(anti-ferromagnets)，相邻结点的自旋方向趋近于相反。
- 可以使用MRF建模：连接相邻变量，然后定义团(clique)之间的势函数：

$$\varphi_{st}(y_s, y_t) = \begin{pmatrix} e^{w_{st}} & e^{-w_{st}} \\ e^{-w_{st}} & e^{w_{st}} \end{pmatrix}$$

势函数的系数 $\varphi_{st}(y_s, y_t) = \begin{pmatrix} e^{w_{st}} & e^{-w_{st}} \\ e^{-w_{st}} & e^{w_{st}} \end{pmatrix}$

- w_{st}是结点s和t之间的耦合强度(coupling strength)。如果两个结点没有连接，则设置$w_{st}=0$。假定权值矩阵W是对称阵，即$w_{st}=w_{ts}$；进一步假定所有的边有相同的强度，即$w_{st}=J\neq 0$。
- 如果所有的权值都为正(J>0)，则相邻结点的自旋趋向于同向，能够对铁磁现象建模；如果权值足够强，则结点的概率分布将只有两种状态：一部分结点是1状态，一部分结点是-1状态，这被称作系统的基态(groud states)。
  - 类比：将某状态认为是实际观测的图像，基态认为是去噪后的“干净”的图像。
- 同理，如果J<0可以对反铁磁现象建模。

使用变分做图像去噪

- 考虑图像的去噪问题：$x_i \in \{-1,+1\}$ 是隐藏在观测图像背后的“干净”图像的像素取值。为简洁方便，假定是二值图。X的联合分布假定具有如下先验形式：

$$p(x) = \frac{1}{Z_0}\exp\left(-E_0(x)\right), \qquad \text{其中}, E_0(x) = -\sum_{i=1}^{D}\sum_{j\in nbr_i} W_{ij} x_i x_j$$

- 似然函数 $p(y|x) = \prod_i p(y_i|x_i) = \prod_i \left(\exp\left(\ln p(y_i|x_i)\right)\right)$

$$= \exp\sum_i \ln p(y_i|x_i) = \exp\sum_i \left(L_i(x_i)\right)$$

后验概率

- 后验概率 $p(x\mid y)=\dfrac{p(y\mid x)p(x)}{p(y)}\propto p(y\mid x)p(x)$

$$\propto\left(\exp\sum_i\left(L_i(x_i)\right)\right)\left(\exp\left(-E_0(x)\right)\right)$$

$$=\exp\left(-E_0(x)+\sum_i L_i(x_i)\right)$$

$$\Rightarrow p(x\mid y)=\frac{1}{Z}\exp\left(-E(x)\right)$$

- 其中，$E(x)=E_0(x)-\sum_i L_i(x_i)$

近似概率

- 根据后验概率 $p(x|y)=\frac{1}{Z}\exp(-E(x))=\frac{1}{Z}\exp\left(-E_0(x)+\sum_i L_i(x_i)\right)$
- 得到经验概率的对数：

$$\ln\tilde{p}(x)=-E_0(x)+\sum_i L_i(x_i)=\sum_{i=1}^{D}\sum_{j\in nbr_i}W_{ij}x_ix_j+\sum_i L_i(x_i)$$

- 只考虑与i相关的部分 $\ln\tilde{p}(x)=x_i\sum_{j\in nbr_i}W_{ij}x_j+L_i(x_i)+const$
- 即结点i的均值为μ_i，利用变分结论

$$\log q_i(x_i)\propto\sum_{x_{-i}}\prod_{j\neq i}q_j(x_j)\log\tilde{p}(x)=E_{-q_i}[\log\tilde{p}(x)]$$

- 得：

$$q_i(x_i)\propto\sum_{x_{-i}}\prod_{j\neq i}q_j(x_j)\cdot\left(x_i\sum_{j\in nbr_i}W_{ij}x_j+L_i(x_i)\right)=\exp\left(x_i\sum_{j\in nbr_i}W_{ij}\mu_j+L_i(x_i)\right)$$

根据公式：

- 记平均场对结点i的影响为：$m_i = \sum_{j \in nbr_i} W_{ij}\mu_j$

$$q_i(x_i) \propto \exp\left(x_i \sum_{j \in nbr_i} W_{ij}\mu_j + L_i(x_i) \right) = \exp\left(x_i \cdot m_i + L_i(x_i)\right)$$

- 进一步，记：$L_i^+ = L_i(+1),\ L_i^- = L_i(-1)$
- 则近似边缘后验概率为：

$$\begin{cases} q_i(x_i = 1) = \dfrac{e^{m_i + L_i^+}}{e^{m_i + L_i^+} + e^{-m_i + L_i^-}} = \dfrac{1}{1 + e^{-2m_i + L_i^- - L_i^+}} = sigmiod(2a_i) \\ q_i(x_i = -1) = sigm(-2a_i) \end{cases}$$

其中，$a_i = m_i + \dfrac{L_i^+ - L_i^-}{2}$

更新方程

- 结点i迭代后的期望为：

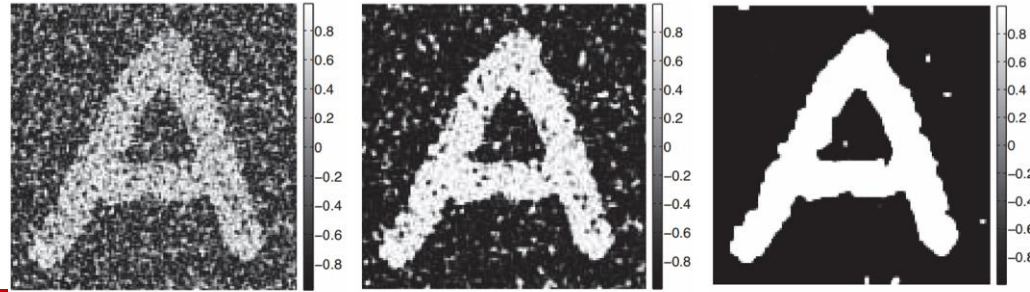
$$\mu_i = E_{q_i}(x_i) = q_i(x_i=+1)\cdot(+1) + q_i(x_i=-1)\cdot(-1)$$

$$= \frac{1}{1+e^{-2a_i}} - \frac{1}{1+e^{2a_i}} = \frac{e^{a_i}}{e^{a_i}+e^{-a_i}} - \frac{e^{-a_i}}{e^{-a_i}+e^{a_i}} = \tanh(a_i)$$

- 因此，更新方程为：

$$\mu_i = \tanh\left(\sum_{j\in nbr_i} W_{ij}\mu_j + \frac{L_i^+ - L_i^-}{2}\right)$$

迭代方程



- 根据上式很容易得到迭代公式：

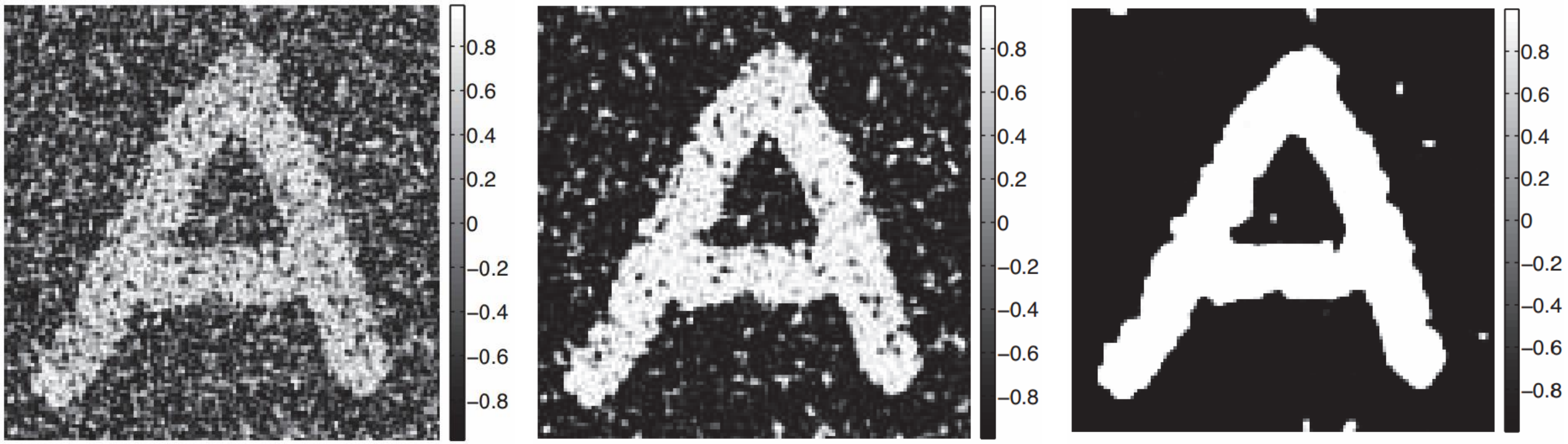
$$\mu_i^{(t)} = \tanh\left(\sum_{j\in nbr_i} W_{ij}\mu_i^{(t-1)} + \frac{L_i^+ - L_i^-}{2}\right)$$

- 实践中，往往需要增加衰减因子，得
  - damped updates：1>λ>0

$$\mu_i^{(t)} = (1-\lambda)\mu_i^{(t-1)} + \lambda\tanh\left(\sum_{j\in nbr_i} W_{ij}\mu_i^{(t-1)} + \frac{L_i^+ - L_i^-}{2}\right)$$

实际效果

- 2维Ising模型：（混入 $\sigma=2$ 的高斯噪声）
  - 先验权值 $W_{ij}=1$，衰减因子 $\lambda=0.5$
  - 左：迭代1次；中：迭代3次；右：迭代15次。



变分贝叶斯Variational Bayesian

- 上述变分实践是给定模型参数推断隐变量。此外，变分方法也可以推断参数本身。使用平均场方法，将后验概率写成参数各自分布的乘积，即得到变分贝叶斯方法(Variational Bayesian, VB)。
- 变分贝叶斯：$p(\theta \mid D) \approx \prod_k q_k(\theta_k)$

高斯分布的变分贝叶斯

- 使用变分贝叶斯推断一维高斯分布$p(\mu,\lambda^{-1}|D)$后验概率的参数。其中，λ为精度(方差的倒数)。为计算方便，使用共轭先验的形式。

$$p(\mu,\lambda)=N\left(\mu|\mu_0,(\kappa_0\lambda)^{-1}\right)\cdot Ga(\lambda|a_0,b_0)$$

  - 该形式可看成混合高斯GMM——思考EM方案.

- 近似分解得到如下形式：

$$q(\mu,\lambda)=q_\mu(\mu)q_\lambda(\lambda)$$

未正则化的对数后验

$$
\begin{cases}
p(D\mid\mu,\lambda)=\sqrt{\dfrac{\lambda}{2\pi}}\cdot e^{-\frac{\lambda(x-\mu)^2}{2}}\\
p\left(\mu\mid\mu_0,(\kappa_0\lambda)^{-1}\right)=\sqrt{\dfrac{\kappa_0\lambda}{2\pi}}\cdot e^{-\frac{\kappa_0\lambda(\mu-\mu_0)^2}{2}}\\
p(\lambda\mid a_0,b_0)=\dfrac{\beta^{a_0}}{\Gamma(a_0)}\cdot\lambda^{a_0-1}e^{-b_0\lambda}
\end{cases}
$$

- 目标函数

$$
\begin{aligned}
&\log\tilde{p}(\mu,\lambda)\\
&=\log p(\mu,\lambda,D)\\
&=\log p(D\mid\mu,\lambda)+\log p(\mu\mid\lambda)+\log p(\lambda)\\
&=\log\prod_{i=1}^{N}\sqrt{\frac{\lambda}{2\pi}}\cdot e^{-\frac{\lambda(x_i-\mu)^2}{2}}+\log\sqrt{\frac{\kappa_0\lambda}{2\pi}}\cdot e^{-\frac{\kappa_0\lambda(\mu-\mu_0)^2}{2}}+\log\frac{\beta^{a_0}\lambda^{a_0-1}e^{-b_0\lambda}}{\Gamma(a_0)}\\
&=\frac{N}{2}\log\lambda-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2+\frac{1}{2}\log(\kappa_0\lambda)-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2\\
&+(a_0-1)\log\lambda-b_0\lambda+const
\end{aligned}
$$

更新 $q_\mu(\mu)$

- 最优形式的 $q_\mu(\mu)$ 是通过计算关于λ的平均值获得的：

$$\log\tilde{p}(\mu,\lambda)=\frac{N}{2}\log\lambda-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2+\frac{1}{2}\log(\kappa_0\lambda)-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2+(a_0-1)\log\lambda-b_0\lambda+const$$

$$\begin{aligned}\log q_\mu(\mu)&=E_{q_\lambda}\left(\log\tilde{p}(\mu,\lambda)\right)\\&=E_{q_\lambda}\left(-\frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2-\frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2\right)+const\\&=-\frac{E_{q_\lambda}(\lambda)}{2}\left(\sum_{i=1}^{N}(x_i-\mu)^2+\kappa_0(\mu-\mu_0)^2\right)+const\end{aligned}$$

由$q_\mu(\mu)$得到的参数等式

- 对比正态分布的对数形式：

$$\ln p(x) = \ln\frac{1}{\sqrt{2\pi}} - \ln\sigma - \frac{(x-\mu)^2}{2\sigma^2}$$

$$\log q_\mu(\mu) = -\frac{E_{q_\lambda}(\lambda)}{2}\left(\sum_{i=1}^{N}(x_i-\mu)^2 + \kappa_0(\mu-\mu_0)^2\right) + const$$

- 得，

$$\begin{cases} \mu_N = \dfrac{\kappa_0\mu_0 + N\overline{x}}{\kappa_0 + N} \\ \kappa_N = (\kappa_0 + N)E_{q_\lambda}(\lambda) \end{cases} \qquad \begin{cases} E_{q(\mu)}(\mu) = \mu_N \\ E_{q(\mu)}(\mu^2) = \dfrac{1}{\kappa_N} + \mu_N^2 \end{cases}$$

  - 目前尚未知$q_\lambda(\lambda)$，因此无法计算$E_{q_\lambda}(\lambda)$，继续考察$q_\lambda(\lambda)$。

更新 $q_\lambda(\lambda)$

- 最优形式的$q_\lambda(\lambda)$是通过计算关于μ的平均值获得的：

$$\log \tilde{p}(\mu,\lambda) = \frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\log(\kappa_0\lambda) - \frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2 + (a_0-1)\log\lambda - b_0\lambda + const$$

$$\log q_\lambda(\lambda) = E_{q_\mu}\left(\log \tilde{p}(\mu,\lambda)\right)$$

$$= E_{q_u}\left(\frac{N}{2}\log\lambda - \frac{\lambda}{2}\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{1}{2}\log(\kappa_0\lambda) - \frac{\kappa_0\lambda}{2}(\mu-\mu_0)^2 + (a_0-1)\log\lambda - b_0\lambda\right) + const$$

$$= \frac{N}{2}\log\lambda + \frac{1}{2}\log\lambda + (a_0-1)\log\lambda - b_0\lambda - \frac{\lambda}{2}E_{q_u}\left(\kappa_0(\mu-\mu_0)^2 + \sum_{i=1}^{N}(x_i-\mu)^2\right) + const$$

由$q_\lambda(\lambda)$得到的参数等式

- 对比正态分布的对数形式

$$\ln p(x;\alpha,\beta)=\alpha\ln\beta+(\alpha-1)\ln x-\beta x-\ln\Gamma(\alpha)$$

$$\log q_\lambda(\lambda)=\frac{N}{2}\log\lambda+\frac{1}{2}\log\lambda+(a_0-1)\log\lambda-b_0\lambda-\frac{\lambda}{2}E_{q_u}\left(\kappa_0(\mu-\mu_0)^2+\sum_{i=1}^{N}(x_i-\mu)^2\right)+const$$

- 得，

$$\begin{cases} a_N=a_0+\dfrac{N+1}{2} \\ b_N=b_0+\dfrac{1}{2}E_{q_u}\left(\kappa_0(\mu-\mu_0)^2+\displaystyle\sum_{i=1}^{N}(x_i-\mu)^2\right) \\ =b_0+\kappa_0\left(E(\mu^2)+\mu_0^2-2E(\mu)\mu_0\right)+\dfrac{1}{2}\displaystyle\sum_{i=1}^{N}\left(x_i^2+E(\mu^2)-2E(\mu)x_i\right) \end{cases}$$

- 同时：$E_{q(\lambda)}(\lambda)=\dfrac{a_N}{b_N}$

迭代公式

- 根据

$$\begin{cases} \mu_N = \dfrac{\kappa_0\mu_0 + N\cdot\bar{x}}{\kappa_0 + N} \\ \kappa_N = (\kappa_0 + N)E_{q_\lambda}(\lambda) \end{cases}$$

$$\begin{cases} a_N = a_0 + \dfrac{N+1}{2} \\ b_N = b_0 + \kappa_0\left(E(\mu^2) + \mu_0^2 - 2E(\mu)\mu_0\right) + \dfrac{1}{2}\sum_{i=1}^{N}\left(x_i^2 + E(\mu^2) - 2E(\mu)x_i\right) \end{cases}$$

- 以及

$$\begin{cases} E_{q(\mu)}(\mu) = \mu_N \\ E_{q(\mu)}(\mu^2) = \dfrac{1}{\kappa_N} + \mu_N^2 \end{cases}$$

$$E_{q(\lambda)}(\lambda) = \frac{a_N}{b_N}$$

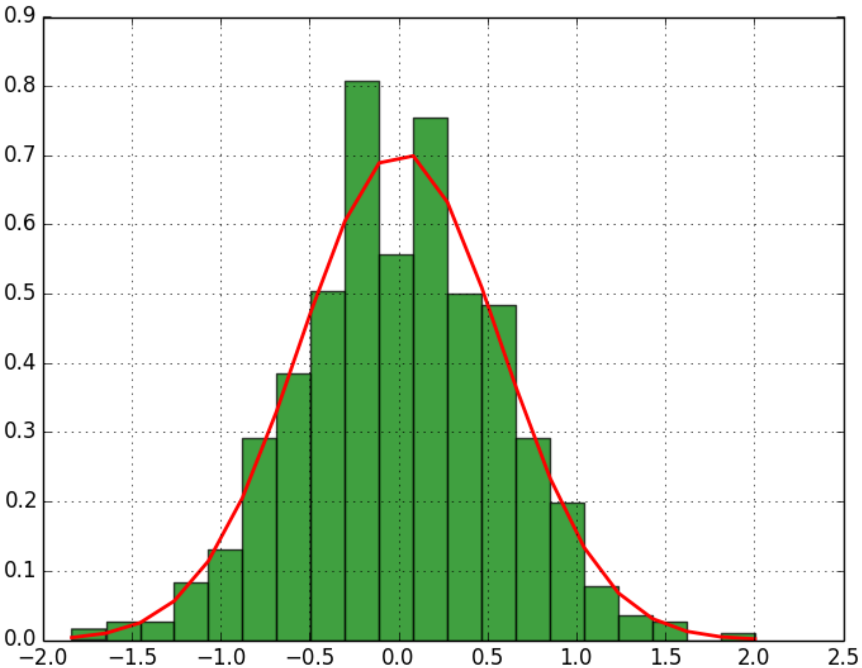
- 得：

$$\begin{cases} \mu_N = \dfrac{\kappa_0\mu_0 + N\cdot\bar{x}}{\kappa_0 + N} \\ \kappa_N = (\kappa_0 + N)\dfrac{a_N}{b_N} \end{cases}$$

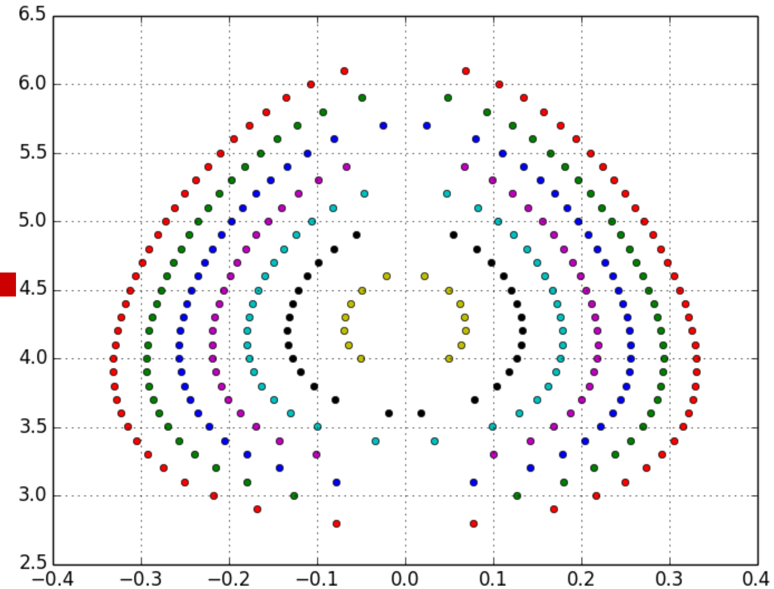
Code

```python
# 需要估算的四个参数
a, b = 9., 2.        # a-1幂参数，-b指数参数
mu, kappa = 0., 3.
# 显示等高线
show_parameter(mu, kappa, a, b)

# 生成随机数
s = np.zeros(1000)  # 样本
for i in range(len(s)):
    lamda = np.random.gamma(a, 1/b)
    u = np.random.normal(mu, math.sqrt(1/(kappa * lamda)))  # 样本高斯分布的均值
    s[i] = np.random.normal(u, math.sqrt(1/lamda))

# 显示样本的直方图
t1, bins, t2 = plt.hist(s, 20, normed=True, color='g', alpha=0.75)
plt.grid(True)
sigma = s.std()
y = np.exp(-(bins - s.mean()) ** 2 / (2 * sigma**2)) / (math.sqrt(2*math.pi)*sigma)
plt.plot(bins, y, 'r-', linewidth=2)
plt.show()

# 变分迭代
a, b, mu, kappa = 3, 1, 6, 2   # 随机初始化
kappa_n = kappa
mu_n = (kappa * mu + s.sum()) / (kappa + len(s))
a_n = a + (len(s) + 1) / 2
b_n = b
for t in range(10):
    e_u2 = mu_n ** 2 + 1/kappa_n    # E[u^2]
    b_n = b + kappa*(e_u2 + mu**2 - 2*mu_n*mu) + (sum(s*s) + e_u2*len(s) - 2*mu_n*s.sum())/2
    kappa_n = (kappa + len(s)) * a_n / b_n
show_parameter(mu_n/b_n, kappa_n/b_n, a_n/b_n, 1)
```

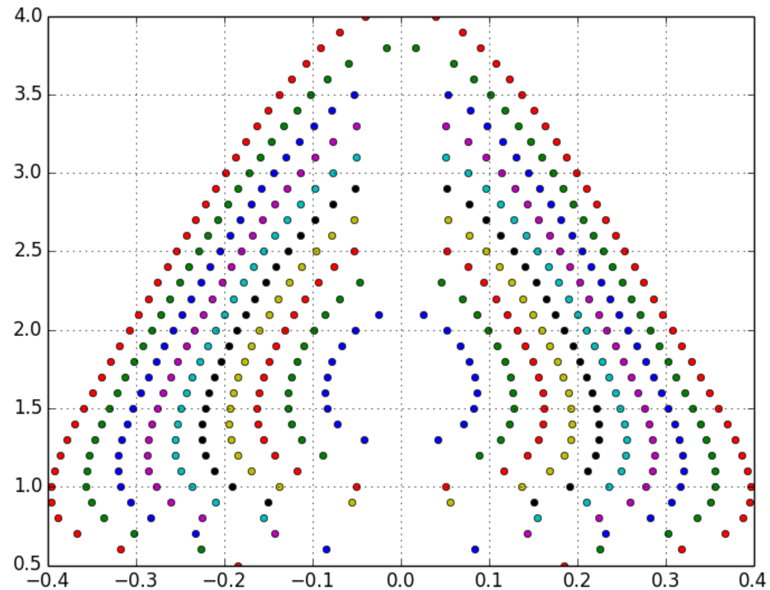
变分贝叶斯实验结果



0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5



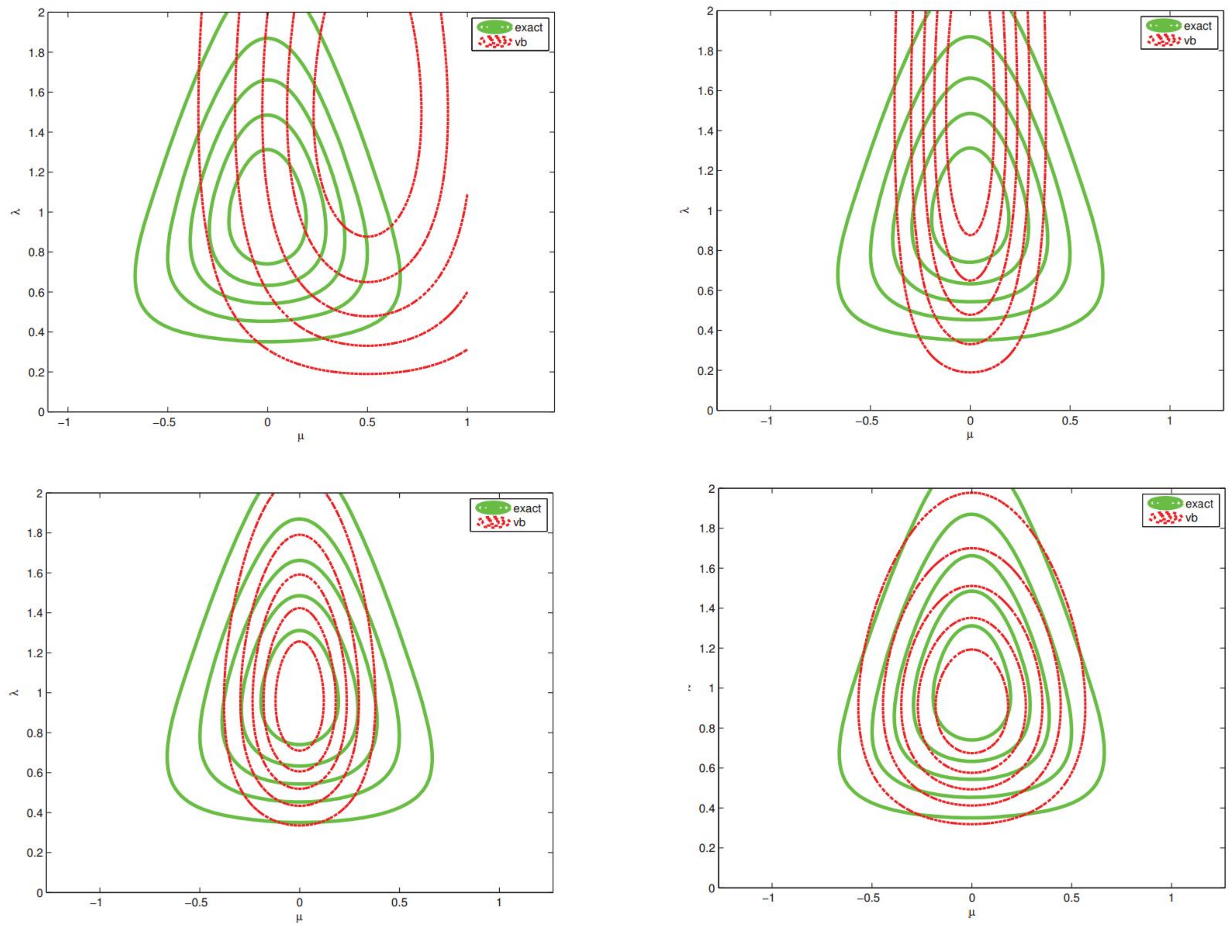
6.5
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4



4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
−0.4 −0.3 −0.2 −0.1 0.0 0.1 0.2 0.3 0.4

变分参数估计实例

初始状态	更新$q_\mu(\mu)$
更新$q_\lambda(\lambda)$	5次迭代后



变分总结

- 变分既能够推断隐变量，也能推断未知参数，是非常有力的参数学习工具。其难点在于公式演算略复杂，和采样相对：一个容易计算但速度慢，一个不容易计算但运行效率高。
- 平均场方法的变分推导，对离散和连续的隐变量都适用。在平均场方法的框架下，变分推导一次更新一个分布，其本质为坐标上升。可以使用模式搜索(pattern search)、基于参数的扩展(parameter expansion)等方案加速。
  - 思考：EM的梯度上升理解。
- 有时假定所有变量都独立不符合实际，可使用结构化平均场(structured mean field)，将变量分成若干组，每组之间独立。
- 变分除了能够和贝叶斯理论相配合得到VB，还能进一步与EM算法结合，得到VBEM，用于带隐变量和未知参数的推断。
  - 如GMM、LDA

参考文献

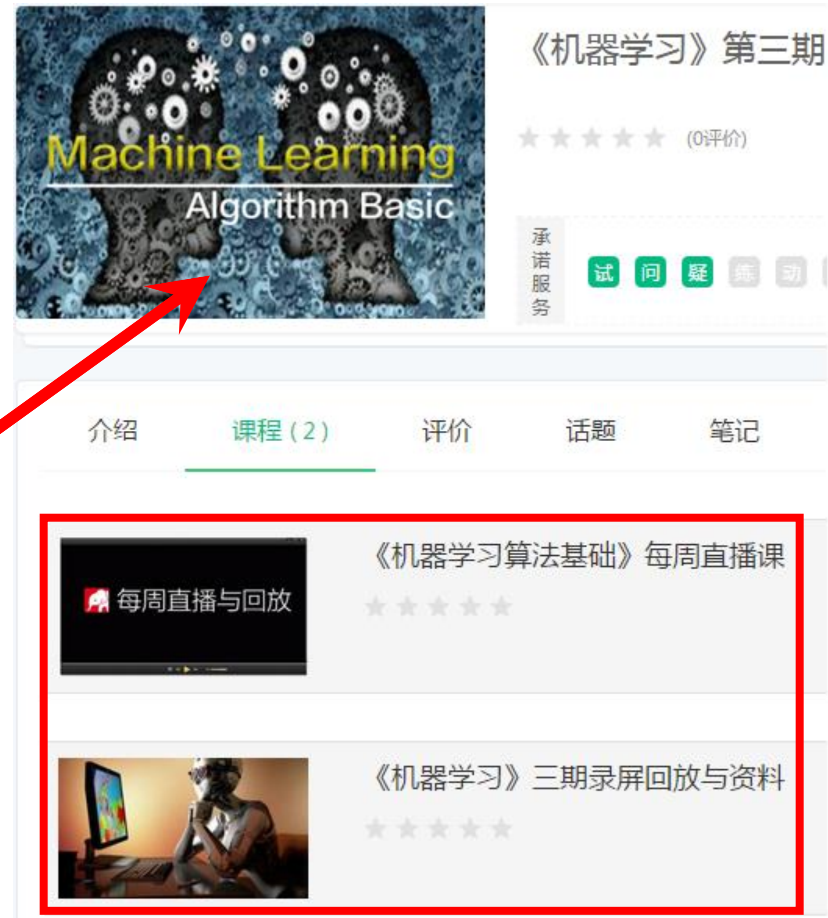
- Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective, Chapter 21*. MIT Press, 2012
- Christopher M. Bishop. *Pattern Recognition and Machine Learning Chapter 10*. Springer-Verlag, 2006

我们在这里

- http://wenda.ChinaHadoop.cn
  - 视频/课程/社区
- 微博
  - @ChinaHadoop
  - @邹博_机器学习
- 微信公众号
  - 小象
  - 大数据分析挖掘

课程资源

- 直播课的入口
- 录播视频和讲义资料



感谢大家！

恳请大家批评指正！