5 Bayesian statistics

主讲人：吴晓晖

Introduction

1.参数估计（estimation）

- 目标：求 p(θ|D)
- 方法：MLE、MAP、经验贝叶斯、全贝叶斯等

2.贝叶斯派（Bayesian） vs 频率派（Frequentist）

- 区别：在求p(θ|D)时，频率派求一个定值，贝叶斯派求一个分布MLE、MAP属于**点估计（point estimation）**，属于频率派。

2.贝叶斯派 vs 频率派

1.关于参数的解释不同

经典估计方法认为待估参数具有确定值，它的估计量才是随机的，如果估计量是无偏的，该估计量的期望等于那个确定的参数。

而贝叶斯方法认为待估参数是一个服从某种分布的随机变量。

2.所利用的信息不同

经典方法只利用样本信息；

贝叶斯方法要求事先提供一个参数的先验分布，即人们对有关参数的主观认识，被称为先验信息，是非样本信息，在参数估计过程中，这些非样本信息与样本信息一起被利用。

3.对随机误差项的要求不同*

经典方法，除了最大似然法，在参数估计过程中并不要求知道随机误差项的具体分布形式，但是在假设检验与区间估计时是需要的；

贝叶斯方法需要知道随机误差项的具体分布形式。

4.选择参数估计量的准则不同*

经典估计方法或者以残差平方和最小，或者以似然函数值最大为准则，构造极值条件，求解参数估计量；

贝叶斯方法需要构造一个损失函数，并以损失函数最小化为准则求得参数估计量。

Reference：
参数估计：频率学派与贝叶斯学派

MLE和MAP估计

最大似然分布（MLE）

$$\hat{\theta} = argmax_{\theta} p(\mathcal{D}|\theta)$$

最大后验估计（MAP）

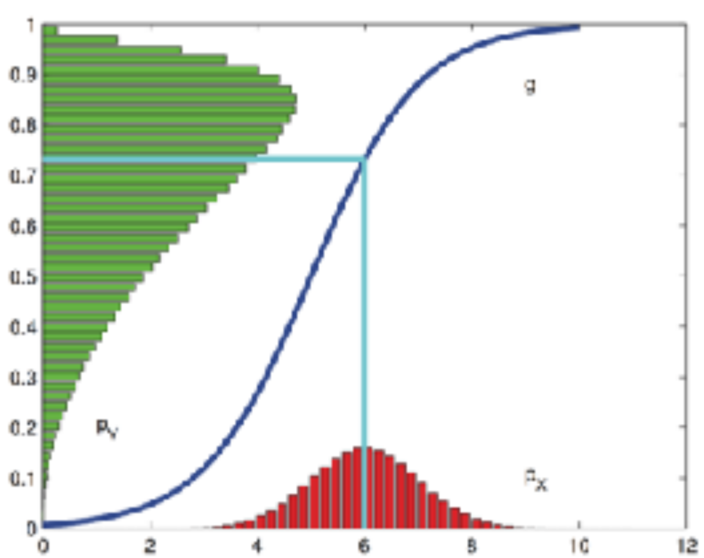
$$\begin{aligned}\hat{\theta} &= argmax_{\theta} p(\theta|\mathcal{D}) \\ &= argmax_{\theta} \frac{p(\mathcal{D}|\theta)p(\theta|\eta)}{p(\mathcal{D})} \\ &= argmax_{\theta} p(\mathcal{D}|\theta)p(\theta|\eta)\end{aligned}$$

Reference：

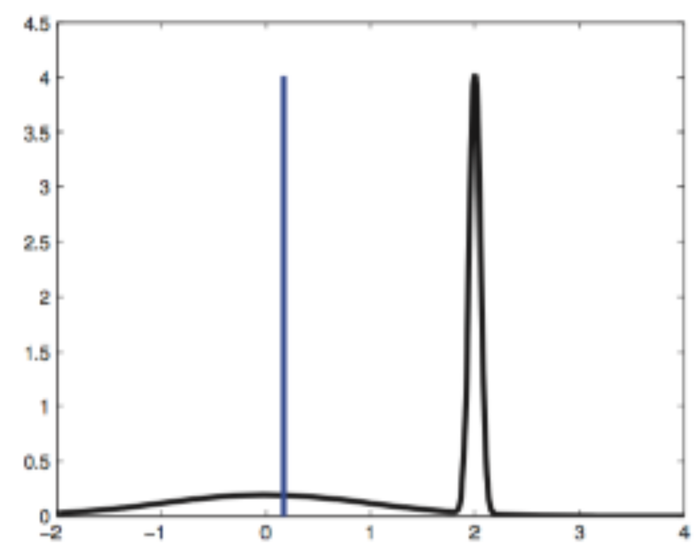
参数估计：最大似然估计（MLE），最大后验估计(MAP)，贝叶斯估计，经验贝叶斯(Empirical Bayes)与全贝叶斯(Full Bayes)

MAP的缺点

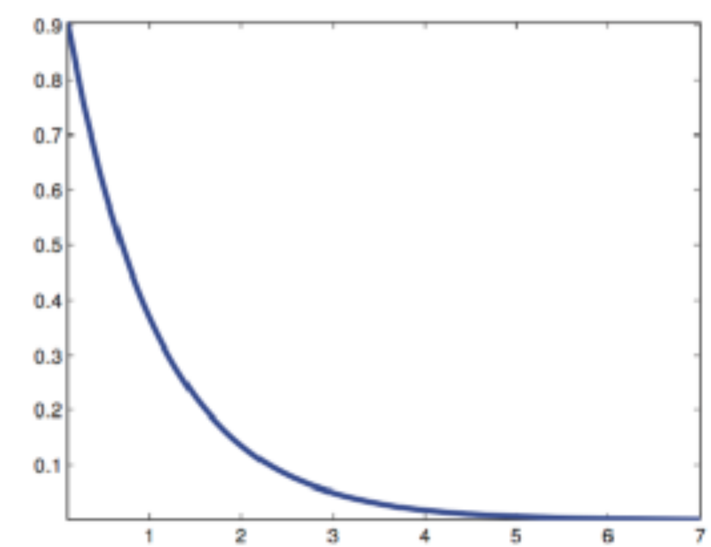
1. 没有衡量不确定性（uncertainty）
2. 导致结果过拟合（overfit）
3. 众数（mode）不具代表性（看图a、b）
4. 重新参数化后结果受影响*（看图c）



(c)



(a)



(b)

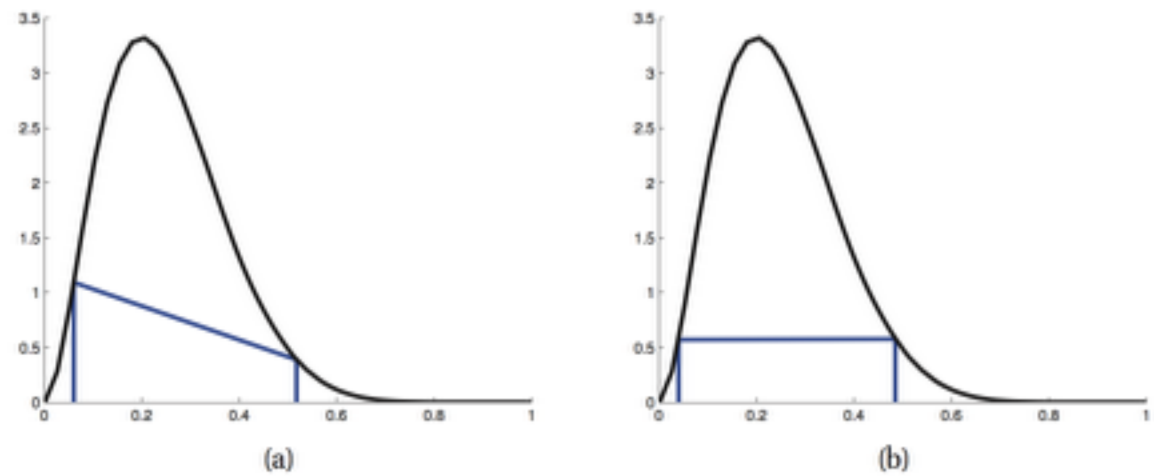
置信区间（Credible intervals）

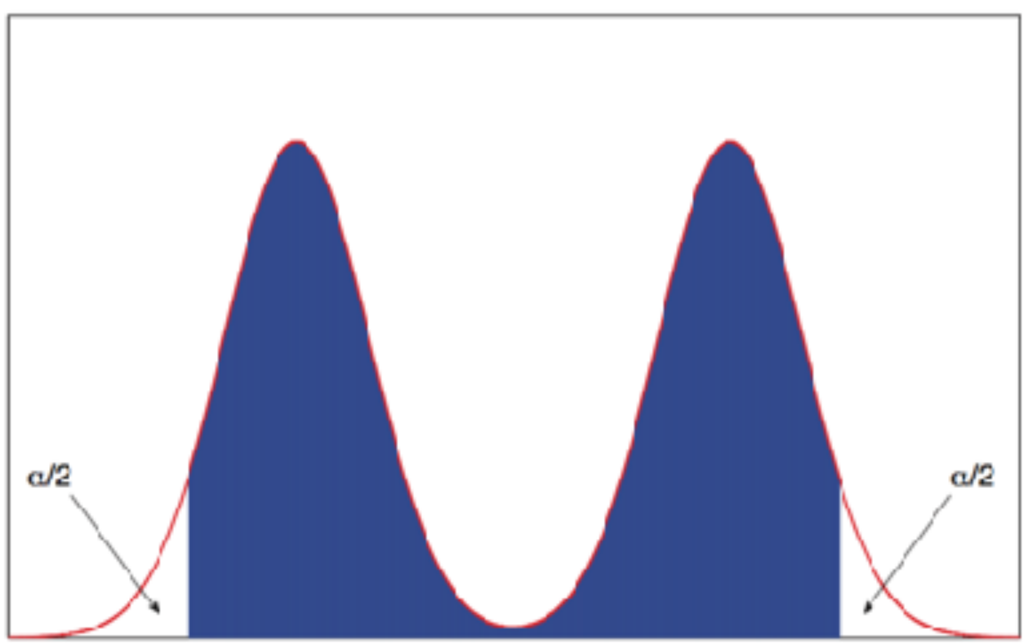
置信区间的定义就是包含 $1-\alpha$ 概率质量的区间：

$$C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \le \theta \le u|\mathcal{D}) = 1-\alpha$$

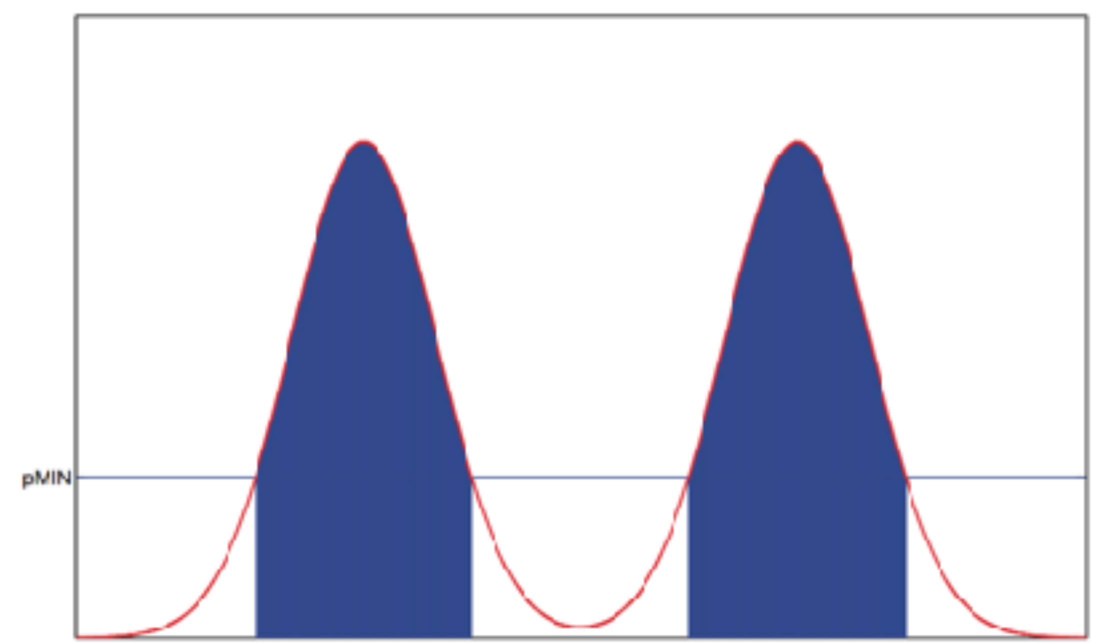
$$\ell = F^{-1}(\alpha/2)$$

$$u = F^{-1}(1-\alpha/2)$$





中心区间（Central interval）



最高密度区间（HDI）

不同比例时如何推论

举个例子：

卖家一有90个好评10个差评，卖家二有2个好评0个差评，你卖谁的?

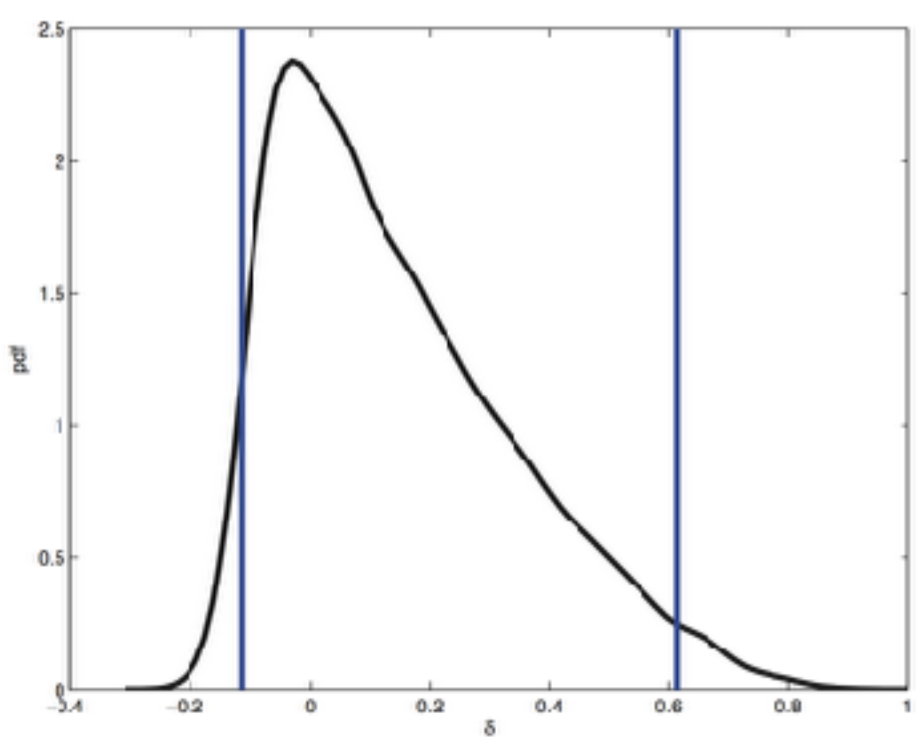
假设 $\theta_i \sim \mathrm{Beta}(1,1)$

得到 $p(\theta_1|\mathcal{D}_1) = \mathrm{Beta}(91,11)$ 和 $p(\theta_2|\mathcal{D}_2) = \mathrm{Beta}(3,1)$

求 $p(\delta > 0|\mathcal{D}) = \int_0^1 \int_0^1 \mathbb{I}(\theta_1 > \theta_2)\mathrm{Beta}(\theta_1|y_1+1, N_1 - y_1 + 1)\,\mathrm{Beta}(\theta_2|y_2+1, N_2 - y_2 + 1)d\theta_1 d\theta_2$

使用蒙特卡罗求解

$p(\delta > 0|\mathcal{D}) = 0.718$



贝叶斯模型选择
（Bayesian model selection）

我们如何选择较好的模型?

1.**交叉验证（cross-validation）**

2.**计算模型的后验** $p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m\in\mathcal{M}} p(m,\mathcal{D})}$

计算该后验需要计算**边缘似然(marginal likelihood、evidence)**

$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

贝叶斯奥卡姆剃刀
（Baysian Occam’s razor）

奥卡姆剃刀：《箴言书注》2卷15题说“切勿浪费较多东西去做，用较少的东西，同样可以做好的事情。”

贝叶斯奥卡姆剃刀：更多的参数可以拟合的更好，但模型有更多的参数不一定有更高的边缘似然。

$$p(\mathcal{D}) = p(y_1)p(y_2|y_1)p(y_3|y_{1:2})\ldots p(y_N|y_{1:N-1})$$

这个过程可以理解为，我们先计算模型生成 y1 的概率，然后乘以 y1 为训练集时 y2 的预测分布，依次类推。显然，如果一个模型过于复杂，那么预测分布值会较小（因为预测性能不好），那么在连乘后，得到的边缘似然也很小。

reference：知乎

边缘似然
（marginal likelihood）

1. 使用共轭先验可以计算，如下：

假设q是p未归一化的形

$$p(\boldsymbol{\theta}) = q(\boldsymbol{\theta})/Z_0$$
$$p(\mathcal{D}|\boldsymbol{\theta}) = q(\mathcal{D}|\boldsymbol{\theta})/Z_\ell$$
$$p(\boldsymbol{\theta}|\mathcal{D}) = q(\boldsymbol{\theta}|\mathcal{D})/Z_N$$

因此 $$q(\boldsymbol{\theta}|\mathcal{D}) = q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})$$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$
$$\frac{q(\boldsymbol{\theta}|\mathcal{D})}{Z_N} = \frac{q(\mathcal{D}|\boldsymbol{\theta})q(\boldsymbol{\theta})}{Z_\ell Z_0 p(\mathcal{D})}$$

所以 $$p(\mathcal{D}) = \frac{Z_N}{Z_0 Z_\ell}$$

边缘似然

（marginal likelihood）

2. 使用**贝叶斯信息准则（BIC、Bayesian information criterion）**可以简化计算，如下：

$$\mathrm{BIC} \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) - \frac{\mathrm{dof}(\hat{\boldsymbol{\theta}})}{2} \log N \approx \log p(\mathcal{D})$$

或者用**BIC-cost**来计算，如下：

$$\text{BIC-cost} \triangleq -2 \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}) + \mathrm{dof}(\hat{\boldsymbol{\theta}}) \log N \approx -2 \log p(\mathcal{D})$$

贝叶斯因子（Bayes factors）

假设M0是零假设（null hypothesis），M1是择一假设（alternative hypothesis）

贝叶斯因子定义为：

$$BF_{1,0} \triangleq \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} / \frac{p(M_1)}{p(M_0)}$$

当 $p(M_1) = p(M_0) = 0.5$ 时

$$p(M_0|\mathcal{D}) = \frac{BF_{0,1}}{1 + BF_{0,1}} = \frac{1}{BF_{1,0} + 1}$$

不提供信息的先验
（Uninformation prior）

1.Haldane prior 是一个improper prior，如下：

$$\lim_{c\to 0} \mathrm{Beta}(c,c) = \mathrm{Beta}(0,0)$$

2.Jeffreys prior 认为p(θ)不提供信息时，对其重新参数化后的先验也是不提供信息的

证明：

Fisher information的定义：$I(\phi) \triangleq -\mathbb{E}\left[\left(\frac{d\log p(X|\theta)}{d\theta}\right)^2\right]$

假设：$p_\phi(\phi) \propto (I(\phi))^{\frac{1}{2}}$

由于：$p_\theta(\theta) = p_\phi(\phi)|\frac{d\phi}{d\theta}|$

$$\frac{d\log p(x|\theta)}{d\theta} = \frac{d\log p(x|\phi)}{d\phi}\frac{d\phi}{d\theta}$$

所以：
$$I(\theta) = -\mathbb{E}\left[\left(\frac{d\log p(X|\theta)}{d\theta}\right)^2\right] = I(\phi)\left(\frac{d\phi}{d\theta}\right)^2$$

$$I(\theta)^{\frac{1}{2}} = I(\phi)^{\frac{1}{2}}|\frac{d\phi}{d\theta}|$$

$$p_\theta(\theta) = p_\phi(\phi)|\frac{d\phi}{d\theta}| \propto (I(\phi))^{\frac{1}{2}}|\frac{d\phi}{d\theta}| = I(\theta)^{\frac{1}{2}}$$

因此：$p_\theta(\theta)$ 和 $p_\phi(\phi)$ 是一样的

不提供信息的先验
（Uninformation prior）

3.Robust prior

使用具有厚尾的先验，避免后验受先验均值的影响，但是计算复杂度较高。

4.Mixtures of conjugate priors

可以逼近任何类型的先验，可以在计算复杂度和灵活性之间做权衡，如下：

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k)$$

其中 $p(\theta|z=k)$是第k个共轭先验，$p(z=k)$是每个共轭先验的权重。

层次贝叶斯（Hierarchical Bayes）

在计算 $p(\boldsymbol{\theta}|\mathcal{D})$ 时，需要用到先验 $p(\theta)$ 由于θ也有参数，所以假设θ的概率为 $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ 这样，就会形成一条链：

$$\boldsymbol{\eta} \to \boldsymbol{\theta} \to \mathcal{D}$$

这是一个二级的层次贝叶斯：

$$p(\boldsymbol{\eta}, \boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})$$

经验贝叶斯（Empirical Bayes）

在估计$\boldsymbol{\eta}$时，我们可以对$\boldsymbol{\eta}$使用点估计：

$$\hat{\boldsymbol{\eta}} = \operatorname{argmax} p(\boldsymbol{\eta}|\mathcal{D})$$

对$\boldsymbol{\eta}$使用均匀先验分布，$\boldsymbol{\eta}$的估计变为：

$$\hat{\boldsymbol{\eta}} = \operatorname{argmax} p(\mathcal{D}|\boldsymbol{\eta}) = \operatorname{argmax}\left[\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})d\boldsymbol{\theta}\right]$$

Method	Definition
Maximum likelihood	$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}\|\boldsymbol{\theta})$
MAP estimation	$\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})$
ML-II (Empirical Bayes)	$\hat{\boldsymbol{\eta}} = \operatorname{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})d\boldsymbol{\theta} = \operatorname{argmax}_{\boldsymbol{\eta}} p(\mathcal{D}\|\boldsymbol{\eta})$
MAP-II	$\hat{\boldsymbol{\eta}} = \operatorname{argmax}_{\boldsymbol{\eta}} \int p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})d\boldsymbol{\theta} = \operatorname{argmax}_{\boldsymbol{\eta}} p(\mathcal{D}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$
Full Bayes	$p(\boldsymbol{\theta}, \boldsymbol{\eta}\|\mathcal{D}) \propto p(\mathcal{D}\|\boldsymbol{\theta})p(\boldsymbol{\theta}\|\boldsymbol{\eta})p(\boldsymbol{\eta})$

贝叶斯决策理论
(Bayesian decision theory)

假设我们观测到 $\mathbf{x} \in \mathcal{X}$，其代表着未知的一个状态、参数或标签 $y \in \mathcal{Y}$ 然而，我们为了做出决策，需要预测 $a \in \mathcal{A}$ ，所以我们需要引入损失函数来衡量a和y的吻合程度，**损失函数（loss）**表示为 $L(y, a)$ 。

决策过程（decision procedure）：

$$\delta(\mathbf{x}) = \operatorname*{argmin}_{a \in \mathcal{A}} \mathbb{E}\left[L(y, a)\right]$$

贝叶斯估计（Bayes estimator ）：

$$\delta(\mathbf{x}) = \arg\min_{a \in \mathcal{A}} \sum_{y} L(y, a) p(y|\mathbf{x})$$

常见的损失函数(loss)

0-1损失：最好估计是后验的**众数**（**mode**）

$$L(y,a) = \mathbb{I}(y \neq a) = \begin{cases} 0 & \text{if } a = y \\ 1 & \text{if } a \neq y \end{cases}$$

均方误差:最好估计是后验的**平均数**（**mean**）

$$L(y,a) = (y-a)^2$$

绝对误差：最好的估计是后验的**中位数**（**median**）

$$L(y,a) = |y-a|$$

二分类决策问题

在二分类问题中，有四种情况，把**正类预测为正类（TP）**，把**负类预测为正类（FP）**，把**负类预测为负类（TN）**，把**正类预测为负类（FN）**。

		Truth	
		1	0
Estimate	1	TP	FP
	0	FN	TN

准确率（Accuracy）： $A = \frac{TP+TN}{TP+TN+FP+FN}$

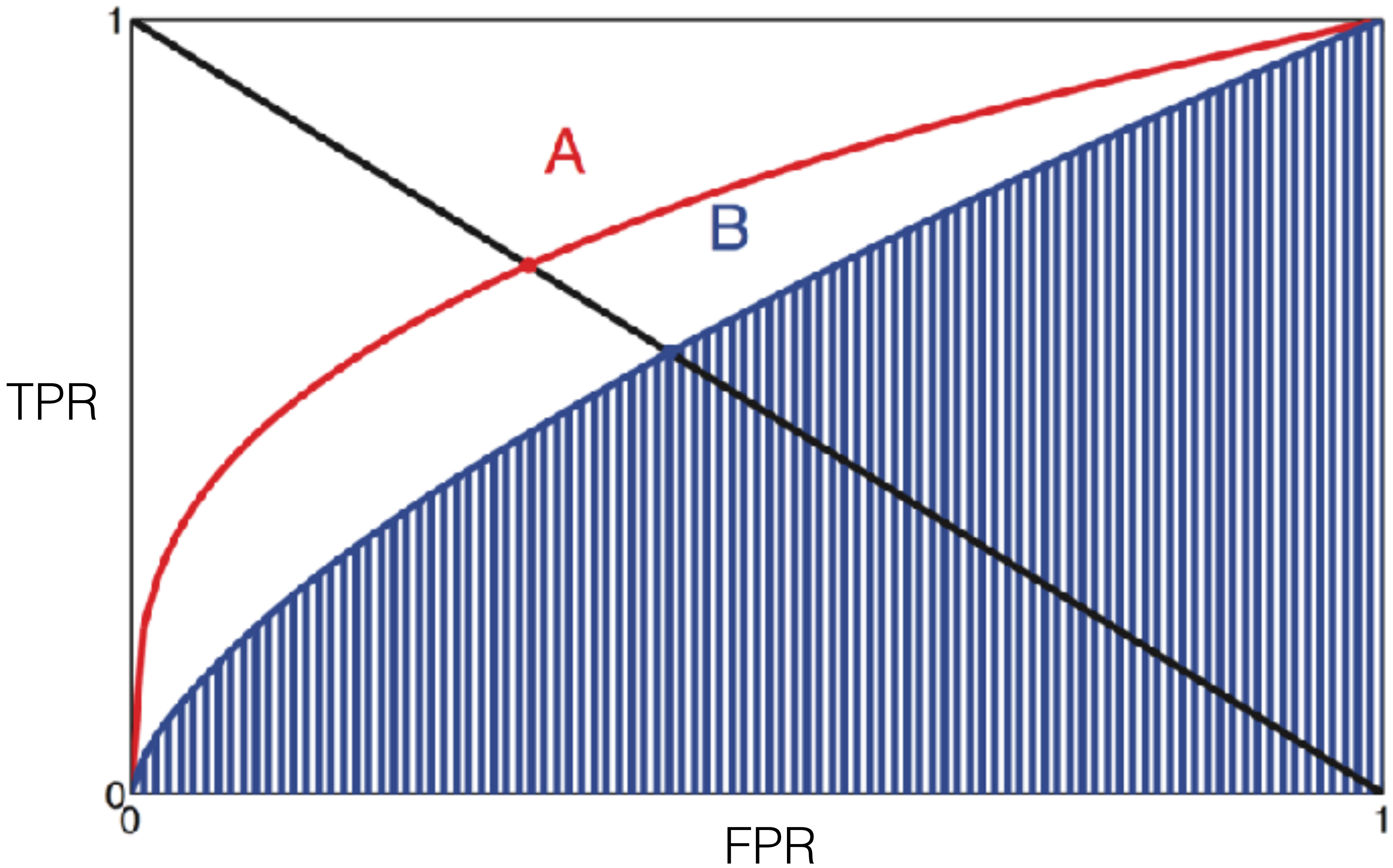
精确率（Precision）： $P = \frac{TP}{TP+FP}$

召回率（Precision）： $R = \frac{TP}{TP+FN}$

FPR（Precision）： $FPR = \frac{FP}{FP+TN}$

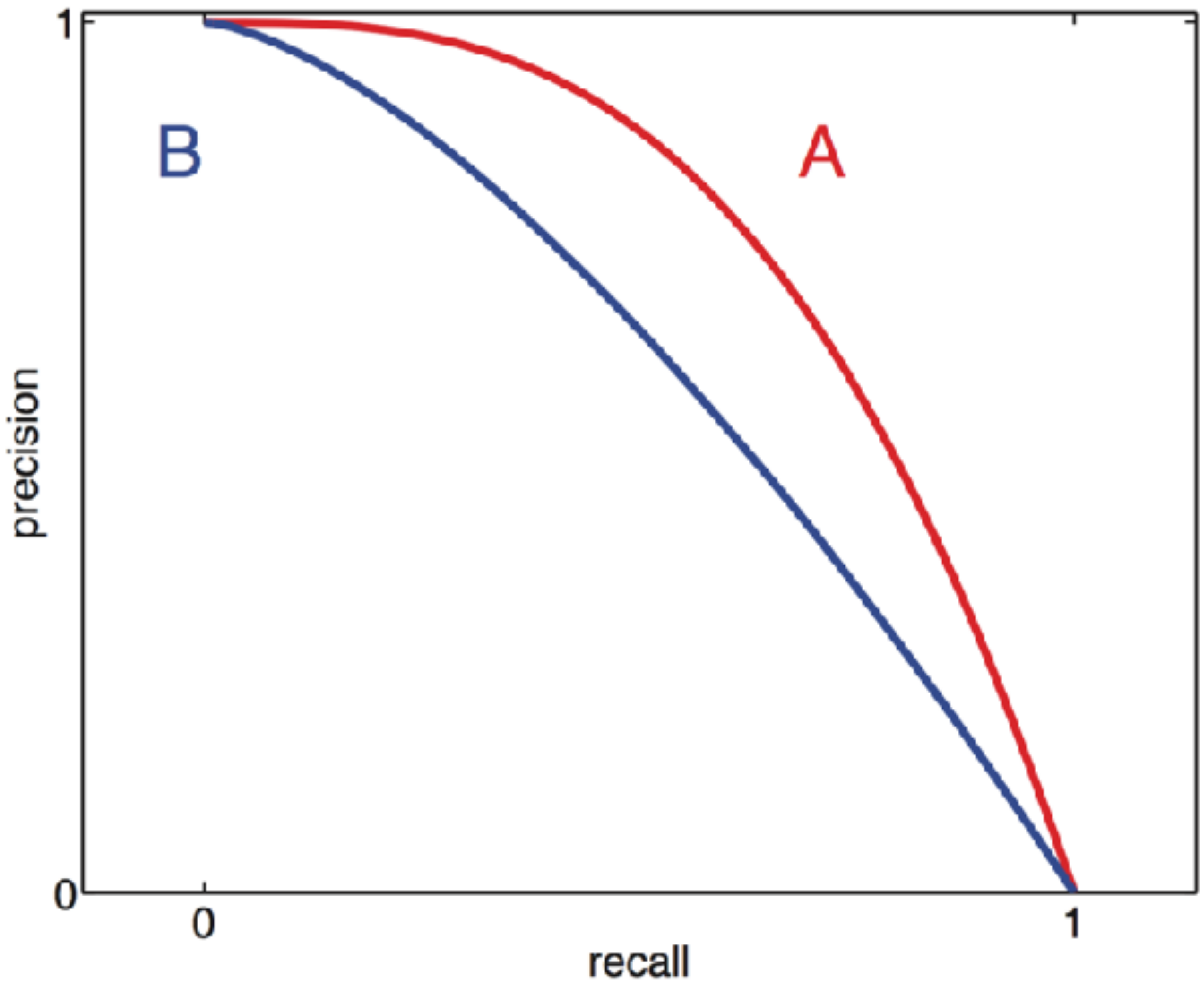
F1： $F_1 \triangleq \frac{2}{1/P + 1/R} = \frac{2PR}{R+P}$

ROC曲线



曲线下方的面积，称为**AUC（area under the curve ）**

PR曲线



Summary

1.贝叶斯派vs频率派：两者之间究竟有何不同

2.后验分布：
- MAP估计的缺点
- 置信区间
- 不同比例的推论

3.模型选择:
- 贝叶斯奥卡姆刀
- 边缘似然
- 贝叶斯因子

4.不提供信息的先验：Haldane先验，Jeffreys先验

5.层次贝叶斯

6.经验贝叶斯

7.贝叶斯决策理论：
- 损失函数
- 二分类决策

Thank you!