附录

(3.11) (3.12) (3.19) (3.20) (3.21)

赋予线性基函数的高斯白噪声, 其输出t的相关极大似然估计:

- 解:

3.1.4 - 最小化 q 的一般正则化目标函数(3.29)形式 $\Leftrightarrow$ 在 $\sum_{j=1}^{M}|w_j|^q \leqslant \eta$ 条件下 最小化未带正则项的平方误差函数(3.12).

3.1.5 - 从2.3.4节中知道多元高斯分布 均值的最大似然解与协方差无关.

- 解:

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}_n - \boldsymbol{\mu}\right)$$

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

3.2 - 同时关于权值 w 和正则化系数 λ 来最小化正则化的误差函数会导致不正确: $\lambda = 0$.

- 解:

  因为关于 λ 求导后为 $\frac{1}{2}\sum_{j=1}^{M}|w_j|^q$，恒正.

3.3.1 - 式(3.48) 推导, (3.10) 定义的似然函数 $p(t \mid w)$ 是w二次函数的指数形式, 所以对应的共轭先验是高斯分布.

- 解:

  忘了好多，还要再补.

$$\int h(x)\exp\left(T(x)^T\eta\right) = 1 \Rightarrow A(\eta) = \log\int h(x)\exp\{T(x)^T\eta\}$$

与x无关 积分时 直接拿到分母 $\leftarrow \exp(A(\eta))$

$A(\eta)$: log normaliser (唯一不含x)

八、指数族分布 (Exponential Family Distribution)

一 背景

统计量：任意形式的函数？ 比如样本均值

某一类分布

指数族分布

充分统计量

$p(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$ → 参数和x函数的线性组合

η 为参数向量，$x \in R^p$

★ $A(\eta)$: log partition function（配分函数）

eg $p(x|\theta) = \frac{1}{z}\hat{p}(x|\theta)$ z是归一化因子[redacted]，$z = \int_x \hat{p}(x|\theta)dx$

$\int_x p(x|\theta)dx = \frac{1}{z}\int \hat{p}(x|\theta)dx = 1$

$\Rightarrow$

- $p(x|\eta) = \frac{1}{\exp(A(\eta))} h(x)\exp(\eta^T\phi(x))$

$= \frac{1}{z}\hat{p}(x|\eta)$ $\Rightarrow \exp(A(\eta)) = z$

$\Rightarrow A(\eta) = \log z$

$\therefore A(\eta)$ 叫作 log partition function

$\phi(x)$: 充分统计量

以高斯分布为例，我们可以令

$$\phi(x) = \begin{pmatrix} \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i^2 \end{pmatrix},$$

"统计"：即表示对样本的统计值

"充分"：通过上面两个统计值，我们可以求得均值和方差，故而可求得高斯分布表达式。

这种性质对于 online learning 特别有用，不用再存储每一个样本了。来一个算一个

常见的指数族分布：

- Gaussian
- Bernoulli — 类别分布
- 二项分布 — 多项式分布
- 泊松分布
- Beta
- Drichlet
- Gamma

共轭先验

$$P(z|x) = \frac{P(x|z)P(z)}{\int_z P(x|z)P(z)dz}$$

积分难以计算 → 用近似推断 { 变分推断, MCMC } 算后验 $P(z|x)$

为了解决积分难计算的问题，一个思路是能否绕过积分呢？我们知道存在如下关系

$$\underbrace{P(z|x)}_{\text{后验分布}} \propto \underbrace{P(x|z)}_{\text{似然}}\underbrace{P(z)}_{\text{先验}}$$

在已知似然函数的情况下，选取什么样的先验分布能够使得后验分布与先验分布具有相同的数学形式呢?

如果存在这样的一个先验分布，那么上一时刻的输出可以作为下一时刻计算的先验分布，那么这样整个计算就可以形成闭环。也就是说如果后验分布和先验分布是同分布，此时我们称先验分布和后验分布是共轭分布，且称先验分布是似然函数的共轭先验。比如高斯分布家族在高斯似然函数下与其自身共轭，也叫自共轭。

共轭先验的好处主要在于代数上的方便性，可以直接给出后验分布的封闭形式，否则的话只能做数值计算。共轭先验也有助于活的关于似然函数如何更新先验分布的直观印象。

最大熵 (无信息先验)

无信息先验是指先验分布尽可能对对后验分布产生小的影响，即尽可能的随机性，其实就是最大熵原理。

指数族分布 — 三大性质

重要应用

广义线性模型
- 线性组合 $w^T x$
- link function 激活函数的反函数
- 指数族分布：$y|x \sim$ 指数族分布 < 线性回归：$y|x \sim N(\mu, \Sigma)$；分类：{ $y|x \sim$ Bernoulli, $y|x \sim$ possion }

概率图模型 无向图 RBM 受限玻尔兹曼机

变分推断 后面介绍

八、指数族分布 (Exponential Family Distribution)

一 高斯分布的指数族形式

高斯分布的概率密度函数改写成标准指数族分布的形式

上节回顾

$p(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$

- η 为参数向量，$x \in \mathbb{R}^p$
- $A(\eta)$: log partition function（对数配分函数）
- $\phi(x)$: sufficient statistics（充分统计量）

高斯分布

$\theta(\mu,\sigma^2)$ θ改写成 μ,σ^2 的形式

$\frac{1}{2\pi\sigma}$ 含参 σ 的一定要塞到exp里

$$
\begin{aligned}
p(x|\theta) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x^2-2x\mu+\mu^2)\right) \\
&= \exp\log(2\pi\sigma^2)^{-\frac{1}{2}} \cdot \exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x) - \frac{\mu^2}{2\sigma^2}\right\} \\
&= \exp\log(2\pi\sigma^2)^{-\frac{1}{2}} \cdot \exp\left(-\frac{1}{2\sigma^2}(-2\mu \quad 1)\begin{pmatrix} x \\ x^2 \end{pmatrix} - \frac{\mu^2}{2\sigma^2}\right) \\
&= \exp\left\{\underbrace{\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)}_{\eta^T}\underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{\phi(x)} - \underbrace{\left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2)\right)}_{A(\eta)}\right\}
\end{aligned}
$$

$h(x) = 1$

建议写成 $\phi(x)^T\eta$

定义（设）

$$\Rightarrow \eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \mu/\sigma^2 \\ -\frac{1}{2\sigma^2} \end{pmatrix} \Rightarrow \begin{cases} \mu = -\frac{\eta_1}{2\eta_2} \\ \sigma^2 = -\frac{1}{2\eta_2} \end{cases}$$

代入

$$\begin{cases} \phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \\ A(\eta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \\ \qquad = \frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{2\eta_2}\right) \end{cases}$$

因为要消掉 $A(\eta)$ 里的 σ^2 和 μ

$\Rightarrow p(x|\theta) = 1 \cdot \exp\{\eta^T\phi(x) - A(\eta)\}$ 其中 η $\phi(x)$ $A(\eta)$ 见上

八、指数族分布 (Exponential Family Distribution)

—— 对数配分函数与充分统计量的关系

指数族分布定义：

$p(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$

- η 为参数向量，$x \in R^p$
- $A(\eta)$: log partition function (对数配分函数)
- $\phi(x)$: sufficient statistics (充分统计量)

高斯分布

$p(x|\theta) = \exp\{\eta^T\phi(x) - A(\eta)\}$

- $\eta = \begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}\mu/\sigma^2\\ -1/2\sigma^2\end{pmatrix}$, $\phi(x) = \begin{pmatrix}x\\ x^2\end{pmatrix}$
- $A(\eta) = -\eta_1^2/4\eta_2 + \frac{1}{2}\log(-2/\eta_2)$

$A(\eta)$ log partition function.

当概率密度函数 $f(x,\eta) = h(x)\exp\{\eta^T\phi(x) - A(\eta)\}$ 时，又 $\because \int f(x,\eta)dx = 1$，

那么 $\phi(x)$ 和 $A(\eta)$ 肯定存在某种关系，下面进行探索：

$$p(x|\eta) = h(x)\exp\{\eta^T\phi(x) - A(\eta)\}$$
$$= \frac{1}{\exp(A(\eta))} h(x)\exp(\eta^T\phi(x))$$

形如 $p(x|\eta) = \frac{1}{z}\hat{p}(x|\eta)$，$z = \int \hat{p}(x,\eta)dx$ 上式两边积即可

$$\Rightarrow \exp(A(\eta)) = \int h(x)\exp(\eta^T\phi(x))dx$$

对 η 求偏导

两边求偏导 $\Rightarrow$

$$\exp(A(\eta))A'(\eta) = \frac{\partial}{\partial\eta}\left(\int h(x)\exp(\eta^T\phi(x))dx\right)$$
$$= \int h(x)\exp(\eta^T\phi(x))\cdot\phi(x)\cdot dx$$

$\phi(x) \rightarrow \frac{\partial}{\partial\eta}(\eta^T\phi(x))$

$$\Rightarrow A'(\eta) = \frac{\int h(x)\exp(\eta^T\phi(x))\cdot\phi(x)\cdot dx}{\exp(A(\eta))}$$

放进去

$$= \int \underbrace{h(x)\exp(\eta^T\phi(x) - A(\eta))}_{p(x|\eta)}\cdot\phi(x)\, dx$$
$$= \int p(x|\eta)\cdot\phi(x)\, dx \quad \text{(由期望定义)}$$
$$= E_{p(x|\eta)}[\phi(x)]$$

即存在这样的关系 $A'(\eta) = E_{p(x|\eta)}[\phi(x)]$

同理 $A''(\eta) = Var[\phi(x)] \geq 0$

故 $A(\eta)$ 一定是凸函数.

我们可以上面结果代入高斯分布进行检验：

$E[\phi(x)] = \begin{bmatrix}E(x)\\ E(x^2)\end{bmatrix}$

易知 $E(x) = \mu$，对应的是 $\frac{\partial A(\eta)}{\partial\eta_1} = -2\eta_1/4\eta_2 = \frac{-2\cdot\mu/\sigma^2}{4\cdot(-\frac{1}{2\sigma^2})} = \mu$

$E(x^2) = (Ex)^2 + Var(x) = \mu^2 + \sigma^2$，$\frac{\partial A(\eta)}{\partial\eta_2} = \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} = \frac{(\mu/\sigma^2)^2}{4(-\frac{1}{2\sigma^2})^2} - \frac{1}{2\cdot(-\frac{1}{2\sigma^2})}$

$= \mu^2 + \sigma^2$

八、指数族分布 (Exponential Family Distribution)

—— 极大似然估计与充分统计量的关系

指数族分布定义：

$p(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta))$

- η 为参数向量，$x \in R^p$
- $A(\eta)$：log partition function（对数配分函数）
- $\phi(x)$：sufficient statistics（充分统计量）

其中 $A'(\eta) = E_{p(x|\eta)}[\phi(x)]$　　$A''(\eta) = Var[\phi(x)]$

$D = \{x_1, \cdots, x_N\}$　Data

$$
\begin{aligned}
\eta_{MLE} &= \arg\max_{\eta} \log P(D|\eta) \\
&= \arg\max_{\eta} \log \prod_{i=1}^{N} P(x_i|\eta) \\
&= \arg\max_{\eta} \sum_{i=1}^{N} \log P(x_i|\eta) \\
&= \arg\max_{\eta} \sum_{i=1}^{N} \log[h_i(x)\exp\{\eta^T\phi_i(x) - A(\eta)\}] \\
&= \arg\max_{\eta} \sum_{i=1}^{N} [\log h_i(x) + \eta^T\phi_i(x) - A(\eta)\} \\
&= \arg\max_{\eta} \sum_{i=1}^{N} [\eta^T\phi_i(x) - A(\eta)]
\end{aligned}
$$

化简

（$\log h_i(x)$ 和 η 无关）

求偏导 $= 0$

$$\frac{\partial \sum_{i=1}^{N}[\eta^T\phi_i(x) - A(\eta)]}{\partial\eta} = \sum_{i=1}^{N}[\phi_i(x) - A'(\eta)]$$

$$= \sum_{i=1}^{N}\phi_i(x) - NA'(\eta) = 0$$

$$\Rightarrow A'(\eta^*) = \frac{1}{N}\sum_{i=1}^{N}\phi_i(x)$$

（可真是秒啊 ^_^）

η^* 是 $\partial = 0$ 时估计的一个值，即 η_{MLE}

令 $g(\eta) = A'(\eta^*)$

故 $\eta^* = g^{-1}(\eta)$（即 η^* 是求 $A'(\eta^*)$ 的反函数） A' 的

对于高斯分布而言，η 表示 (μ, σ^2) 这些参数，为求解分布表达式，我们不用保存每一个样本，相反我们只需要求出一个值，即 $\frac{1}{N}\sum_{i=1}^{N}\phi_i(x)$ 就能利用反函数求出 η_{MLE}.

$\varphi(x)$ 这就是充分统计量

不需要 D，就只 $\sum\varphi_i(x)$ 即可

八、指数族分布（Exponential Family Distribution）

—— 最大熵角度

熵取最大的推导：

熵：对信息可能性的衡量

西瓜P75 信息熵

信息量：$-\log P$ （P是概率）

熵定义：$E_{p(x)}[-\log P] = H(p) = \int -p(x)\log p(x)\,dx$ （连续）

$= -\sum_{x} p(x)\log p(x)$ （离散）

最大熵 ⟺ 等可能

精确表达

以离散情况为例：

X	1	2	…	k
P	p_1	p_2	…	p_k

约束优化

subject to

转化为优化问题：$\max H(p) = \max\left[-\sum_{i=1}^{k} p(x_i)\log p(x_i)\right]$，s.t. $\sum_{i=1}^{k} p_i = 1$

$\Leftrightarrow \min\left[\sum_{i=1}^{k} p(x_i)\log p(x_i)\right]$，s.t. $\sum_{i=1}^{k} p_i = 1$

$\Rightarrow \hat{p}_i = \arg\max_{p} H(p) = \arg\min_{p} \sum_{i=1}^{k} p(x_i)\log p(x_i)$，s.t. $\sum_{i=1}^{k} p_i = 1$

拉格朗日

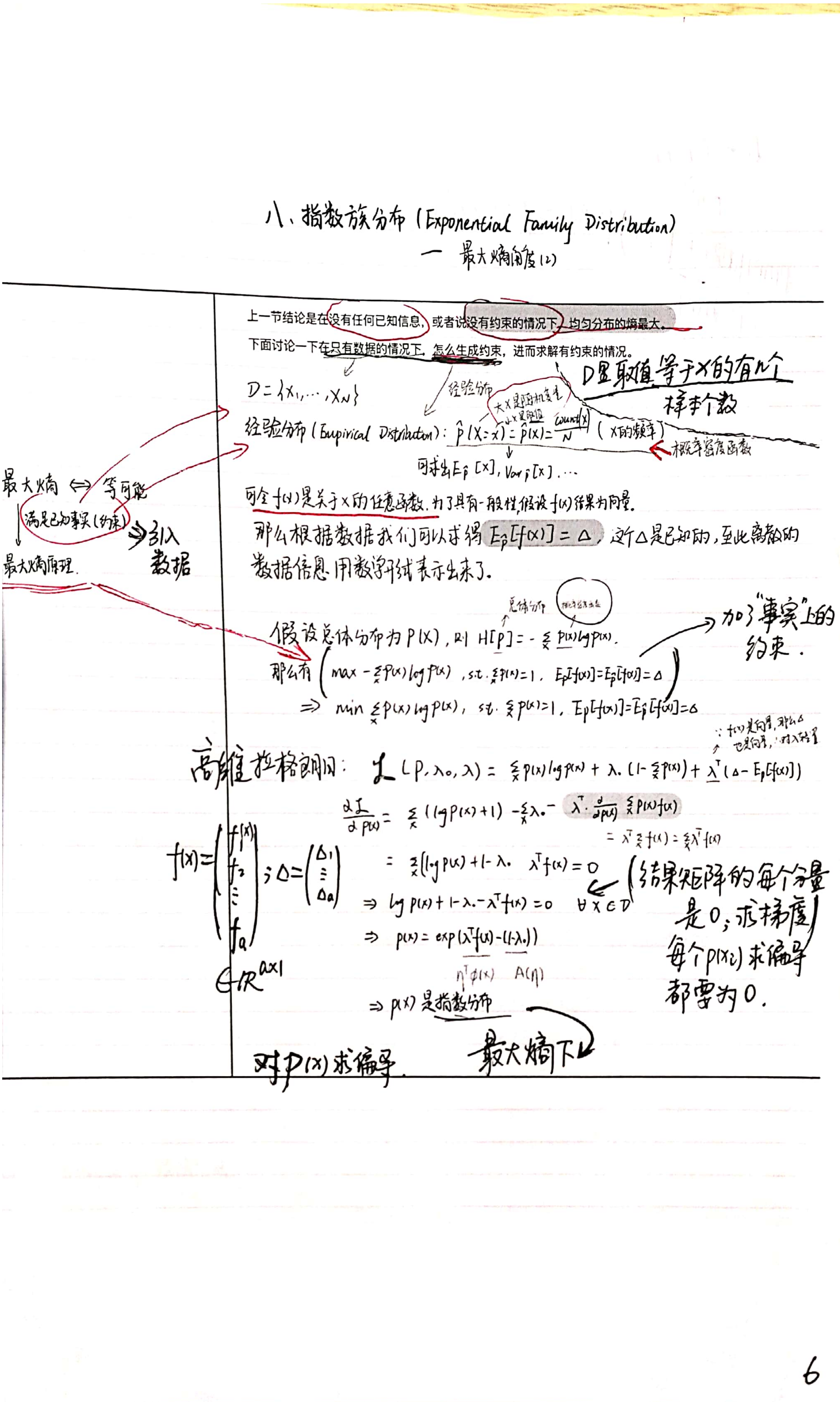
$\Rightarrow \mathcal{L}(p,\lambda) = \sum_{i=1}^{k} p_i\log p_i + \lambda\left(1-\sum_{i=1}^{k} p_i\right)$

$\frac{\partial \mathcal{L}(p,\lambda)}{\partial p_i} = \log p_i + 1 - \lambda = 0$

$\Rightarrow \hat{p}_i = \exp(\lambda - 1)$ 都相等

∵ λ 是常量（constant），故 $\hat{p}_1 = \hat{p}_2 = \cdots \hat{p}_k = \exp(\lambda-1) = \frac{1}{k}$

故 $p(x)$ 为均匀分布



八、指数族分布（Exponential Family Distribution）

—— 最大熵角度（2）

上一节结论是在没有任何已知信息，或者说没有约束的情况下，均匀分布的熵最大。

下面讨论一下在只有数据的情况下，怎么生成约束，进而求解有约束的情况。

$D = \{x_1, \cdots, x_N\}$

D里取值等于x的有N个样本个数

经验分布

大X是随机变量 小x是取值

经验分布（Empirical Distribution）：$\hat{p}(X=x) = \hat{p}(x) = \frac{count(x)}{N}$（x的频率）

概率密度函数

可求出 $E_{\hat{p}}[x]$，$Var_{\hat{p}}[x]$ ……

最大熵 $\Leftrightarrow$ 等可能

满足已知事实（约束）$\Rightarrow$ 引入数据

最大熵原理.

可令 $f(x)$ 是关于 x 的任意函数，为了具有一般性，假设 $f(x)$ 结果为向量。

那么根据数据我们可以求得 $E_{\hat{p}}[f(x)] = \Delta$，这个 Δ 是已知的，至此离散的数据信息用数学形式表示出来了。

总体分布

概率密度函数

假设总体分布为 $p(x)$，则 $H[p] = -\sum_x p(x)\log p(x)$.

加了"事实"上的约束.

那么有 $\max -\sum_x p(x)\log p(x)$，s.t. $\sum_x p(x) = 1$，$E_p[f(x)] = E_{\hat{p}}[f(x)] = \Delta$

$\Rightarrow \min \sum_x p(x)\log p(x)$，s.t. $\sum_x p(x) = 1$，$E_p[f(x)] = E_{\hat{p}}[f(x)] = \Delta$

$\because f(x)$ 是向量，那么 Δ 也是向量，$\therefore$ 对 λ 转置

高维拉格朗日：$\mathcal{L}(p, \lambda_0, \lambda) = \sum_x p(x)\log p(x) + \lambda_0(1 - \sum_x p(x)) + \lambda^T(\Delta - E_p[f(x)])$

$$\frac{\partial \mathcal{L}}{\partial p(x)} = \sum_x(\log p(x) + 1) - \sum_x \lambda_0 - \lambda^T \cdot \frac{\partial}{\partial p(x)} \sum_x p(x) f(x)$$

$= \lambda^T \sum_x f(x) = \sum_x \lambda^T f(x)$

$f(x) = \begin{pmatrix} f_1(x) \\ f_2 \\ \vdots \\ f_a \end{pmatrix}$；$\Delta = \begin{pmatrix} \Delta_1 \\ \vdots \\ \Delta_a \end{pmatrix}$ $\in \mathbb{R}^{a \times 1}$

$= \sum_x (\log p(x) + 1 - \lambda_0 - \lambda^T f(x)) = 0$

（结果矩阵的每个分量是0；求梯度每个 $p(x_i)$ 求偏导都要为0.）

$\Rightarrow \log p(x) + 1 - \lambda_0 - \lambda^T f(x) = 0 \quad \forall x \in D$

$\Rightarrow p(x) = \exp(\underbrace{\lambda^T f(x)}_{\eta^T \phi(x)} - \underbrace{(1-\lambda_0)}_{A(\eta)})$

$\Rightarrow p(x)$ 是指数分布

最大熵下

对 $p(x)$ 求偏导.

6

3.3.1 - 式(3.49) (3.50) (3.51) 推导.

实际上是式(2.116)的推导.

- 解：

3.3.1 - 式(3.51) 附近, 如果我们考虑一个无限宽的先验 $S_0 = \alpha^{-1}\boldsymbol{I}$, 其中 $\alpha \to 0$, 那么后验概率分布的均值就变成了公式(3.15)给出的最大似然估计值 $\boldsymbol{w}_{ML}$.

3.3.1 - 式(3.55) 推导.

- 解:

(3.55)

$p(w|t) = \mathcal{N}(w \mid m_N, S_N)$ 但此时先验为零均值

$$\mathcal{L} = \ln \prod_{i}^{N} \mathcal{N}(w_i \mid m_N, S_N)$$

$=$

3.3.2 - 式(3.58) (3.59) 推导.

- 解:

3.3.2 - 式(3.63) 推导.

- 解:

$$
\begin{aligned}
(3.63)\quad \operatorname{cov}(y(x), y(x')) &= \operatorname{cov}(\phi(x)^T w, w^T\phi(x')) \\
&\overset{①}{=} \phi^T(x) S_N \phi(x') \\
&\overset{②}{=} \beta^{-1} k(x, x')
\end{aligned}
$$

等号②由 (3.62) 易得

等号①：(3.58) 有说到预测分布

即 $y(x,w) = w^T\phi(x) \sim \mathcal{N}(t \mid m_N^T\phi(x), \sigma_N^2(x))$

$$
\begin{aligned}
\text{又} \ \mathbb{E}[ww^T] &= \operatorname{cov}(w) + \mathbb{E}[w]\mathbb{E}[w]^T \\
&= S_N + m_N m_N^T
\end{aligned}
$$

$$
\begin{aligned}
\therefore \text{等式①} \ & \operatorname{cov}(\phi(x)^T w, w^T\phi(x')) \\
&= \mathbb{E}[\phi(x)^T w w^T \phi(x')] - \phi(x)^T m_N m_N^T \phi(x') \\
&= \phi(x)^T (\operatorname{cov}(w) + \mathbb{E}[w]\mathbb{E}[w]^T) \phi(x)^T - \cdots \\
&= \phi(x)^T S_N \phi(x') \qquad \square
\end{aligned}
$$

3.4 - 式(3.70) $\Delta w_{posterior}$ 的积分近似.

- 解：

  因为假设了后验分布近似均匀分布，那么它乘一个分布作为被积函数，就是均匀分布的长度被积函数即可.

3.5 - 式(3.74) 如果定义了 α, β 上的共轭Gamma先验分布, 那么对(3.74)中的w就可以解析地得到w上的t分布.

3.5.1 - 式(3.77) 使用高斯模型条件概率分布的结果(2.115)计算.

3.5.1 - 式(3.78) (3.79) 推导.

- 解：

  由 (3.11) (3.12) (3.52)：

$$
\begin{aligned}
\ln p(\mathbf{t} \mid \boldsymbol{w}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}\left(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}\right) \\
&= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\boldsymbol{w})
\end{aligned}
$$

$$E_D(\boldsymbol{w})=\frac{1}{2}\sum_{n=1}^{N}\left\{t_n-\boldsymbol{w}^T\boldsymbol{\phi}\left(\boldsymbol{x}_n\right)\right\}^2$$

$$p(\boldsymbol{w}\mid\alpha)=\mathcal{N}\left(\boldsymbol{w}\mid\mathbf{0},\alpha^{-1}\boldsymbol{I}\right)$$

(3.78) 中的 $\left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}}$ 是 $p(\boldsymbol{t}\mid\alpha,\beta)$ 的，另一个是 $p(\boldsymbol{w}\mid\alpha)$ 的.

3.5.1 - 式(3.85) 推导: 关系到(3.79)~(3.84), 也是证据函数的计算过程.

3.1.1 - 代码中, 基函数回归里极大似然估计, 基函数(比如高斯)的(初始化)均值会对结果产生影响, 为什么?

3.3 - 贝叶斯线性回归有什么缺点, 好像现在用的很少?

3.5.2 - 式(3.89) 求导推导.

- 解:

  这里不用查cookbook，行列式的求导就是先展开成特征值的乘积，看(3.88)即可.

3.5.2 - 式(3.91) 推导.

- 解:

  由(3.90):

$$\begin{aligned}\gamma&=M-\alpha\sum_i\frac{1}{\lambda_i+\alpha}\\&=\sum_i(1-\frac{\alpha}{\lambda_i+\alpha})\end{aligned}$$

3.5.2 - 如何理解最大化证据函数中迭代计算 α, 直到收敛的过程?

4.1.4 - 式(4.26) 推导.

4.1.5 - 式(4.34) 推导.

- 解:

  式(4.32) 除 N 得到:

$$\begin{aligned}&\sum_n^N\boldsymbol{w}^Tm+w_0-\left(\sum_{n\in N_1}t_n+\sum_{n\in N_2}t_n\right)/N\\=&\sum_n^N\boldsymbol{w}^Tm+w_0=0\end{aligned}$$

4.1.5 - 式(4.37) 推导 (习题4.6).

- 解:

4.1.6 - S_B 秩最多 $K-1$ 个, 向由S_B张成的 $K-1$ 维空间上的投影不会改变 $J(W)$ 的值，因此通过这种方法我们不能够找到多于 $K-1$ 个线性“特征”. 如何理解?