机器学习

3. MLE&MAP

3. MLE & MAP

- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

3. MLE & MAP

- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

基本概念

- 概率
  - 采样空间，事件，σ代数
  - 概率公理，概率度量
  - 随机变量
  - 期望与方差
  - 联合分布
  - 条件分布
  - 独立，条件独立

基本概念

- 采样空间

定义: 一个采样空间Ω对应一个(抽象概念或具有物理实体的)随机实验所有可能的输出. Ω可包含有限或无限元素.

■ 举例:

- 一个骰子的所有可能输出
- 一本书随机打开的页数
- 温度，坐标，时间等

基本概念

- 事件

定义：事件 A 是采样空间Ω的一个子集。

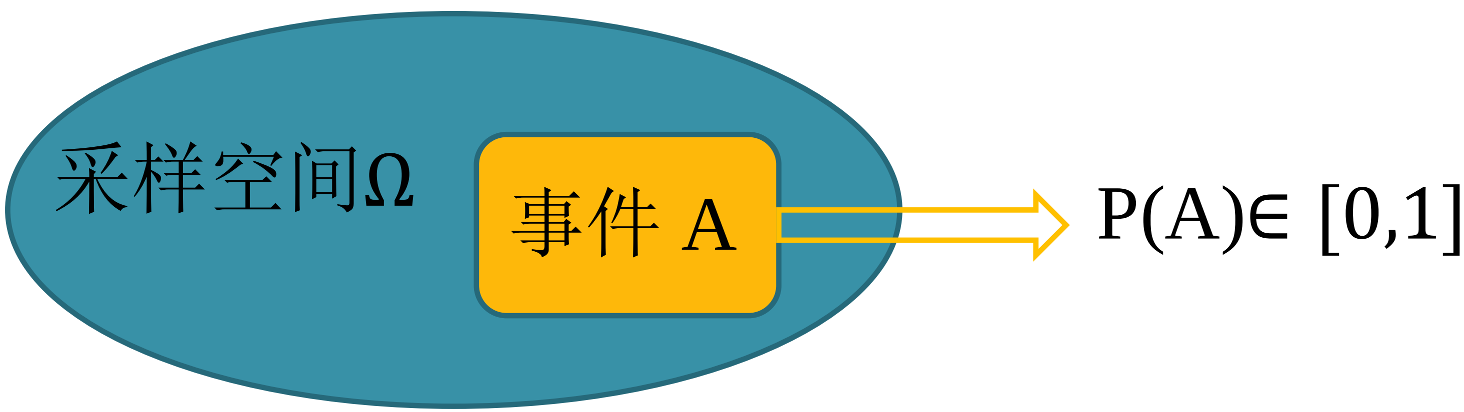
- 举例：
  - 骰子6点
  - 一本书在整10页打开
  - 中国人身高超过1米8

基本概念

➢ **事件概率**

定义：概率 P(A),又称事件 A 发生的概率，是一个将事件 A 映射到区间[0,1]的映射函数。P(A)又成为 A 的概率度量。

- **举例**：
  - ➢**骰子摇到2-4点的概率**



基本概念

- 可行事件集 = σ 代数

定义:Ω的一系列子集的集合，记为 M，可称为σ代数，若其满足以下三个条件：

(i) $\phi \in M$，即空集为 M 元素

(ii) 若 $A \in M$，则$A^c \in M$，即补集运算在 M 中为封闭运算

(iii) 若$A_1, A_2, \cdots \in M$，则$\bigcup_{i=1}^{\infty} A_i \in M$，即 M 在可列并运算下封闭

基本概念

- **概率公理 （Kolmogorov公理）**

(i) 非负性：对任意事件 A，$P(A) \geq 0$

(ii) $P(\Omega) = 1$

(iii) σ可加：对任意不相交事件A_i，我们有

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

- **一些推论**

$P(\phi) = 0$

$P(A^c) = 1 - P(A)$

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

基本概念

- 随机变量

定义:随机变量对应于一个由采样空间产生的事件映射到实数(或整数)空间的函数，即

$$X: \Omega \rightarrow R$$

$$P(a < X < b) = P(w: a < X(w) < b)$$
$$P(X = a) = P(w: X(w) = a)$$

- 例如：
  - X(w) = 1，我们班(Ω)上课第一个睡着的同学(w)是男生
  - X(w)<60，在可能获得的成绩(Ω)里，最后不及格的分数(w)

基本概念

- 离散分布
  - 伯努利分布: Ber(p)

$$\Omega = \{正面，反面\}，X(正面) = 1，X(反面) = 0$$

$$P(X=a) = P(w{:}\,X(w)=a) = \begin{cases} p, & 若\ a=1 \\ 1-p, & 若\ a=0 \end{cases}$$

  - 二项分布: Bin(n,p)

假设一个硬币正面朝上的概率为p，抛此硬币 n 次，得到 k 次正面（或 n-k 次反面）的概率多大？

$$\Omega = \{所有可能\ n\ 长度的正面/反面序列\}，|\Omega| = 2^n$$

$$w = (w_1, w_2, \cdots, w_n) \in \{正面,反面\}^n, K(2) = w\ 中出现正面的次数$$

$$P(K=k) = P(w{:}\,K(w)=k) = \sum_{w:K(w)=k} p^k(1-p)^{n-k} = \binom{n}{k} p^k (1-p)^{n-k}$$

基本概念

- 连续分布

Cumulative distribution function (CDF)

定义:累积分布函数:

$F_X(z) = P(X \leq z)$ or $F_X(z) = P(X < z)$

定义:连续分布：累积分布函数为绝对连续的随机变量概率

$F(-\infty) = 0$，$F(+\infty) = 1$

当存在某个函数p使得$F(x) = \int_{-\infty}^{x} p(t)dt$,则$F: (-\infty, \infty) \rightarrow R$为绝对连续。

Probability Density function (PDF)

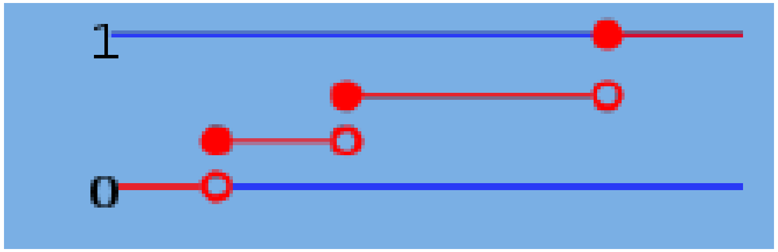
定义:上述 p(x)成为分布 F 的概率密度函数

性质：$\frac{d}{dx}F(x) = p(x)$.

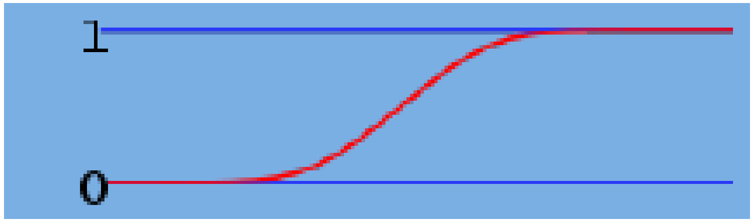
基本概念

- 累积分布图

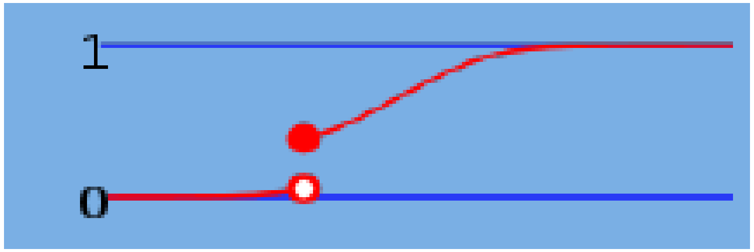
?



?



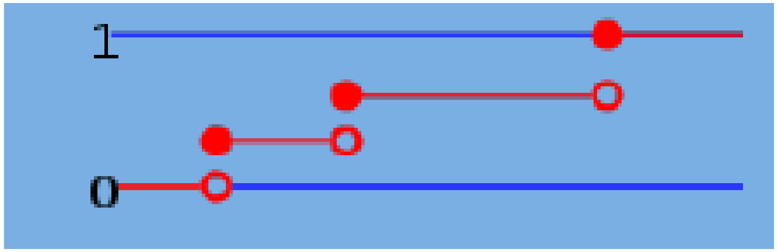
?



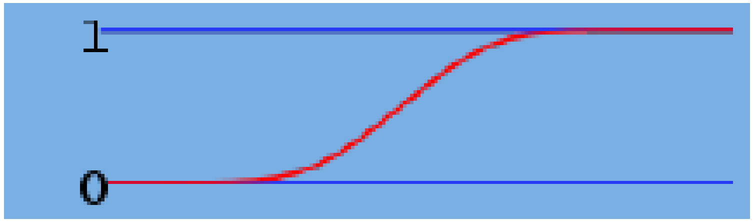
基本概念

➢ 累积分布图

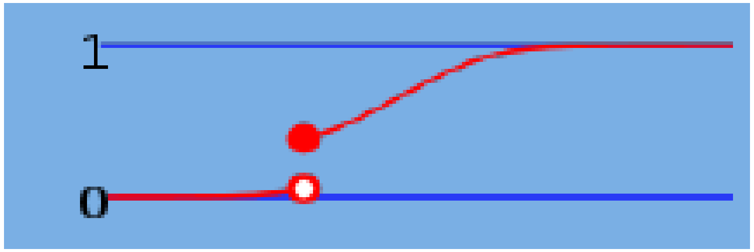
离散概率分布的CDF



连续概率分布的CDF



连续+离散概率分布的CDF



基本概念

- 概率密度分布

PDF 性质：

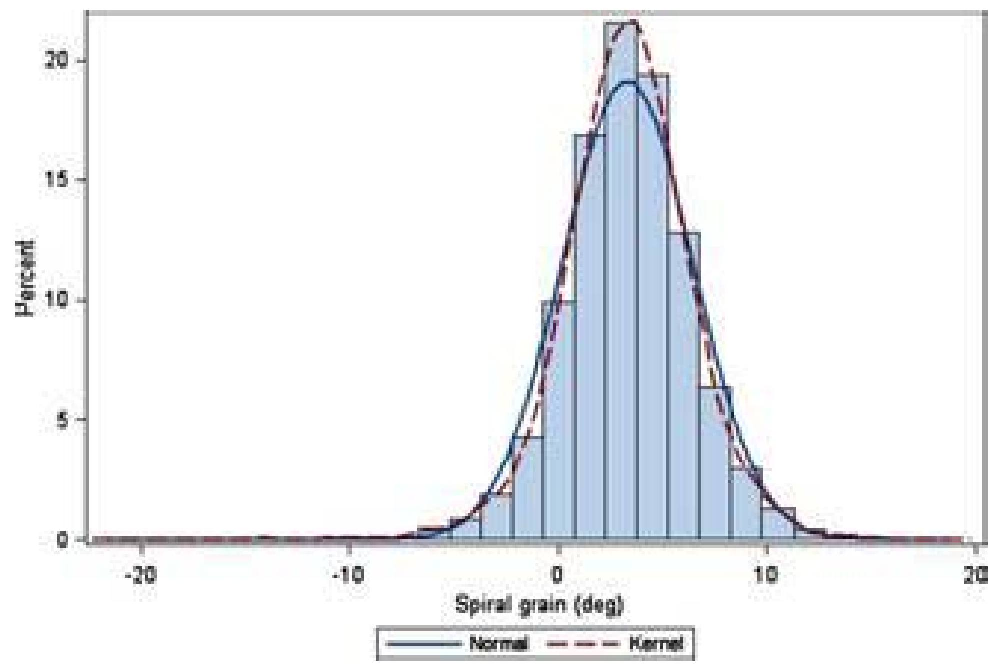
$p(x) \geq 0$

$\int_{-\infty}^{\infty} p(x)dx = 1$

$p(x) = \frac{d}{dx} dF(x)$

$F(x) = \int_{-\infty}^{x} p(t)dt$

$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx$

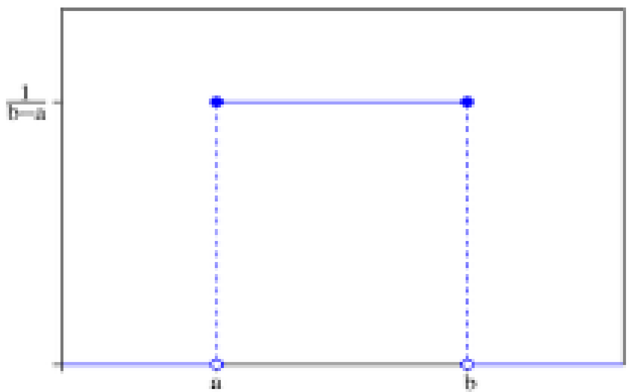


- 直观上，可认为p(x)dx为随机变量 X 落于无穷小区间[x,x+dx]的概率

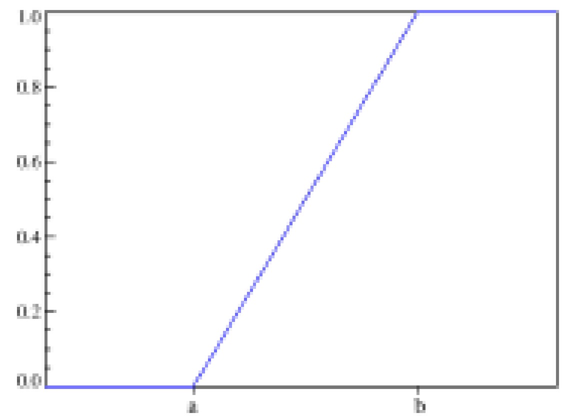
基本概念

Uniform Distribution

- 均匀分布



PDF



CDF

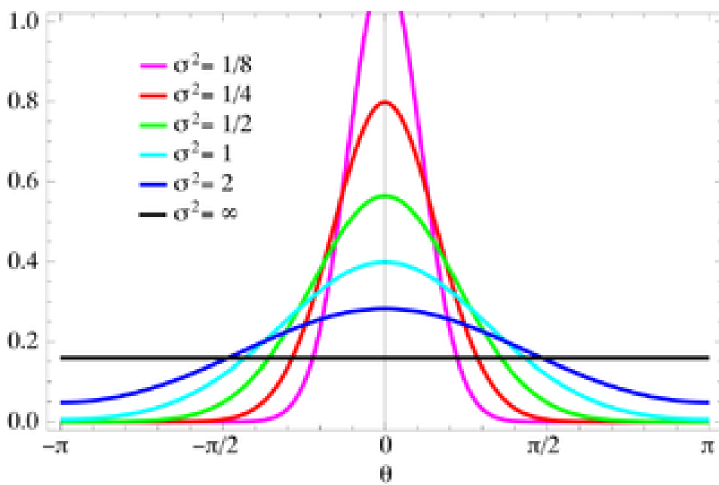
$$p(X)=\begin{cases}1/(b-a), & a\le x\le b\\ 0, & 其它\end{cases}$$

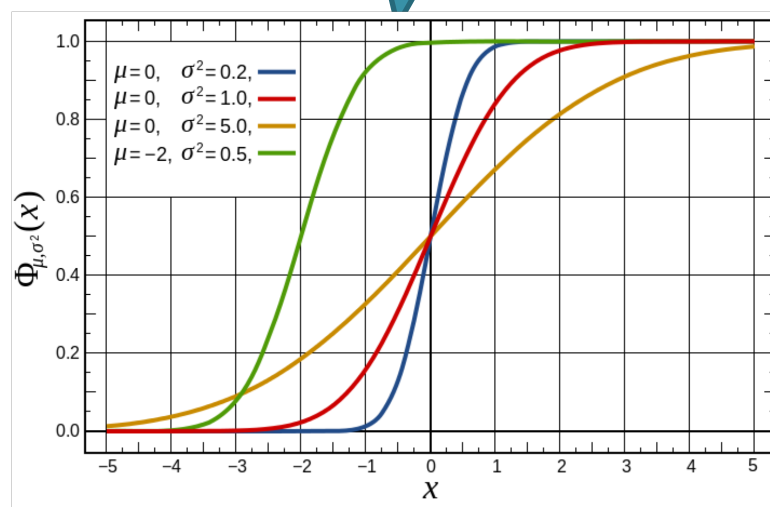
$$F(X)=\begin{cases}0, x\le a\\ (x-a)/(b-a), & a\le x\le b\\ 1, & x>b\end{cases}$$

基本概念

Normal/ Gaussian Distribution

➢ 正态/高斯分布





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

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$$

基本概念

- 矩
  - 期望：平均值，1阶矩

Expectation

$$E(X)=\begin{cases}\sum_{x_i\in\Omega} x_i p(x_i) & \text{，离散概率分布}\\ \int_{-\infty}^{\infty} xp(x)dx & \text{，连续概率分布}\end{cases}$$

  - 方差：平均值，2阶矩

Variance

$$V(X)=\begin{cases}\sum_{x_i\in\Omega}(x_i-E(X))^2 p(x_i) & \text{，离散概率分布}\\ \int_{-\infty}^{\infty}(x-E(X))^2 p(x)dx & \text{，连续概率分布}\end{cases}$$

基本概念

- 多变量（联合）分布

$$P(a \le X \le b, c \le Y \le d) = ?$$

$$P(a_1 \le X_1 \le b_1, \cdots, a_d \le X_d \le b_d) = ?$$

基本概念

- 多变量（联合）分布

  $P(a \le X \le b, c \le Y \le d) = ?$

  $P(a_1 \le X_1 \le b_1, \cdots, a_d \le X_d \le b_d) = ?$

- 离散多变量分布

  P(A1,B1) = 5/20

  边际分布：P(A1) = 16/20, P(B1) = 8/20

Marginal distribution

	B1坐前五排	B2不坐前五排
A1男生	5/20	11/20
A2女生	3/20	1/20

基本概念

- 连续多变量分布

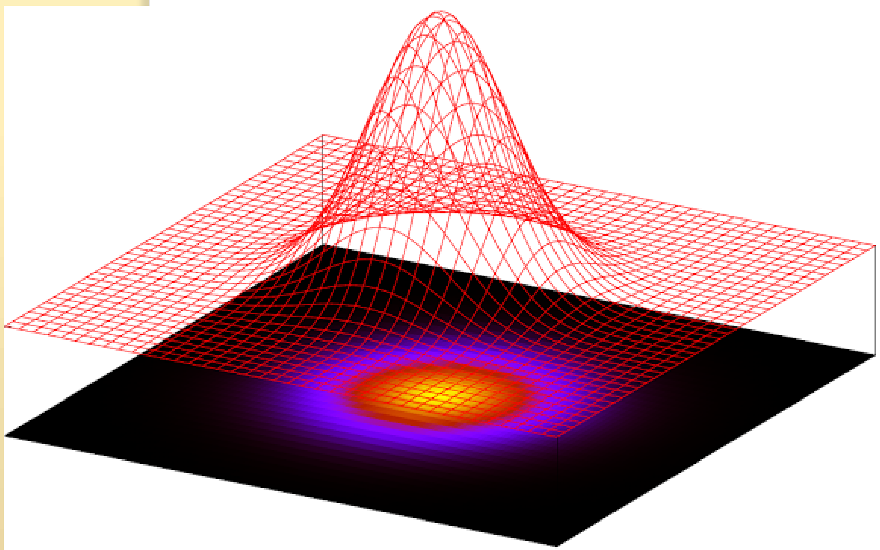
多变量累积分布函数

对于$A \subset R^d$,$P([X_1,\cdots,X_d] \in A) = \int_A p(t_1,\cdots,t_d)dt_1 \cdots dt_d$

$F_X(x_1,\cdots,x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} p(t_1,\cdots,t_d)dt_1 \cdots dt_d$

基本概念

- 多变量高斯分布



$\mu \in R^d$: 均值

$\Sigma \in R^{d\times d}$:协方差矩阵

$$p(x_1,\cdots,x_d) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

基本概念

Conditional probability

- 条件概率

P(X|Y): 在给定Y事件为真的前提下X事件为真的概率

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

P(B1|A1)

P(B1|A2)

	B1坐前五排	B2不坐前五排
A1男生	5/20	11/20
A2女生	3/20	1/20

基本概念

Conditional probability

- 条件概率

P(X|Y): 在给定Y事件为真的前提下X事件为真的概率

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

P(B1|A1) = (5/20) / (16/20) = 5/16

P(B1|A2) = (3/20) / (4/20) = 3/4

	B1坐前五排	B2不坐前五排
A1男生	5/20	11/20
A2女生	3/20	1/20

基本概念

Independence

- 独立性

独立随机变量：

$$\boxed{P(X,Y)=P(X)P(Y);\ \ P(X|Y)=P(X)}$$

含义：

- ✓Y与X不包含互相的信息
- ✓观察到事件Y不能帮助预测X事件信息
- ✓观察到事件X不能帮助预测Y事件信息

例子：

- ✓ 学校举办的报告
- ✓ 一门课程的某一节课

基本概念

- 条件独立

Z使得X与Y独立：

$$P(X,Y|Z) = P(X|Z)P(Y|Z)$$
$$P(X|Y,Z) = P(X|Z)$$

例子：

- ✓ 相关：鞋码大小与读书能力

基本概念

- 条件独立

Z使得X与Y独立：

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$
$$P(X|Y, Z) = P(X|Z)$$

例子：

- ✓ 相关：鞋码大小与读书能力
- ✓ 条件独立：在给定年龄下的鞋码大小与读书能力
- ✓ 伦敦出租车司机：
    - ✓ 调查指出交通事故的发生概率与出租车司机是否穿外衣有强正相关关系。

3. MLE & MAP

- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

贝叶斯规则

- 链式法则

$$P(X,Y) = P(X|Y)P(Y) = P(Y|X)P(X)$$

- 贝叶斯规则

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- 贝叶斯规则构成了整个机器学习的重要统计理论基础！

贝叶斯规则

观测到飞碟|有外星人	观测到飞碟|无外星人
0.6	0.1
0.4	0.9
未观测到飞碟|有外星人	未观测到飞碟|无外星人

有外星人	0.001
无外星人	0.999

- 当观测到飞碟有外星人的概率：

A=1：观测到飞碟

B=1：有外星人

$$P(B=1|A=1)=\frac{P(A=1|B=1)P(B=1)}{P(A=1)}$$

$$=\frac{P(A=1|B=1)P(B=1)}{P(A=1|B=1)P(B=1)+P(A=1|B=0)P(B=0)}$$

$$=\frac{0.6*0.001}{0.6*0.001+0.1*0.999}=0.057$$

- 当观测到飞碟无外星人的概率：

$$P(B=0|A=1)=1-0.057=0.943$$

贝叶斯规则

观测到飞碟|有外星人	观测到飞碟|无外星人
0.6	0.1
0.4	0.9
未观测到飞碟|有外星人	未观测到飞碟|无外星人

有外星人	0.001
无外星人	0.999

- 观察1
  - 即使观测到飞碟，有外星人的概率仍然很小
  - 原因？
- 观察2
  - 观测到飞碟使有外星人的先验概率显著增大
  - 原因？

贝叶斯规则

先验 vs. 后验

3. MLE & MAP

- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

最大似然估计

Maximum Likelihood Estimation

- 硬币实验
  - 抛一个硬币，正面朝上的概率?
  - 以下的抛投结果:

  - 正面朝上的概率是
  - 3/5
  - 为什么?

最大似然估计

- 伯努利分布

$$D=\{X_i\}_{i=1}^{n}，\ X_i\in\{正面，\ 背面\}$$

$$P(正面)=\theta，\ P(背面)=1-\theta$$

- 所有抛掷为i.i.d.事件
  - 独立(independent)事件
  - 同分布(identically distributed)事件
- 最大似然估计目标：选择 θ 使得观察数据的发生概率最大！

最大似然估计

- 最大似然函数估计（MLE）：

$$D = \{X_i\}_{i=1}^{n}，\ X_i \in \{正面，背面\}$$
$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- 计算使得观察数据的发生概率最大的 θ

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$
$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta) = ?$$

最大似然估计

- 计算使得观察数据的发生概率最大的 θ

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(X_i|\theta)$$

$$\theta_{MLE} = \frac{n_{正面}}{n_{正面} + n_{背面}}$$

最大似然估计

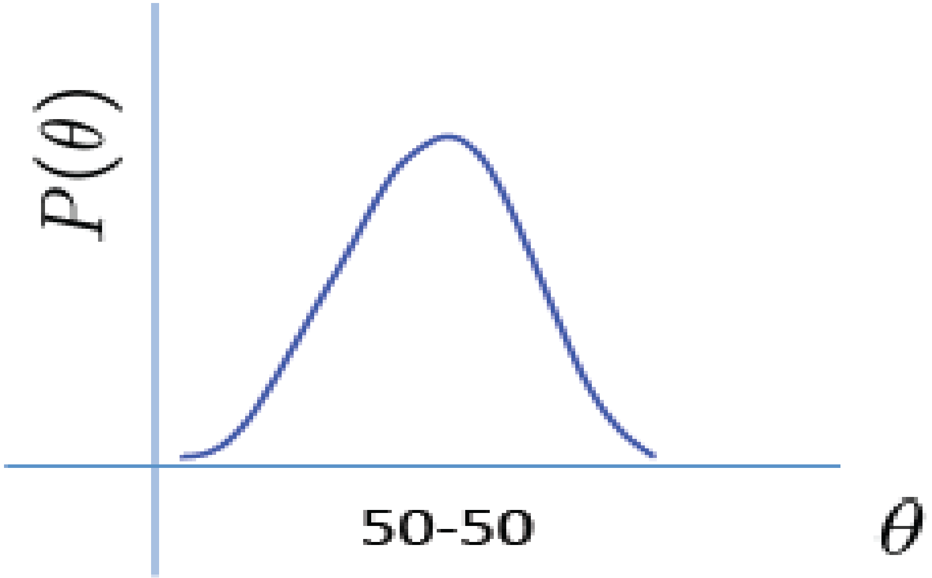
- 一个问题：如果抛投五次硬币，全部为正面，MLE结果是?

最大似然估计

- 一个问题：如果抛投五次硬币，全部为正面，MLE结果是?
- P(正面) = 1 对吗?

最大似然估计

- 一个问题：如果抛投五次硬币，全部为正面，MLE结果是?
- P(正面) = 1 对吗?
- 我们默认有一个先验在脑中



3. MLE & MAP

- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

最大后验估计

Maximum A Posteriori Estimation

- 贝叶斯规则

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

- 或者。。。

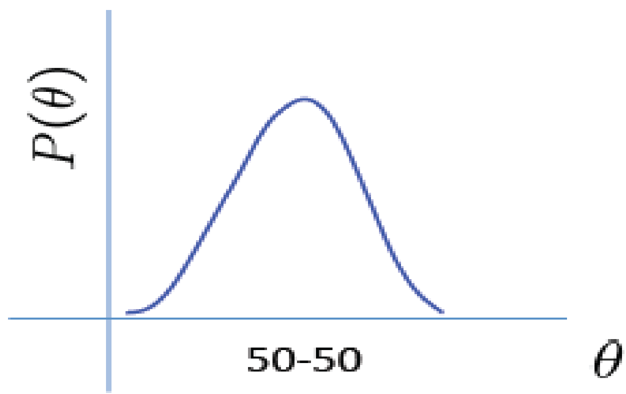
$$P(\theta|D) \sim P(D|\theta)P(\theta)$$

后验分布　　似然函数　　先验分布

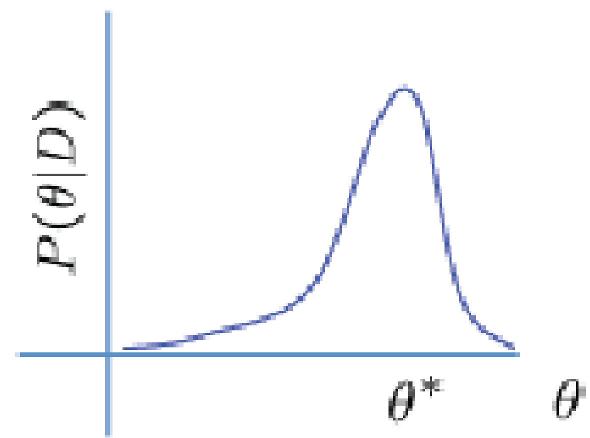
最大后验估计

- 硬币实验

$$P(\theta|D) \sim P(D|\theta)P(\theta)$$



先验
分布



后验
分布

先验从哪里来？？？

- 关于先验
  - 通过专家知识而来（哲学系方法）
  - 简单易算形式（工程系方法）
- 无信息先验
  - 均匀分布
  - 近似均匀分布
- 共轭先验
  - 后验能够闭合形式表达
  - 先验与后验具有同样形式

最大后验估计

- 共轭先验
  - 先验与后验具有相同的参数化形式

例 1：抛掷硬币

似然函数：二项分布

$$P(D|\theta) = C_n^{n_Z}\theta^{n_Z}(1-\theta)^{n-n_Z}$$

Beta 分布(先验)：

$$P(\theta) = \frac{\theta^{\beta_Z-1}(1-\theta)^{\beta_B-1}}{B(\beta_Z,\beta_B)} \sim Beta(\beta_Z,\beta_B)$$

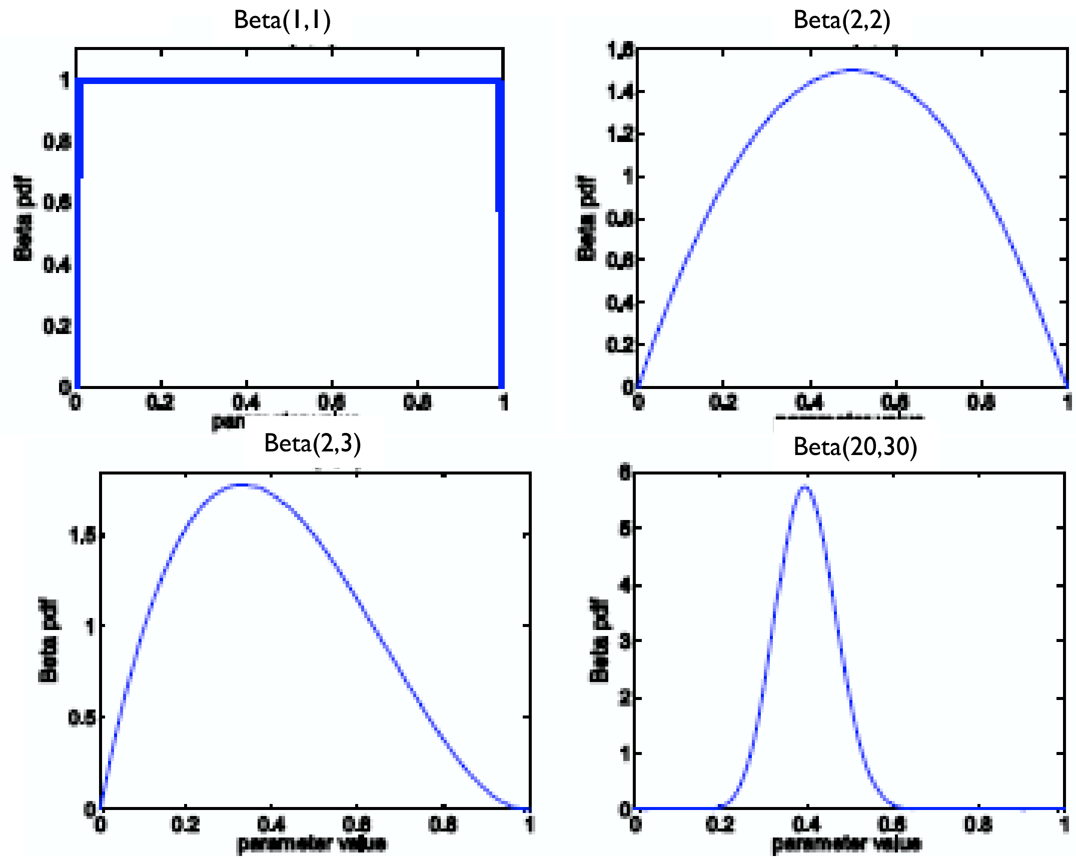
后验形式：

$$P(\theta|D) \sim Beta(\beta_Z+n_Z,\beta_B+n_B)$$

- 二项分布的共轭先验分布为Beta分布

最大后验估计

- 典型Beta分布图



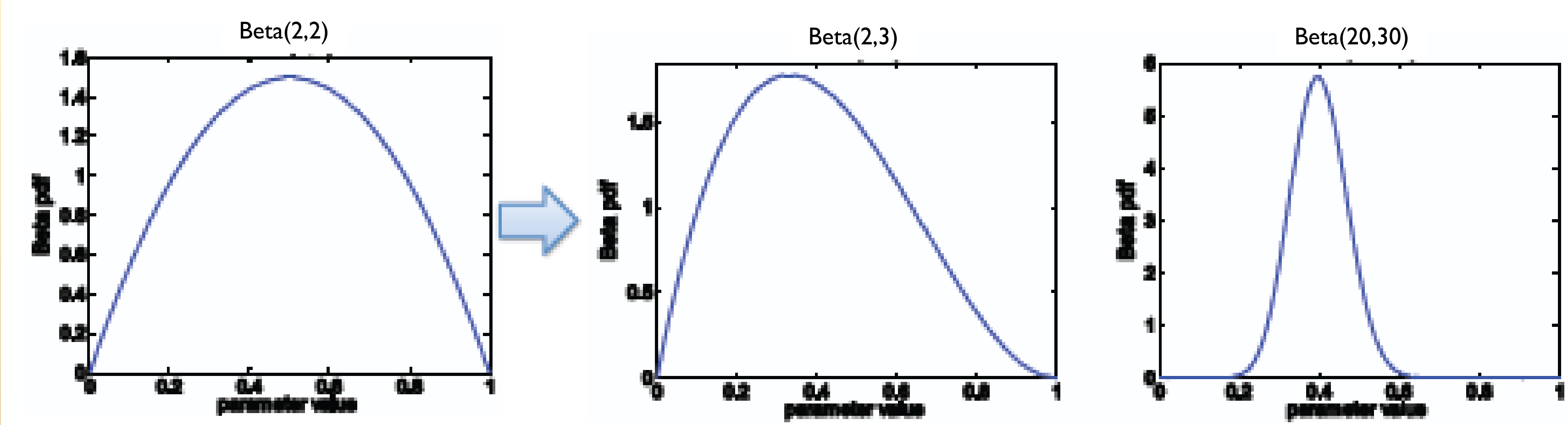
- β_Z,β_B越大，分布越集中

最大后验估计

- Beta先验计算效果

$P(\theta) \sim Beta(\beta_Z, \beta_B)$

$P(\theta|D) \sim Beta(\beta_Z + n_Z, \beta_B + n_B)$



- 直观意义？
- 当越来越多的样本加入时，先验的作用会？

最大后验估计

- 最大后验原理

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|D)$$

$$= \arg\max_{\theta} P(D|\theta)P(\theta)$$

- 对二项分布的MAP估计：

$$P(\theta|D) \sim Beta(\beta_Z + n_Z, \beta_B + n_B)$$

$$\theta_{MAP} = \frac{\beta_Z + n_Z - 1}{\beta_Z + n_Z + \beta_B + n_B - 2}$$

- 与其MLE相比有什么样的不同?

3. MLE & MAP

- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

MLE vs. MAP

- MLE:
  选择最大化观察数据概率的参数值

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$

- MAP:
  在给定观察数据与先验前提下，选择后验最大的参数值

$$\theta_{MAP} = \arg\max_{\theta} P(\theta|D)$$
$$= \arg\max_{\theta} P(D|\theta)P(\theta)$$

MLE vs. MAP

$$\theta_{MLE} = \frac{n_Z}{n_Z + n_B}$$

- 三次正面，结果为?
- MLE估计：正面概率为1

$$\theta_{MAP} = \frac{\beta_Z + n_Z - 1}{\beta_Z + n_Z + \beta_B + n_B - 2}$$

- MAP估计：$$\frac{(\beta_Z - 1) + 3}{(\beta_Z + \beta_B - 2) + 3}$$

- 总投掷次数n趋于无穷时，会发生什么情况?
- 何时先验更能体现出重要性?

MLE vs. MAP

贝叶斯学派:
在小数据时
你做的都是
错的!

频域学派:
你太依赖先验,
而且先验不同,
结果也不同!

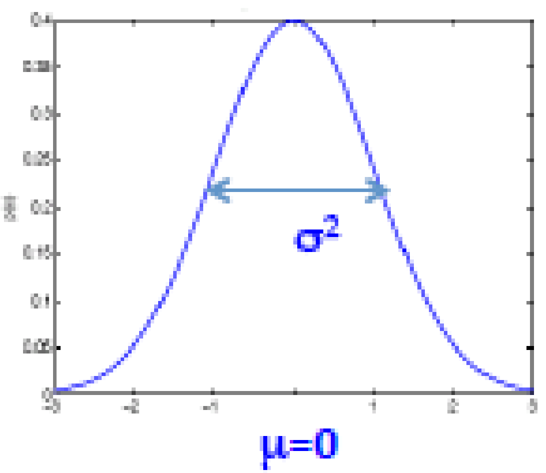
3. MLE & MAP

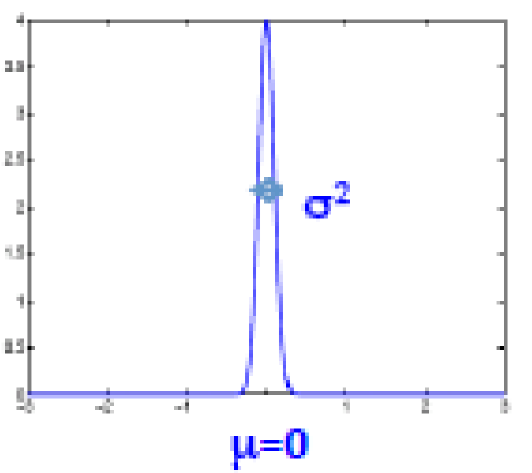
- 统计/概率 基本概念与知识
- 贝叶斯准则
- 最大似然估计 (MLE)
- 最大后验估计 (MAP)
- MLE VS. MAP
- 高斯分布情形

高斯分布情形

- 高斯分布的MLE与MAP估计如何求?

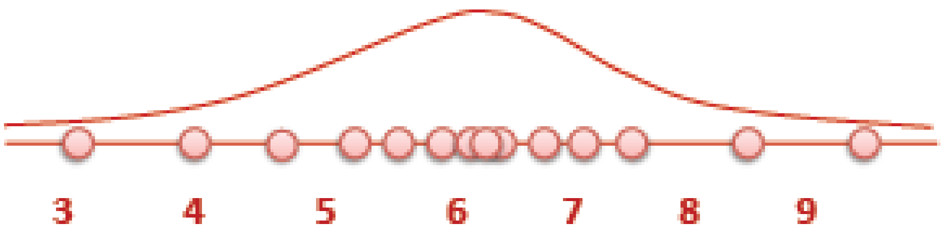
$$P(x|\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}=N(\mu,\sigma)$$





高斯分布情形

- 数据:



- 参数: μ - 均值, σ^2 - 方差

- 数据为i.i.d.:
  - 独立事件 independent
  - 根据同一个高斯分布产生 identically distributed

高斯分布情形

- 高斯分布的性质
  - 仿射变换的闭合性

$$x \sim N(\mu, \sigma^2)$$

$$y = ax + b \sim N(a\mu + b, a^2\sigma^2)$$

  - 高斯分布的加和

$$x \sim N(\mu_x, {\sigma_x}^2)$$

$$y \sim N(\mu_y, {\sigma_y}^2)$$

$$z = x + y \sim N(\mu_x + \mu_y, {\sigma_x}^2 + {\sigma_y}^2)$$

高斯分布情形

- 高斯分布参数的MLE估计

$$\theta_{MLE} = \arg\max_{\theta} P(D|\theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} P(x_i|\theta)$$

$$= \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

$$= \arg\max_{\theta=(\mu,\sigma^2)} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x_i-\mu)^2}{2\sigma^2}}$$

高斯分布情形

- 高斯分布参数的MLE估计

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma_{MLE}{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_{MLE})^2$$

高斯分布情形

- 高斯分布的共轭先验
  - 均值变量：高斯分布
  - 方差变量：Wishart分布
- 高斯先验形式

$$P(\mu|\eta,\lambda)=\frac{1}{\lambda\sqrt{2\pi}}e^{\frac{-(\mu-\eta)^2}{2\lambda^2}}=N(\eta,\lambda^2)$$

高斯分布情形

- 高斯分布均值变量的MLE与MAP估计

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu_{MAP} = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i + \frac{\eta}{\lambda^2}}{\frac{n}{\sigma^2} + \frac{1}{\lambda^2}}$$

要求

1. 搞清楚机器学习涉及的统计基础
2. 搞清楚贝叶斯定理的含义
3. 搞清楚MLE与MAP各自的特点与本质

阅读：

[1] Pattern Recognition and Machine Learning, Christopher , M. Bishop, Springer, 2006. II. Probabilistic Distribution