

第四讲 贝叶斯估计

CSDN学院
2017年7月



► 参数估计

- 给定模型类别 $p(x|\theta)$ 和数据 \mathcal{D} , 选择与数据最匹配的参数 θ : 参数估计
- 有多种方法可用来估计模型的参数
 - 矩估计法
 - 极大似然估计 : 频率学派
 - 贝叶斯方法 : 贝叶斯学派
 - 参数也是随机变量 , 可以讨论其分布

► Outline

- 贝叶斯估计的基本思想
 - 先验、似然、后验、后验预测
- 常见分布参数的贝叶斯估计
 - 正态分布、Binomial分布、Multinomial分布
- 一些机器学习模型的参数估计
 - 线性回归、Logistic回归



一、贝叶斯估计基本思想

► 贝叶斯估计

- MLE是频率学派估计参数的方法
 - MLE只是一个点估计
 - 估计的不确定性可以通过计算其方差确定（估计的分布只能根据其渐近正态的性质假设为正态分布）
- 贝叶斯估计：参数也是随机变量，用概率分布描述其性质
 - 先验分布（prior） $p(\theta)$ ：在没有看到数据之前，参数的分布
 - 同MLE相同，似然为 $p(\mathcal{D}|\theta)$
 - 后验分布（posterior） $p(\theta|\mathcal{D})$ ：在看到数据后 \mathcal{D} ，对参数分布的更新

► 贝叶斯估计

- 设**先验**为： $p(\theta)$
 - 先验反映我们对参数取值的信念：偏好更简单或更光滑的模型
 - 如果不太确定参数的取值范围，实践中通常取一个分布范围较宽的分布，反映我们对参数的不确定性
- **似然**为： $p(\mathcal{D}|\theta)$
- 则根据贝叶斯公式，得到参数的**后验**为
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \propto p(\mathcal{D}|\theta)p(\theta)$$
 - 参数估计不再是一个点估计，而是一个分布（信息更多）

► 贝叶斯估计

- 参数的后验估计为 $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$
- 则后验预测为： $p(X=\tilde{x}|\mathcal{D}) = \int p(X=\tilde{x}|\theta)p(\theta|\mathcal{D})d\theta$
 - 对参数 θ 积分：用参数后验进行加权平均



二、常见分布的贝叶斯估计

► Beta-binomial模型——先验

- 假设 $X_i \sim \text{Ber}(\theta)$, 则IID数据 $\mathcal{D} = \{X_1, \dots, X_N\}$ 似然为 $p(\mathcal{D}|\theta) = \theta^{N_1} (1-\theta)^{N_0}$
- 其中 $N_1 = \sum_{i=1}^N \mathbb{I}(x_i = 1), N_0 = \sum_{i=1}^N \mathbb{I}(x_i = 0)$
 - 注意: $N_1 \sim \text{Bin}(N_1 | \theta, N_1 + N_0)$
- 为了计算方便, 参数 θ 先验的形式最好与似然相同: **共轭先验**, 即 $p(\theta) = \theta^{\gamma_1} (1-\theta)^{\gamma_2}$
- 在Binomial模型中, 共轭先验为**Beta分布**: $\text{Beta}(\theta | a, b) \propto \theta^{a-1} (1-\theta)^{b-1}$
 - a, b : **超参数**

► Beta-binomial模型——后验

$$\text{Beta}(\theta | a, b) \propto \theta^{a-1} (1-\theta)^{b-1}$$

- 似然为 $p(\mathcal{D} | \theta) = \theta^{N_1} (1-\theta)^{N_0}$
- 先验为 **Beta分布**: $p(\theta | a, b) = \theta^{a-1} (1-\theta)^{b-1}$
- 则后验

$$\begin{aligned} p(\theta | \mathcal{D}) &\propto p(\mathcal{D} | \theta) p(\theta) \\ &\propto \theta^{N_1} (1-\theta)^{N_0} \theta^{a-1} (1-\theta)^{b-1} \\ &= \theta^{N_1+a-1} (1-\theta)^{N_0+b-1} \end{aligned}$$

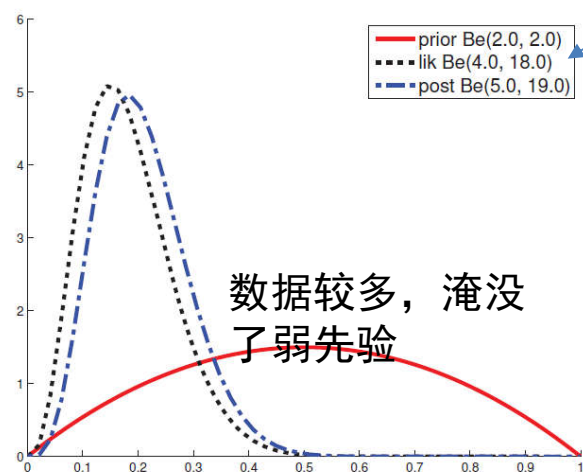
$$\propto \text{Bin}(N_1 | N_1 + N_0, \theta) \text{Beta}(\theta | a, b)$$

$$\propto \text{Beta}(\theta | N_1 + a, N_0 + b)$$
- 为Beta分布 $\text{Beta}(\theta | N_1 + a, N_0 + b)$
 - 将超参数加到经验计数上：先验的强度（先验的有效样本大小）为伪计数的和 $a+b$ ，与样本数 $N_1 + N_0$ 的作用类似

► 例：

• 例：

MLE等价于先验为Beta (1,1) 的后验

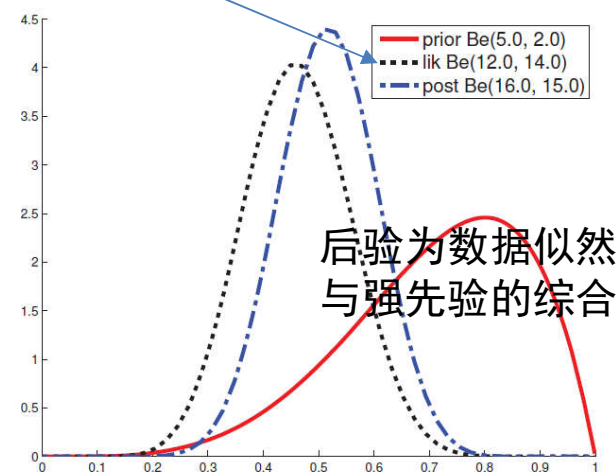


(a)

$$N_1 = 3, N_0 = 17$$

先验 Beta(2,2)

后验为 Beta(5,19)



(b)

$$N_1 = 11, N_0 = 13$$

先验 Beta(5,2)

后验为 Beta(16,15)

► Beta-binomial模型——后验点估计

- 后验为 $p(\theta|\mathcal{D}) \propto \text{Beta}(\theta|N_1+a, N_0+b)$ ，则**最大后验估计** (Maximum A Posteriori, MAP)为

$$\hat{\theta}_{MAP} = \frac{a+N_1-1}{a+b+N-2}$$

- 若采用均匀先验，则MAP退化为MLE: $\hat{\theta}_{MAP} = \frac{N_1}{N} = \hat{\theta}_{MLE}$
- 后验均值**: $\bar{\theta} = \frac{a+N_1}{a+b+N}$
- 可视为先验均值和MLE的加权平均，令 $m = \frac{a}{a+b}$

$$\begin{aligned} \bar{\theta} &= \frac{a+N_1}{a+b+N} = \frac{am+N_1}{a+b+N} = \frac{a}{a+b+N}m + \frac{\frac{a+b}{N}}{a+b+N} \times \frac{N_1}{N} \\ &= \lambda m + (1-\lambda)\hat{\theta}_{MLE} \end{aligned}$$

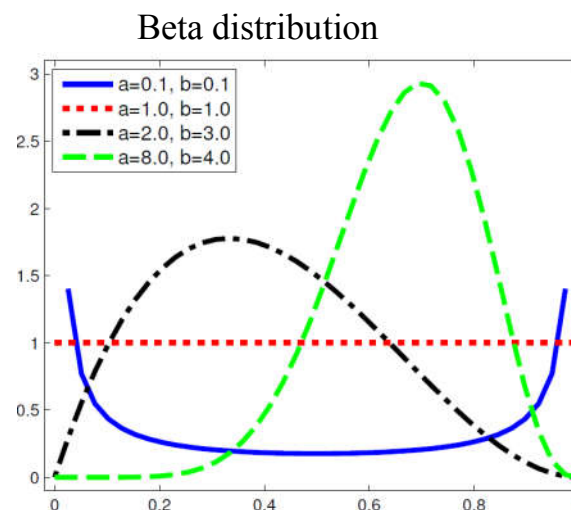
► Recall: Beta分布

- $$\text{Beta}(\theta|a,b) = \frac{1}{B(a,b)} \theta^{a-1} (1-\theta)^{b-1}$$

$$B(a,b) := \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$\mathbb{E}(\theta) = \frac{a}{a+b}, \quad \text{mode}[\theta] = \frac{a-1}{a+b-2}$$

$$\mathbb{V}(\theta) = \frac{ab}{(a+b)^2(a+b+1)}$$



► Beta-binomial模型——后验方差

- 后验为 $p(\theta|\mathcal{D}) \propto \text{Beta}(\theta|N_1+a, N_0+b)$ ，则方差为

$$\mathbb{V}(\theta|\mathcal{D}) = \frac{(a+N_1)(b+N_0)}{(a+b+N_1+N_0)^2(a+b+N_1+N_0+1)}$$

- 当 $N \gg a, b$ 时,

$$\mathbb{V}(\theta|\mathcal{D}) = \frac{N_1 N_0}{N^2 N} = \frac{\hat{\theta}(1-\hat{\theta})}{N}$$

- 结果同MLE。
- 标准差为 $\sigma = \sqrt{\mathbb{V}(\theta|\mathcal{D})}$

► Beta-binomial模型——在线学习

- 序列训练：有两批数据 \mathcal{D}_a 和 \mathcal{D}_b 顺序来更新模型，其充分统计量分别为 N_1^a, N_0^a 和 N_1^b, N_0^b ，则

$$\begin{aligned}
 p(\theta | \mathcal{D}_a, \mathcal{D}_b) &\propto p(\mathcal{D}_b | \theta) p(\theta | \mathcal{D}_a) \\
 &\propto \text{Bin}(N_1^b | \theta, N_1^b + N_0^b) \text{Beta}(\theta | N_1^a + a, N_0^a + b) \\
 &\propto \text{Beta}(\theta | N_1^a + N_1^b + a, N_0^a + N_0^b + b)
 \end{aligned}$$

- 这两批数据合在一起更新模型时，令 $N_1 = N_1^a + N_1^b, N_0 = N_0^a + N_0^b$

$$\begin{aligned}
 p(\theta | \mathcal{D}_a, \mathcal{D}_b) &\propto \text{Bin}(N_1 | \theta, N_1 + N_0) \text{Beta}(\theta | a, b) \\
 &\propto \text{Beta}(\theta | N_1 + a, N_0 + b) \\
 &\propto \text{Beta}(\theta | N_1^a + N_1^b + a, N_0^a + N_0^b + b)
 \end{aligned}$$



二者的结果相同：贝叶斯方法和很适合与online learning结合

► Beta-binomial模型——后验预测

- 后验为 $p(\theta | \mathcal{D}) \propto \text{Beta}(N_1 + a, N_0 + b)$
- 后验预测为

$$\begin{aligned} p(\tilde{x}=1 | \mathcal{D}) &= \int_0^1 p(\tilde{x}=1 | \theta) p(\theta | \mathcal{D}) d\theta \\ &= \int_0^1 \theta \text{Beta}(\theta | N_1 + a, N_0 + b) d\theta = \mathbb{E}(\theta | \mathcal{D}) = \frac{a + N_1}{a + b + N} \end{aligned}$$

► Dirichlet-multinomial模型——先验

- 将两种输出可能（投掷硬币）推广到投掷 K 面骰子
- 观测到骰子投掷结果 $\mathcal{D}=\{x_1,\dots,x_N\}$ ，其中 $x_i \in \{1,\dots,K\}$ ，则 multinomial似然为

$$p(\mathcal{D}|\boldsymbol{\theta})=\prod_{k=1}^K \theta_k^{N_k}$$

- 其中 $N_k = \sum_{i=1}^N \mathbb{I}(x_i = k)$ 表示 N 次试验中第 k 面出现的次数
- 共轭先验为Dirichlet分布： $\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})=\frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^K \theta_k^{\alpha_k-1}$

► Dirichlet-multinomial模型——后验

- 则后验为: $p(\theta | \mathcal{D}) \propto p(\mathcal{D} | \theta) p(\theta)$

$$\begin{aligned} &\propto \prod_{k=1}^K \theta_k^{N_k} \prod_{k=1}^K \theta_k^{a_k-1} = \prod_{k=1}^K \theta_k^{N_k+a_k-1} \\ &\propto \text{Dir}(\boldsymbol{\theta} | \alpha_1 + N_1, \dots, \alpha_K + N_K) \end{aligned}$$

– 后验：将超参数 α_k 加到经验计数 N_k 上

- MAP为 $\hat{\theta}_k = \frac{\alpha_k + N_k - 1}{\alpha_0 + N - K}$

- 当 $\alpha_k = 1$ 时，退化为MLE $\hat{\theta}_{MAP} = \frac{N_k}{N} = \hat{\theta}_{MLE}$



► Dirichlet-multinomial模型——后验预测

- 后验预测为

$$\begin{aligned} p(\tilde{x}=j|\mathcal{D}) &= \int p(\tilde{x}=j|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\ &= \int p(\tilde{x}=j|\theta_j) \left[\int \dots \int p(\boldsymbol{\theta}_{-j}|\theta_j) d\boldsymbol{\theta}_{-j} \right] d\theta_j \\ &= \int \theta_j p(\theta_j|\mathcal{D}) d\theta_j = \mathbb{E}(\theta_j|\mathcal{D}) = \frac{\alpha_j + N_j}{\sum_k \alpha_k + N} \end{aligned}$$

► 例：Bag of Words语言模型

- 例：假设词典为

mary	lamb	little	big	fleece	white	black	snow	rain	unk
1	2	3	4	5	6	7	8	9	10

- 给定序列

Mary had a little lamb, little lamb, little lamb,
 Mary had a little lamb, its fleece as white as snow

- 得到每个单词的使用次数(直方图)为

Token	1	2	3	4	5	6	7	8	9	10
Word	mary	lamb	little	big	fleece	white	black	snow	rain	unk
Count	2	4	4	0	1	1	0	1	0	4

- 用 N_j 表示单词j出现的次数， θ 的先验用Dir(α)表示，则后验预测为

$$p(\tilde{x}=j|\mathcal{D})=\mathbb{E}(\theta_j|\mathcal{D})=\frac{\alpha_j+N_j}{\sum_k \alpha_k+N}$$

令 $\alpha_j=1$ ，得到 $p(\tilde{X}=j|D)=\frac{(3/27, \boxed{5/27}, 5/27, 1/27, 2/27, 2/27, 1/27, 2/27, 1/27, 5/27)}{\text{lamb}}$

► 共轭先验

- 若后验 $p(\theta|\mathcal{D}) \in \mathcal{F}$ ，则称先验 $p(\theta)$ 与似然 $p(\mathcal{D}|\theta) \in \mathcal{F}$ **共轭**

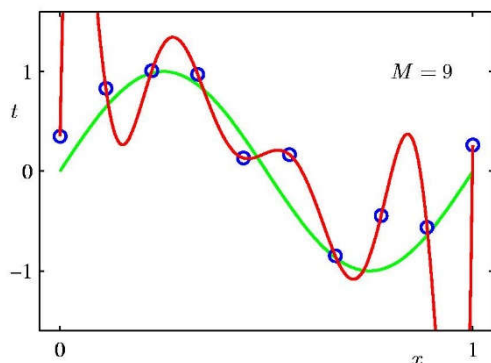
Likelihood	Prior
Binomial/ Bernoulli	Beta
Multinomial/ multinoulli	Dirichlet
Poisson	Gamma
MVN (fixed Σ)	MVN (Multiple Variables Normal)
MVN (fixed μ)	Wishart
MVN (general case)	MVN-Wishart
Exponential family	Conjugate
Linear regression (fixed σ^2)	MVN
Linear regression (general case)	MVN-Gamma



三、线性回归

► Recall : 线性回归

- 例: $Y = \sin(2\pi X) + \varepsilon, \varepsilon \sim \mathcal{N}(0, 0.3^2)$
- $N = 10$ 个数据点, 用9阶多项式拟合



系数:

0.35, 232.37, -5321.83, 48568.31, -231639.30

640042.26, -1061800.52, 1042400.18, -557682.99, 125201.43

- 系数有+、有-、绝对值非常大
- 曲线波动很大: 为了拟合数据点
- 如果数据有一点小的变化, 得到的拟合结果会有很大不同 → 结果不稳定

► 高斯先验


- **先验**：偏向较小的系数值，从而得到的曲线比较平滑：0均值的
高斯： $w_j \sim \mathcal{N}(0, \tau^2)$

$$p(\mathbf{w}) = \prod_{j=1}^D \mathcal{N}(w_j | 0, \tau^2) \propto \exp\left(-\frac{1}{2\tau^2} \sum_{j=1}^D \mathbf{w}_j^2\right) = \exp\left(-\frac{1}{2\tau^2} [\mathbf{w}^T \mathbf{w}]\right)$$

- 其中 $1/\tau^2$ 控制先验的强度

- **似然**： $p(y_i | \mathbf{x}_i, \mathbf{w}, \sigma^2) = \mathcal{N}(y_i | \mathbf{w}^T \mathbf{x}_i, \sigma^2)$
 $p(\mathbf{y} | \mathbf{X}, \mathbf{w}, w_0, \sigma^2) = \mathcal{N}(\mathbf{y} | \mathbf{X}\mathbf{w} + w_0 \mathbf{1}_N, \sigma^2 \mathbf{I}_N)$

$$\propto \exp\left(-\frac{1}{2\sigma^2} \underbrace{(\mathbf{y} - (\mathbf{X}\mathbf{w} + w_0 \mathbf{1}_N))^T (\mathbf{y} - (\mathbf{X}\mathbf{w} + w_0 \mathbf{1}_N))}_{RSS(\mathbf{w})}\right)$$

 **后验**： $p(\mathbf{w}, w_0 | \mathbf{X}, \mathbf{y}, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} [(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0 \mathbf{1}_N)^T (\mathbf{y} - \mathbf{X}\mathbf{w} - w_0 \mathbf{1}_N)] - \frac{1}{2\tau^2} [\mathbf{w}^T \mathbf{w}]\right)$

岭回归

$$p(\mathbf{w}, w_0 | \mathbf{X}, \mathbf{y}, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2}[(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_N)^T(\mathbf{y} - \mathbf{X}\mathbf{w} - w_0\mathbf{1}_N)] - \frac{1}{2\tau^2}[\mathbf{w}^T\mathbf{w}]\right)$$

- MAP等价于最小目标函数：

$$J(\mathbf{w}) = \underbrace{\sum_{i=1}^N \left(y_i - (\mathbf{w}^T \mathbf{x}_i + w_0)\right)^2}_{RSS(\mathbf{w})} + \lambda \underbrace{\|\mathbf{w}\|_2^2}_{\substack{\text{正则项,} \\ \text{复杂性惩罚}}}$$
$$= (\mathbf{y} - (\mathbf{X}\mathbf{w} + w_0\mathbf{1}_N))^T (\mathbf{y} - (\mathbf{X}\mathbf{w} + w_0\mathbf{1}_N)) + \lambda \mathbf{w}^T \mathbf{w}$$

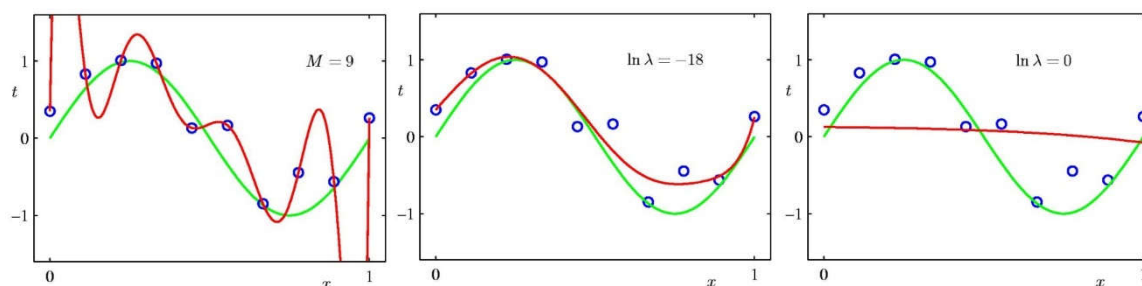
- 其中 $\lambda = \sigma^2 / \tau^2$
- 称为岭回归，或正则化的最小二乘
- 注意： w_0 没有被正则（ w_0 只影响函数的高度，不影响复杂性）



求解可通过对矩阵X进行SVD分解实现

► 例：多项式回归

• 例：

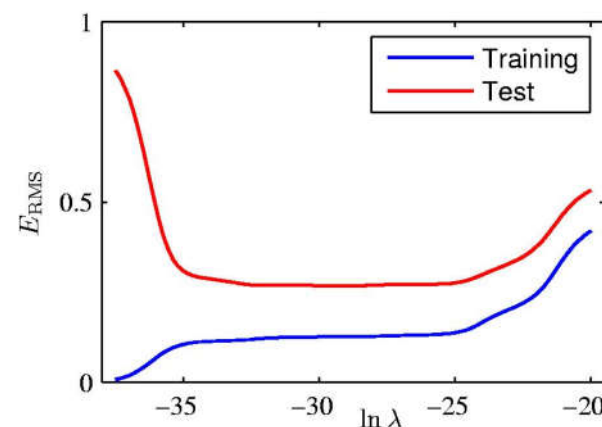


• 系数值：

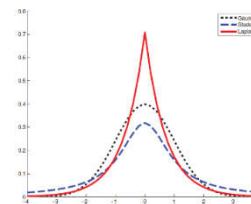
	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
w_0^*	0.35	0.35	0.13
w_1^*	232.37	4.74	-0.05
w_2^*	-5321.83	-0.77	-0.06
w_3^*	48568.31	-31.97	-0.05
w_4^*	-231639.30	-3.89	-0.03
w_5^*	640042.26	55.28	-0.02
w_6^*	-1061800.52	41.32	-0.01
w_7^*	1042400.18	-45.95	-0.00
w_8^*	-557682.99	-91.53	0.00
w_9^*	125201.43	72.68	0.01

► 例：多项式回归（cont.）

- 训练误差 vs 测试误差：
 - λ 增加(模型越简单)，训练误差总是增加
 - λ 增加，测试误差是U形，有一个最低点 → 最佳模型（模型复杂度适中）
 - λ 的选择：交叉验证(CV)



► Laplace先验



- 线性回归中参数的先验还可以设置为Laplace先验：

$$p(\mathbf{w} | \lambda) = \prod_{j=1}^D \text{Lap}(w_j | 0, 1/\lambda) \propto \prod_{j=1}^D \exp(-\lambda |w_j|) = \exp\left(-\lambda \sum_{j=1}^D |w_j|\right)$$

- 似然为：

$$p(y_i | \mathbf{x}_i, \mathbf{w}, \sigma^2) = \mathcal{N}(y_i | \mathbf{w}^T \mathbf{x}_i, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} (y_i - (\mathbf{w}^T \mathbf{x}_i + w_0))^2\right)$$

- 后验为：

$$p(\mathbf{w}, w_0 | \mathbf{X}, \mathbf{y}, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (\mathbf{w}^T \mathbf{x}_i + w_0))^2 - \sum_{j=1}^D \lambda |w_j|\right)$$

► Laplace先验 (cont.)

- 后验 $p(\mathbf{w}, w_0 | \mathbf{X}, \mathbf{y}, \sigma^2) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - (\mathbf{w}^T \mathbf{x}_i + w_0))^2 - \sum_{j=1}^D \lambda |w_j|\right)$ 不是已知分布的概率函数(☹)

- 最大后验估计MAP等价于L1正则的线性回归 (Lasso) :

$$J(\mathbf{w}) = \underbrace{\sum_{i=1}^N (y_i - (\mathbf{w}^T \mathbf{x}_i + w_0))^2}_{RSS(\mathbf{w})} + \lambda \underbrace{\|\mathbf{w}\|_1}_{\substack{\text{正则项} \\ \text{复杂性惩罚}}}$$

- 当 λ 取合适值时, \mathbf{w} 变得稀疏 (有些系数为0)
- 相比岭回归, 优化计算更复杂

► 案例：波士顿房价预测



- L2L1LR_bostonhouseprice.ipynb



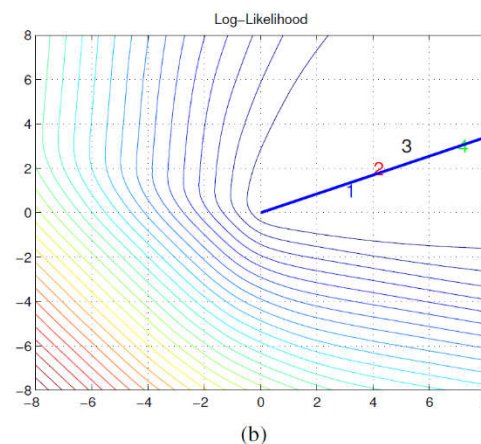
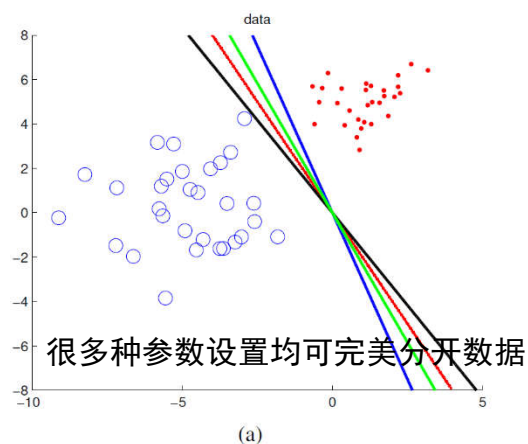


四、Logistic回归

► 贝叶斯Logistics回归

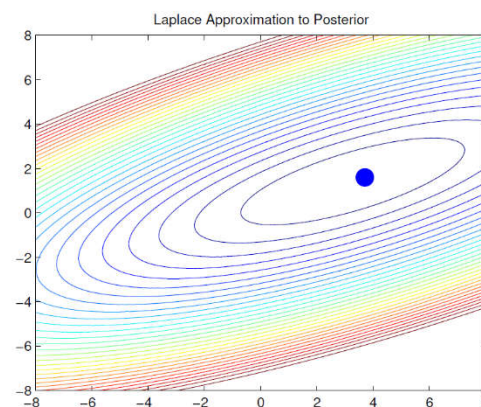
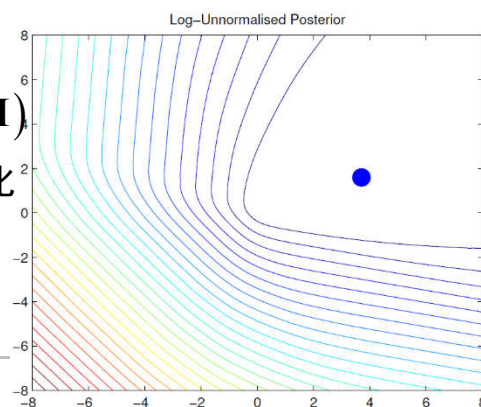
- 类似线性回归，Logistics回归也可以在MLE的基础上加正则
- 似然： $p(y_i = 1 | \mathbf{x}_i, \mathbf{w}) = \text{sigm}(y | \mathbf{w}^T \mathbf{x}_i)$
- 但没有共轭先验，通常取先验： $p(\mathbf{w}) = \mathcal{N}(\mathbf{m}_0, \mathbf{V}_0)$
- 事实上，正则对分类尤其重要
 - 当数据完全线性可分时，似然值在 $\|\mathbf{w}\| \rightarrow \infty$ 时最大（sigmoid函数最陡），但该模型在训练数据上概率最大，推广性不会太好
 - Sklearn中LogisticRegression缺省正则参数 $C=1$ ，必须有正则
- L2正则的目标函数为： $J'(\mathbf{w}) = NLL(\mathbf{w}) + \lambda \mathbf{w}^T \mathbf{w}$
 - 也可以和L1正则结合

► Example: LR



似然函数的最大值为 ∞
因为当数据完美可分时， $\|\mathbf{w}\|$ 越大，sigmoid函数越陡，从而似然值越大

增加先验：
 $\mathcal{N}(\mathbf{w} | \mathbf{0}, 100\mathbf{I})$
相当于正则化



Laplace近似比较粗糙，但大体形状相近

► 案例：蘑菇毒性预测



- L2L1LR_Mushroom.ipynb



► 总结

- 贝叶斯推断
 - 输入：数据 \mathcal{D} 和似然 $p(\mathcal{D}|\theta)$
 - 确定先验：为了计算方便，通常采用共轭先验 $p(\theta)$
 - 根据贝叶斯公式，计算后验 $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$
 - 模型训练
 - 根据学习到的模型进行预测（后验预测）
 - 模型测试

THANK YOU



AI100