

# 模式识别引论

An Introduction to Pattern Recognition

主讲: 李春光

[www.pris.net.cn/teacher/lichunguang](http://www.pris.net.cn/teacher/lichunguang)

模式识别与智能系统实验室

网络搜索教研中心 信息与通信工程学院 北京邮电大学

# 部分数学基础内容回顾

- 概率论
- 决策论
- 信息论

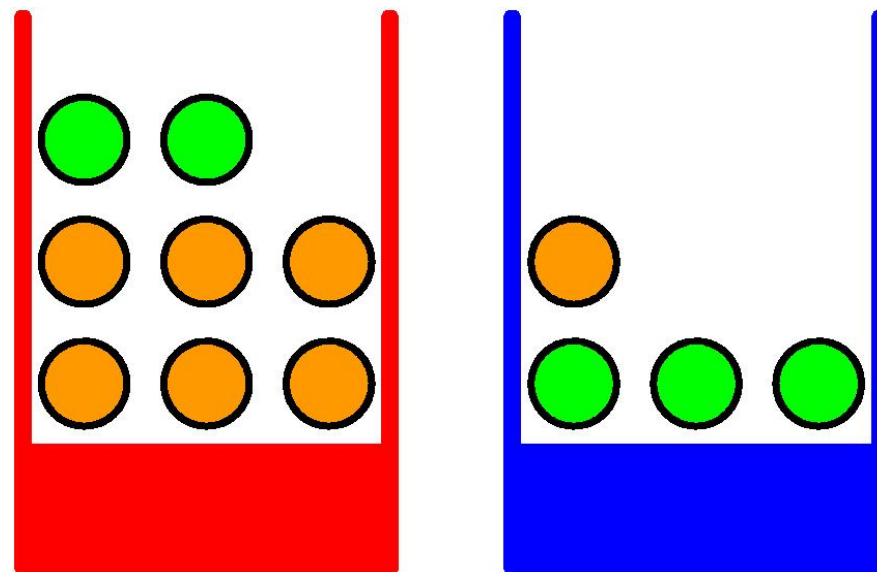
# 概率论

- 基本概念
- 贝叶斯定理
- 高斯分布
  - 参数的最大似然估计
- 从贝叶斯定理到MLE / MAP 估计

# 概率 (Probability)

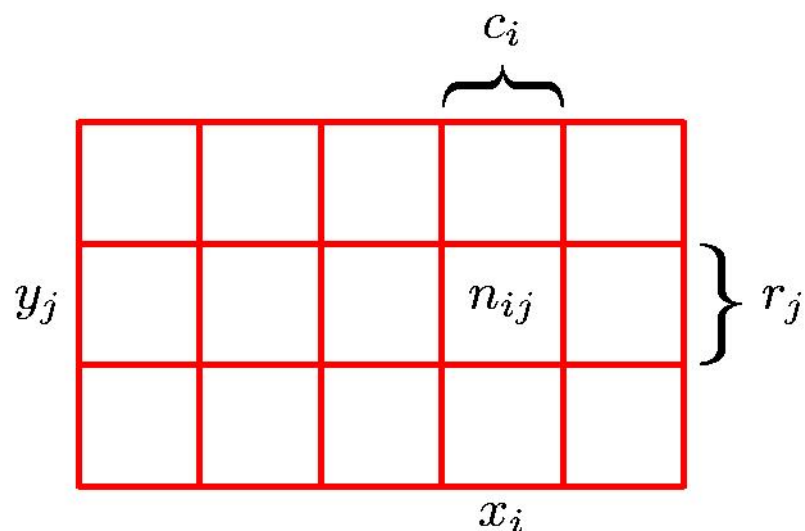
- 考虑一个试验.
  - 随机选择 **red box** 和 **blue box**
  - 随机抽取 **apple or orange**

- 相关的几个概念
  - 概率
  - 联合概率
  - 边缘概率
  - 条件概率
  - 先验概率
  - 后验概率
  - 先验→后验如何转换
  - 概率的运算规则



# 联合 / 边际概率 / 条件概率

- 



## Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

## Joint Probability

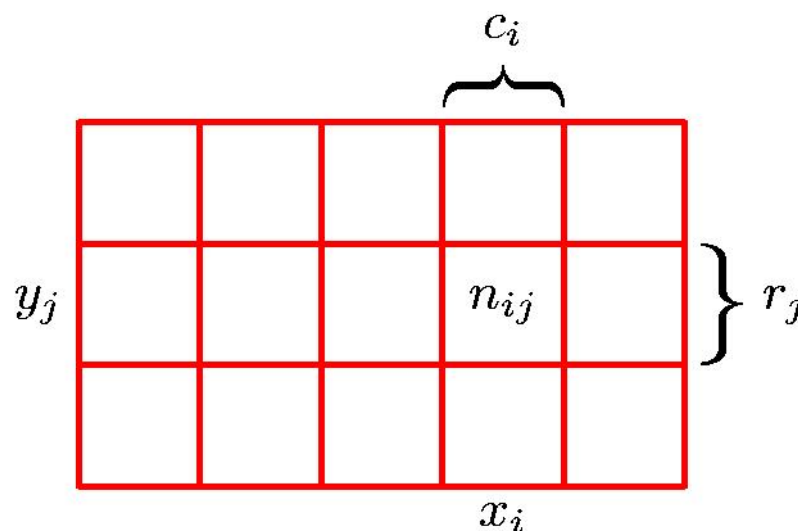
$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

## Conditional Probability

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

# 运算规则

•



## Sum Rule

$$\begin{aligned} p(X = x_i) &= \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^L n_{ij} \\ &= \sum_{j=1}^L p(X = x_i, Y = y_j) \end{aligned}$$

## Product Rule

$$\begin{aligned} p(X = x_i, Y = y_j) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p(Y = y_j | X = x_i) p(X = x_i) \end{aligned}$$

# The Rules of Probability

## Sum Rule

$$p(X) = \sum_Y p(X, Y)$$

## Product Rule

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

# 贝叶斯定理

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

– **where**  $p(X) = \sum_Y p(X|Y)p(Y)$

– **Prior probability)**

– **Posterior probability**

- MLE
- MAP
- Bayesian

posterior  $\propto$  likelihood  $\times$  prior



## 例：几个概率的计算

- 考虑一个试验.

- 以概率**P(B)**随机选择 **red box / blue box**

- $P(B=\text{red}) = 0.4$ ,  $P(B=\text{blue}) = 0.6$

- 随机抽取 **apple or orange**

- 假设选择的是**blue box**

- $P(F=\text{apple} \mid B=\text{blue}) = 3/4$

- $P(F=\text{orange} \mid B=\text{blue}) = 1/4$

- 假如选择的是**red box**

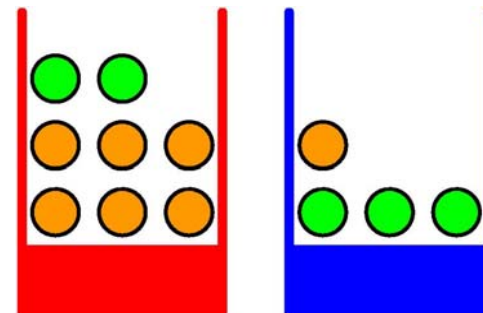
- $P(F=\text{apple} \mid B=\text{red}) = 1/4$

- $P(F=\text{orange} \mid B=\text{red}) = 3/4$

- 计算随机选择一个**Fruit**，结果是**apple or orange**的概率

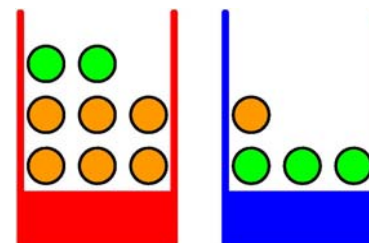
- $$P(F=a) = P(F=a \mid B=r) P(B=r) + P(F=a \mid B=b) P(B=b)$$
$$= 1/4 \times 2/5 + 3/4 \times 3/5 = 11/20$$

- $P(F=o) = ?$



## 例：先验概率 $\rightarrow$ 后验概率

- 考虑一个试验.
  - 以概率  $P(B)$  随机选择 **red box / blue box**
    - $P(B=\text{red})=0.4$ ,  $P(B=\text{blue})=0.6$
  - 随机抽取一个 **Fruit**
  - 请问：
    - (1) 可能是抽取自哪个 **box**
    - (2) 若发现选择的是 **orange**, 则可能抽取自哪个 **box**?



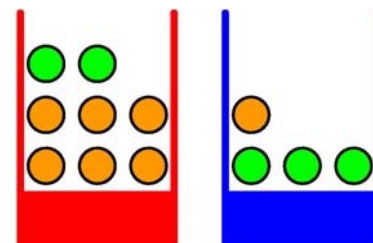
- (直观)答案
  - (1) **blue**
  - (2) **red**

- 计算过程

$$\begin{aligned} & P(B=\text{blue} \mid F=\text{orange}) \\ &= P(B=\text{blue}, F=\text{orange}) / P(F=\text{orange}) \\ &= P(F=\text{orange}, B=\text{blue}) / P(F=\text{orange}) \\ &= P(F=\text{orange} \mid B=\text{blue}) P(B=\text{blue}) / P(F=\text{orange}) \\ &= (1/4 \times 3/5) / (9/20) \\ &= 1/3 \\ \text{So, } & P(B=\text{red} \mid F=\text{orange}) = 2/3 \end{aligned}$$

## 例：先验概率 $\rightarrow$ 后验概率

- 考虑一个试验.
  - 以概率  $P(B)$  随机选择 **red box / blue box**
    - $P(B=\text{red})=0.4$ ,  $P(B=\text{blue})=0.6$
  - 随机抽取一个 **Fruit**
  - 请问：
    - (1) 可能是抽取自哪个 **box**
    - (2) 若发现选择的是 **apple**, 则可能抽取自哪个 **box**?
- (直观)答案
  - (1) **blue**
  - (2) **blue**



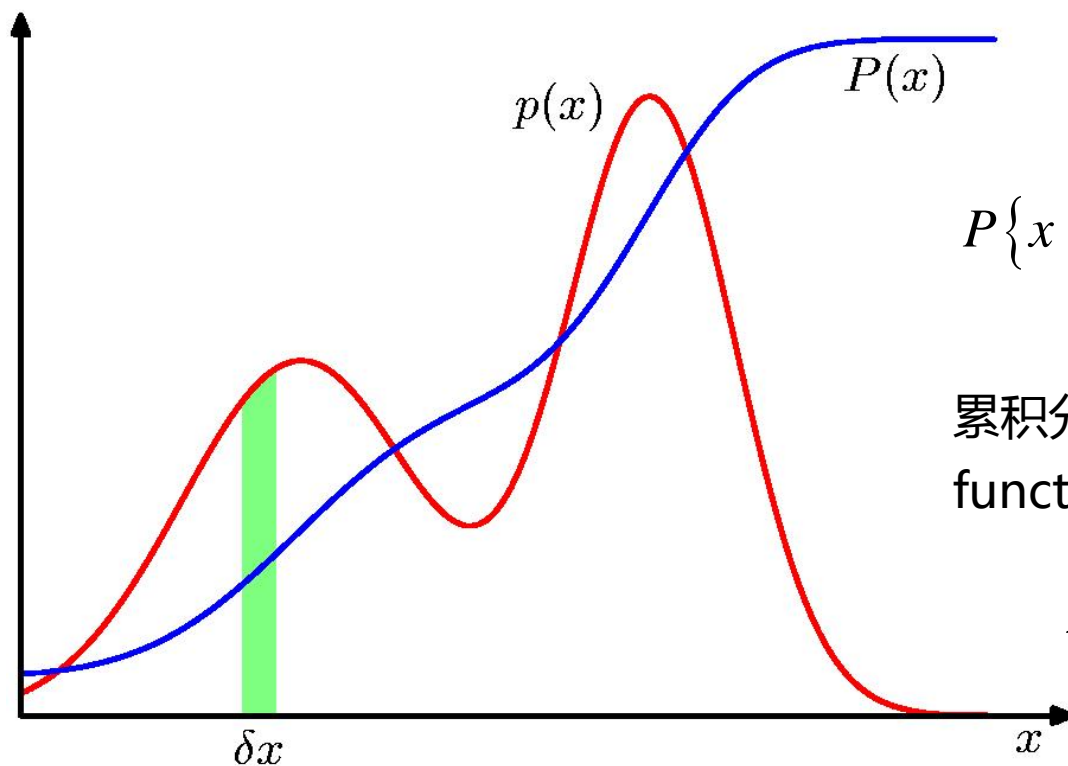
- 计算过程
$$\begin{aligned} & P(B=\text{blue} \mid F=\text{apple}) \\ &= P(B=\text{blue}, F=\text{apple}) / P(F=\text{apple}) \\ &= P(F=\text{apple}, B=\text{blue}) / P(F=\text{apple}) \\ &= P(F=\text{apple} \mid B=\text{blue}) P(B=\text{blue}) / P(F=\text{apple}) \\ &= (3/4 \times 3/5) / (11/20) \\ &= 9/11 \end{aligned}$$

# Q / A

- Any Questions...



# 概率密度 (probability Densities)



$$P\{x \in (a, b)\} = \int_a^b p(x) dx$$

累积分布函数(Cumulative distribution function)

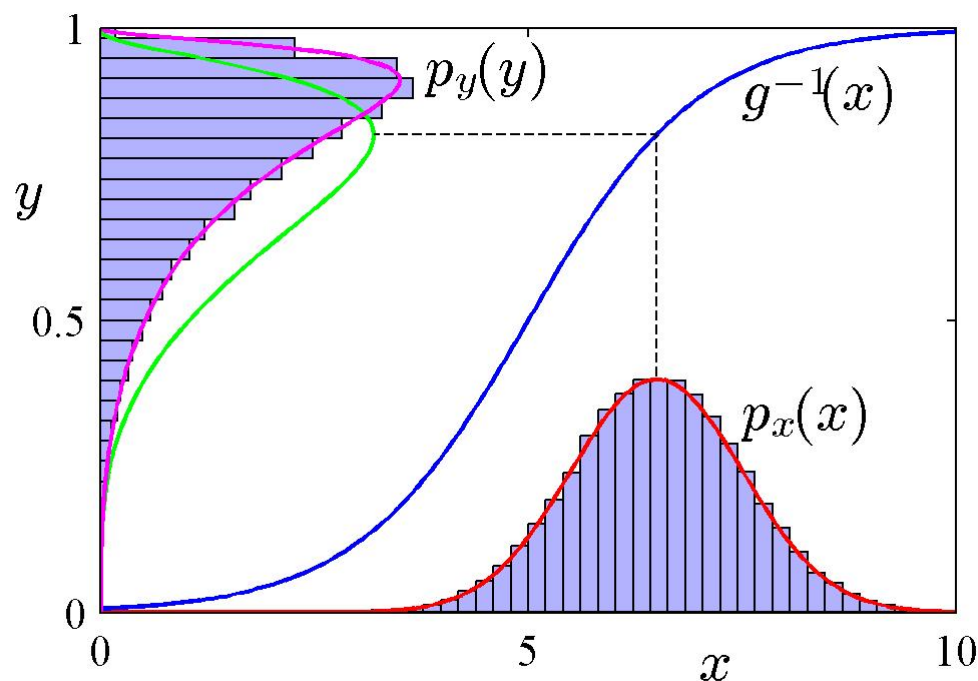
$$P(z) = \int_{-\infty}^z p(x) dx$$

$$p(x) \geq 0 \quad \int_{-\infty}^{\infty} p(x) dx = 1$$

# 概率密度的转换

- $p(x) dx \rightarrow q(y) dy$

$$\begin{aligned} p_y(y) &= p_x(x) \left| \frac{dx}{dy} \right| \\ &= p_x(g(y)) |g'(y)| \end{aligned}$$




高维情况，使用Jacobi行列式

# 函数的数学期望(expectations)

- 函数的数学期望

$$\mathbb{E}[f] = \sum_x p(x)f(x) \qquad \mathbb{E}[f] = \int p(x)f(x) dx$$

- 函数的条件期望

$$\mathbb{E}_x[f|y] = \sum_x p(x|y)f(x)$$


Conditional Expectation  
(discrete)

- 函数的数学期望的样本近似

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

Approximate Expectation  
(discrete and continuous)

# 函数的数学期望(expectations)

- 函数的数学期望 vs. 数学期望的样本近似

$$\mathbb{E}[f] = \int p(x) f(x) dx$$

- 给定**N**个数据点，采样自**p(x)**，则函数的数学期望可以近似计算

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

→ 采样方法(Sampling Method)  
e.g., 用于计算某些数值积分 (面积)



# 函数的数学期望(expectations)

- 函数的数学期望 vs. 数学期望的样本近似

$$\mathbb{E}[f] = \int p(x) f(x) dx \qquad \mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

- 假设  $f$  是损失函数，给定  $N$  个数据点看作训练数据，则数学期望对应于泛化误差，数学期望的样本估计对应于经验误差

$$R_N[f_N] = \frac{1}{N} \sum_{n=1}^N f_N(x_n) \xrightarrow{N \rightarrow \infty} R[f] = \int p(x) f(x) dx$$

- 统计学习理论揭示这个过程的收敛条件、收敛速度、如何控制收敛速度等

# Q / A

- Any Questions...



# 方差(Variance) 和协方差(Covariance)

- 方差

$$\text{var}[f] = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- 协方差

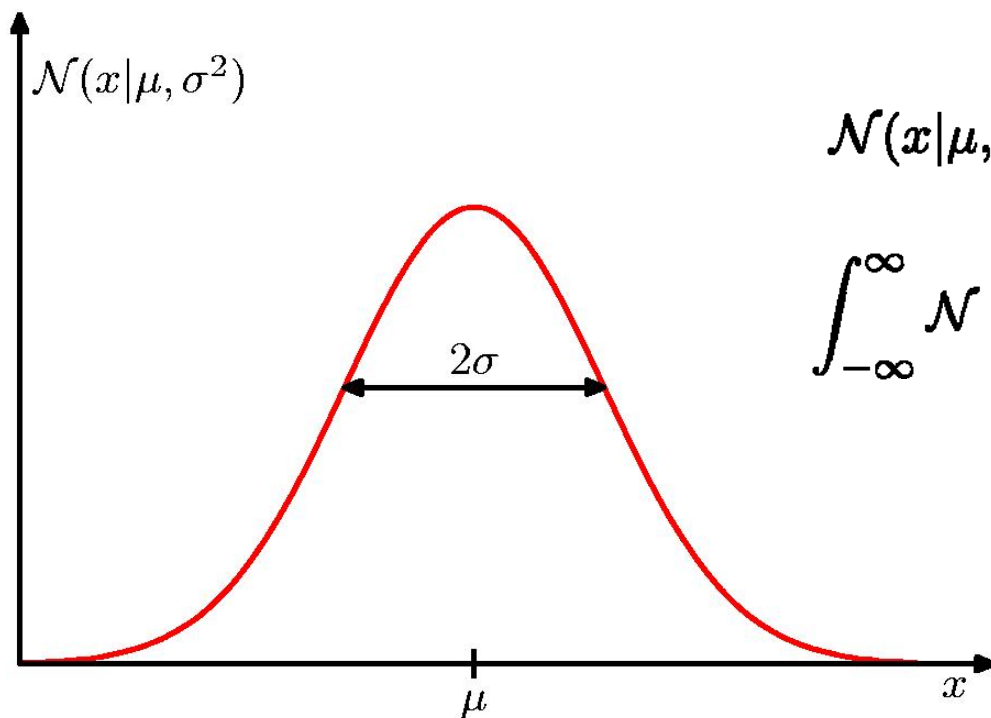
$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y} [\{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\}] \\ &= \mathbb{E}_{x,y} [xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

- 协方差矩阵

$$\begin{aligned} \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\} \{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}] \\ &= \mathbb{E}_{\mathbf{x}, \mathbf{y}} [\mathbf{x} \mathbf{y}^T] - \mathbb{E}[\mathbf{x}] \mathbb{E}[\mathbf{y}^T] \end{aligned}$$

# 高斯分布(Gaussian Distribution)

- $$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$



$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

# 高斯分布中的参数

- 均值

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, dx = \mu$$

- 二阶矩

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \, dx = \mu^2 + \sigma^2$$

- 方差

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

# 多元高斯分布(Multivariate Gaussian)

- 任意协方差矩阵

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$



— 若各个随机变量之间不相关，则协方差矩阵变成对角矩阵

# Q / A

- Any Questions...



- 概率论 vs. 统计学?
  - 分布/分布密度  $\rightarrow$  ...
  - 数据  $\rightarrow$  分布/分布密度

# 高斯分布中的参数估计

- 最大似然估计法(Maximal Likelihood Estimation)

- 给定**N**个数据点的数据集**X**，假设**i.i.d.**, **e.g.** 高斯分布


- independent and identically distributed: i.i.d.

- 数据集由特定参数下的高斯分布生成的概率为

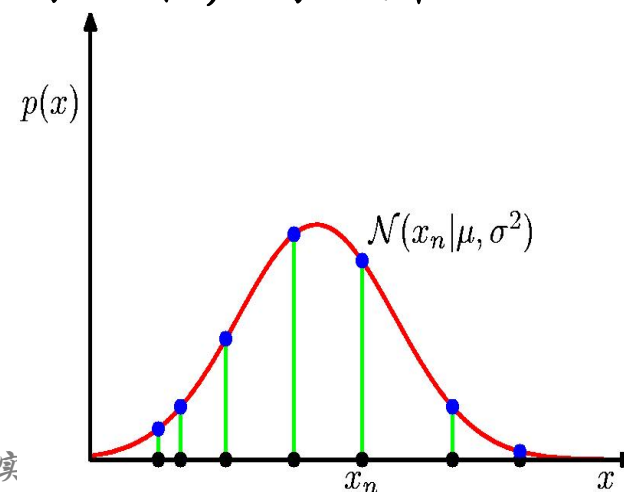
$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

- 称为似然函数(Likelihood function)

- 最大似然估计原则: 在给定数据的情况下，寻找最大化似然函数的参数


$$\max_{\mu, \sigma^2} L(\mu, \sigma^2 | X) = p(X | \mu, \sigma^2)$$

$$\max_{\mu, \sigma^2} l(\mu, \sigma^2 | X) = \log p(X | \mu, \sigma^2)$$





# 高斯分布中参数的最大似然估计

- 目标函数:

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

- 均值的估计

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



- 方差的估计

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

# 高斯分布中参数的最大似然估计的统计性质

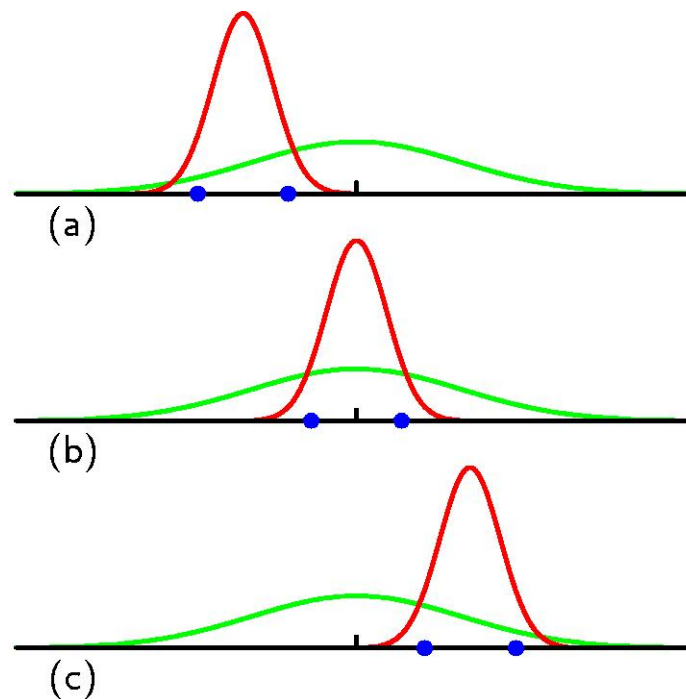
- 均值是无偏估计

$$\mathbb{E}[\mu_{\text{ML}}] = \mu$$

- 方差是有偏估计

$$\mathbb{E}[\sigma_{\text{ML}}^2] = \left(\frac{N-1}{N}\right) \sigma^2$$

$$\begin{aligned}\tilde{\sigma}^2 &= \frac{N}{N-1} \sigma_{\text{ML}}^2 \\ &= \frac{1}{N-1} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2\end{aligned}$$



# 从Bayes定理再看高斯分布中的参数估计

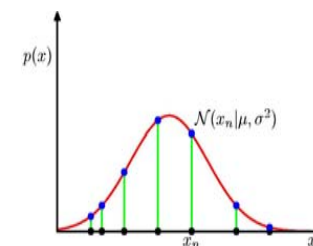
- 把数据集记为 $D$ ，分布中的待估计参数记为 $\mathbf{w}$ ，考虑到数据中的不确定性，则给定数据 $D$ ，估计参数 $\mathbf{w}$ 的问题表示为：

$$p(\mathbf{w} | D) = \frac{p(D | \mathbf{w}) p(\mathbf{w})}{p(D)}$$

- 参数 $\mathbf{w}$ 的估计问题即优化问题

$$\arg \max_{\mathbf{w}} p(\mathbf{w} | D)$$

- 如果关于 $\mathbf{w}$ 没有任何已知信息，则  $p(\mathbf{w}) = \mathbf{c}$

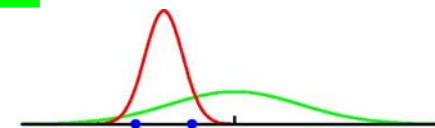


$$\begin{aligned} \arg \max_{\mathbf{w}} p(\mathbf{w} | D) &= \arg \max_{\mathbf{w}} p(D | \mathbf{w}) \\ &= \arg \max_{\mathbf{w}} L(\mathbf{w} | D) = \arg \max_{\mathbf{w}} \log L(\mathbf{w} | D) \end{aligned}$$

- 最大似然估计法(Maximal Likelihood Estimation)
- 如果关于 $\mathbf{w}$ 有一些已知信息，e.g.,  $\mathbf{w}$ 是高斯分布， $\mathbf{w}$ 是Laplacian分布

$$\arg \max_{\mathbf{w}} p(\mathbf{w} | D) = \arg \max_{\mathbf{w}} p(D | \mathbf{w}) p(\mathbf{w})$$

- 最大后验概率估计(Maximum A Posterior Estimation)



# Q / A

- Any Questions...



# 从Bayes 定理到不同的参数学习方法

- 贝叶斯定理

$$P(\mathbf{w} | D) = \frac{P(D | \mathbf{w})P(\mathbf{w})}{P(D)}$$

- 最大似然(ML)法

$$\max l(\mathbf{w} | D) = P(D | \mathbf{w})$$

- 最大后验概率(MAP)法

$$\max P(\mathbf{w} | D) \propto P(D | \mathbf{w})P(\mathbf{w})$$

- 贝叶斯方法

- 不再估计参数，而是估计参数的后验分布  $p(\mathbf{w} | D)$
    - 不是仅仅构造映射函数，而是依靠映射关系构造一个分布
    - 决策阶段利用参数的后验分布加权映射分布密度函数计算

$$p(t | x, D) = \int_{\mathbf{w}} p(t | x, \mathbf{w}) p(\mathbf{w} | D) d\mathbf{w}$$

# 部分数学基础内容回顾

- 概率论
- 决策论
- 信息论

# 决策论

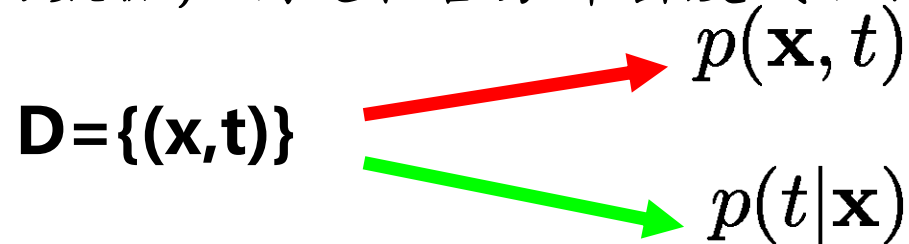
- 推理与决策
- 最小分类错误率
- 最小期望损失
- 回归问题
- 生成式 vs. 鉴别式

# 推理与决策

- 推理

- Inference

- The act or process of deriving logical conclusions from premises known or assumed to be true.
    - The act or process, through observation of patterns of facts, to indirectly see new meanings and contexts for understanding.
  - 基于给定的数据，确定联合分布密度或后验分布密度的过程



- 决策

- Decision

- 给定 $\mathbf{x}$ ，确定最优的输出 $t$

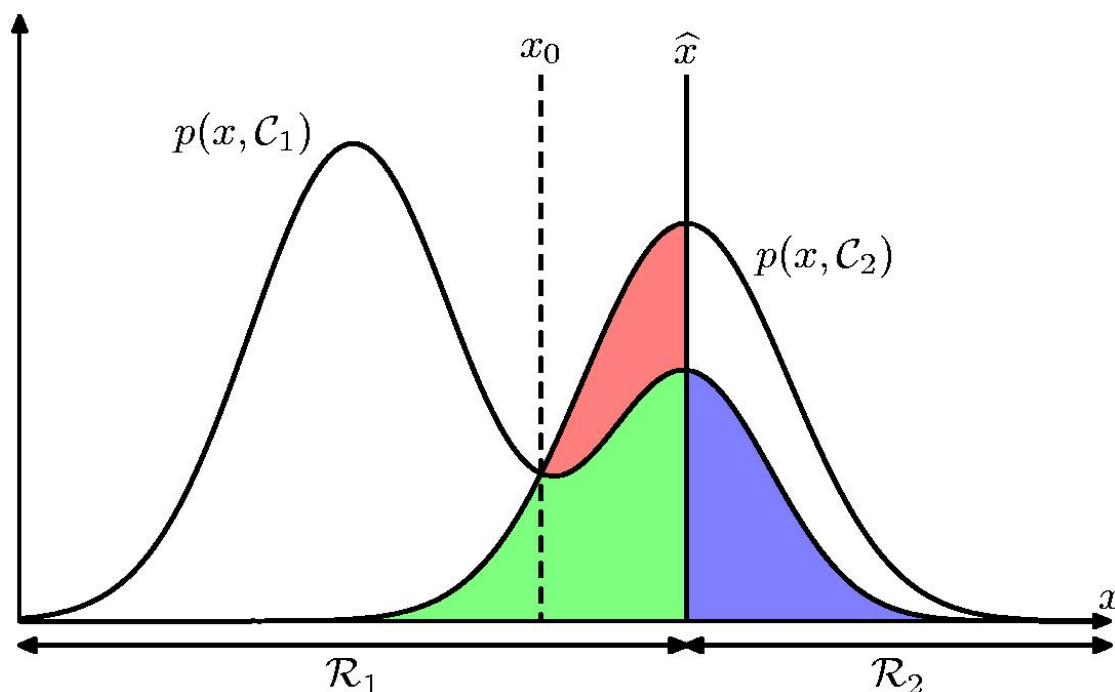




# 最小错误率

- Minimum Misclassification Rate

$$\begin{aligned} P(\text{mistake}) &= P(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + P(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1) d\mathbf{x}. \end{aligned}$$



# 最小期望损失

- **Minimum Expected Loss**

- Introduce a loss matrix to weight the errors
- Minimize

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, C_k) d\mathbf{x}$$

where Regions  $\mathcal{R}_j$  are chosen to minimize

- Note that

$$p(C_k, \mathbf{x}) = p(C_k | \mathbf{x}) p(\mathbf{x})$$

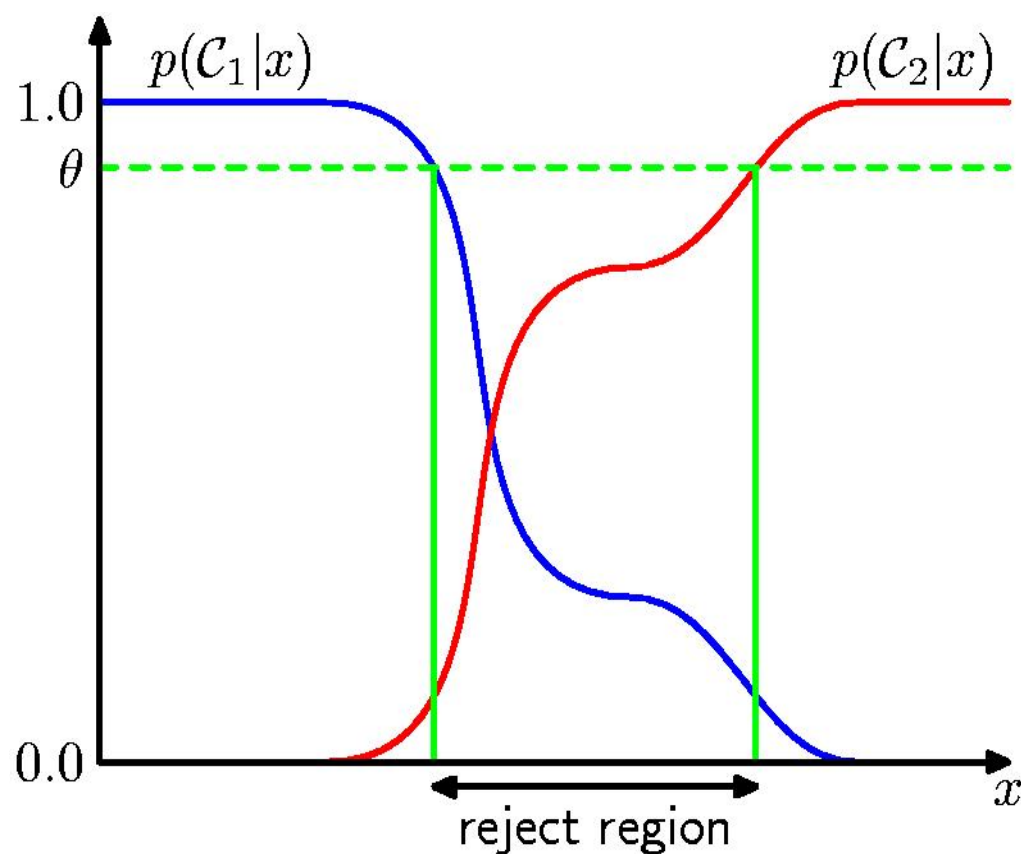
- So, 等价于 minimize  $\mathbb{E}[L] = \sum_k L_{kj} p(C_k | \mathbf{x})$

- **举例: Cancer诊断**

		Decision	
		cancer	normal
Truth	cancer	0	1000
	normal	1	0

# 带拒绝选项 (Reject option)

- 避免做出错误决策



# 分类问题解决方法的划分

- 生成式模型

- **Generative models**

- 推理问题: 建模联合分布  $p(C_k, \mathbf{x}) = p(\mathbf{x} | C_k) p(C_k)$

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) p(C_k)}{p(\mathbf{x})}$$

- 鉴别式模型

- **Discriminative models**

- 推理问题: 建模后验分布  $p(C_k | \mathbf{x})$

- 鉴别函数

- **Discriminant function**

- 推理问题: 无. 直接寻找鉴别函数  $f(\mathbf{x})$



# 为何需要推理后验分布？

- Minimizing risk
  - e.g., the loss matrix may change over time
- Reject option
- Compensating for Unbalanced class priors

$$p(C_k | \mathbf{x}) = \frac{p(\mathbf{x} | C_k) p(C_k)}{p(\mathbf{x})}$$

- Combining models

$$\begin{aligned} p(C_k | \mathbf{x}_1, \mathbf{x}_2) &= \frac{p(\mathbf{x}_1, \mathbf{x}_2 | C_k) p(C_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}_1 | C_k) p(\mathbf{x}_2 | C_k) p(C_k)}{p(\mathbf{x})} \\ &\propto \frac{p(C_k | \mathbf{x}_1) p(C_k | \mathbf{x}_2)}{p(C_k)} \end{aligned}$$

# 回归(Regression)

- Using Squared Loss Function

$$\mathbb{E}[L] = \iint \{y(\mathbf{x}) - t\}^2 p(\mathbf{x}, t) d\mathbf{x} dt$$

where

$$\begin{aligned} \{y(\mathbf{x}) - t\}^2 &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}] + \mathbb{E}[t|\mathbf{x}] - t\}^2 \\ &= \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 + 2\{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}\{\mathbb{E}[t|\mathbf{x}] - t\} + \{\mathbb{E}[t|\mathbf{x}] - t\}^2 \end{aligned}$$

$$\mathbb{E}[L] = \int \{y(\mathbf{x}) - \mathbb{E}[t|\mathbf{x}]\}^2 p(\mathbf{x}) d\mathbf{x} + \int \text{var}[t|\mathbf{x}] p(\mathbf{x}) d\mathbf{x}$$

- So, we have  $y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$

# 回归模型的划分

- Generative approach:

- **Model**  $p(t, \mathbf{x}) = p(\mathbf{x}|t)p(t)$

- **Use Bayes' theorem**

$$p(t|\mathbf{x}) = \frac{p(\mathbf{x}|t)p(t)}{p(\mathbf{x})}$$

- Discriminative approach:

- **Model**  $p(t|\mathbf{x})$  **directly**

**Then compute**  $y(\mathbf{x}) = \mathbb{E}[t|\mathbf{x}]$

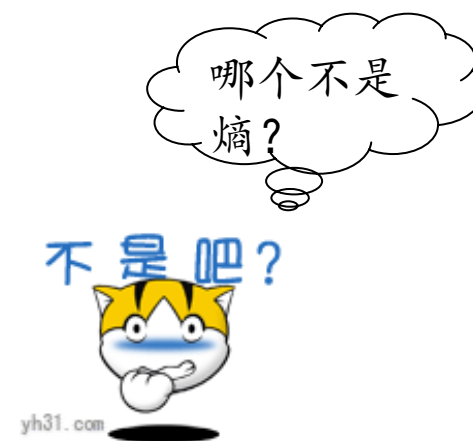
# 部分数学基础内容回顾

- 概率论
- 决策论
- 信息论



# 信息论

- 概率与熵
  - 最大熵
  - 微分熵
  - 条件熵
  - 联合熵
- KL散度
  - 互信息



# 熵 (Entropy)

$$H[x] = - \sum_x p(x) \log_2 p(x)$$

- Important quantity in
  - coding theory
  - statistical physics
  - machine learning

## 举例：熵的计算

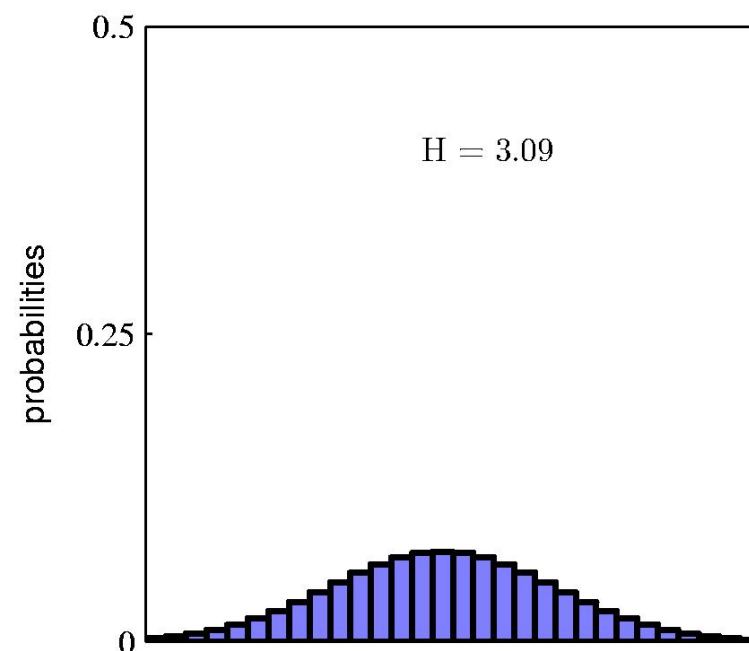
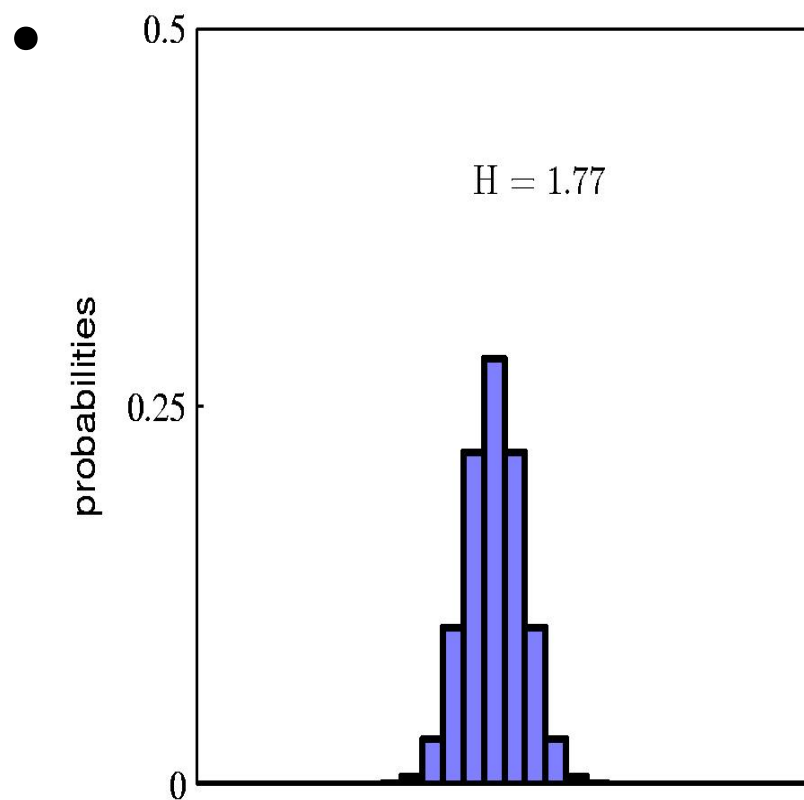
- | $x$    | a             | b             | c             | d              | e              | f              | g              | h              |
|--------|---------------|---------------|---------------|----------------|----------------|----------------|----------------|----------------|
| $p(x)$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ | $\frac{1}{64}$ | $\frac{1}{64}$ | $\frac{1}{64}$ | $\frac{1}{64}$ |
| code   | 0             | 10            | 110           | 1110           | 111100         | 111101         | 111110         | 111111         |

$$\begin{aligned} H[x] &= -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} \\ &= 2 \text{ bits} \end{aligned}$$

$$\begin{aligned} \text{average code length} &= \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 \\ &= 2 \text{ bits} \end{aligned}$$



# 举例：分布概率与熵的大小



# 最大熵原理 (maximum entropy principle)

- 最大熵原理
  - 在**1957**年由**E.T.Jaynes**提出的,
- 主要思想:
  - 在只掌握关于未知分布的部分知识时, 应该选取符合这些知识但熵值最大的概率分布
- 最大熵原理的实质:
  - 在已知部分知识的前提下, 关于未知分布最合理的推断是符合已知知识的最不确定或最随机的推断


# 最大熵问题举例 1:

- 例1: 考虑一个包含 $m$ 个状态的离散分布. 请估计这 $m$ 个状态的概率.

– 直观结果:

- $p(X = k) = \frac{1}{m}$  ?
- 为什么?
- 基于最大熵原理构造优化问题

$$\max E(p_k) \quad \text{s.t.} \quad \sum_{k=1}^m p_k = 1, \quad p_k \geq 0.$$


$$\max_{p_1, \dots, p_m} - \sum_{k=1}^m p_k \log p_k \quad \text{s.t.} \quad \sum_{k=1}^m p_k = 1, \quad p_k \geq 0$$

## 最大熵问题举例 2:

- 例2: 给定一阶矩和二阶矩, 请估计一个分布密度函数.
  - 直观结果:
    - Gaussian ?
    - 为什么?
- 基于最大熵原理构造优化问题

## 最大熵问题举例 2:

- 优化问题构造:

- 目标函数

$$\max_{p(x)} \text{Entrop}[p(x)]$$

- 约束条件

s.t.



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

$$p(x) \geq 0$$

$$\int_{-\infty}^{+\infty} xp(x) = \mu$$

$$\int_{-\infty}^{+\infty} (x - \mu)^2 p(x) = \sigma^2$$



# Q / A

- Any Questions...



# 微分熵 (Differential Entropy)

- 定义:

- Put bins of width  $\Delta$  along the real line

$$\lim_{\Delta \rightarrow 0} \left\{ - \sum_i p(x_i) \Delta \ln p(x_i) \right\} = - \int p(x) \ln p(x) dx$$

- 性质

- 在二阶矩有限的情况下，微分熵的最大值对应于高斯分布，即

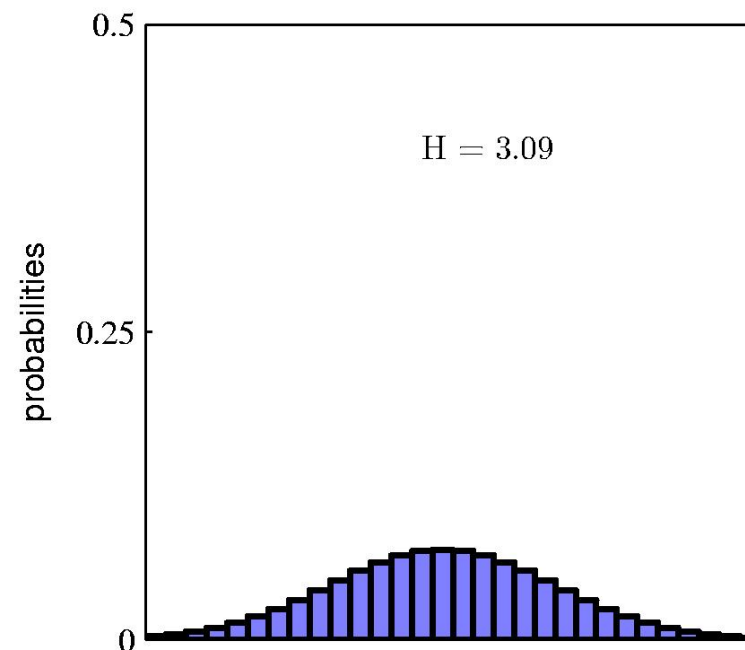
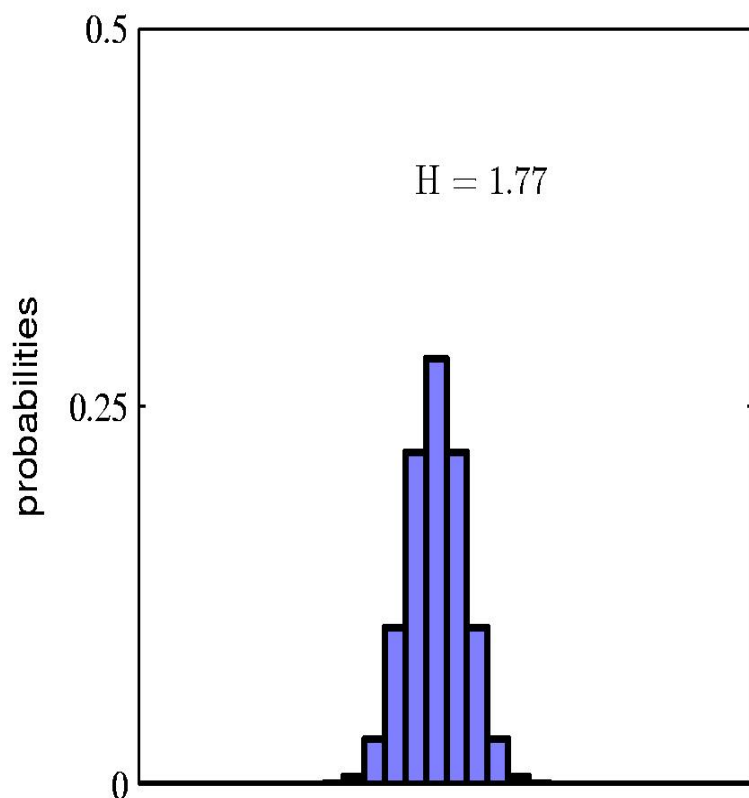
$$p(x) = \mathcal{N}(x|\mu, \sigma^2)$$

- 其中

$$H[x] = \frac{1}{2} \{1 + \ln(2\pi\sigma^2)\}.$$



# 举例：离散分布的概率与熵的大小



- 离散情况:  $\rightarrow$  均匀分布
- 连续情况:  $\rightarrow$  高斯分布



## 最大熵问题举例 2 (续):

- 优化问题构造:

- 目标函数

$$\max_{p(x)} \text{Entrop}[p(x)] = -\int_{-\infty}^{+\infty} p(x) \ln p(x) dx$$

- 约束条件

s.t.



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

$$\int_{-\infty}^{+\infty} xp(x) dx = \mu$$

$$\int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx = \sigma^2$$

$$p(x) \geq 0$$

## 最大熵问题举例 2 (续):

- 优化问题构造:
  - Lagrange 乘子法, 引入3个乘子

$$\min_{p(x)} \int_{-\infty}^{+\infty} p(x) \ln p(x) dx + \lambda_1 \left( \int_{-\infty}^{+\infty} p(x) dx - 1 \right) \\ \rightarrow + \lambda_2 \left( \int_{-\infty}^{+\infty} xp(x) dx - \mu \right) \\ + \lambda_3 \left( \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right)$$

$$\text{s.t. } p(x) \geq 0$$

# Fréchet微分与变分问题的极值条件

- Fréchet微分

- 泛函的**Fréchet**微分可以解释为最佳局部线性逼近

- 定义为：
$$d\varepsilon(f, h) = \left[ \frac{d}{d\beta} \varepsilon(f + \beta \cdot h) \right]_{\beta=0} \quad \text{其中, } h = h(\mathbf{x}) \in H$$

- 函数 $F(x)$ 为泛函的一个相对极值的必要条件

$$d\varepsilon(f, h) = 0$$

- 对所有的线性函数 $h(x)$ ，泛函的**Fréchet**微分在 $\mathbf{f}(\mathbf{x})$ 处均为**0**



## 最大熵问题举例 2 (续):

- 优化问题构造:
  - Lagrange 乘子法, 引入3个乘子

$$\min_{p(x)} \int_{-\infty}^{+\infty} p(x) \ln p(x) dx + \lambda_1 \left( \int_{-\infty}^{+\infty} p(x) dx - 1 \right) \\ \rightarrow + \lambda_2 \left( \int_{-\infty}^{+\infty} xp(x) dx - \mu \right) \\ + \lambda_3 \left( \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right)$$

$$\text{s.t. } p(x) \geq 0$$

$$\rightarrow p(x) = \exp \left\{ -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 \right\}$$

$$\lambda_2 = 0, \lambda_3 = -\frac{1}{2\sigma^2}, \lambda_1 = 1 - \ln \sqrt{2\pi\sigma^2}$$

# 条件熵 (Conditional Entropy)

- 定义:

$$H[\mathbf{y}|\mathbf{x}] = - \iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) d\mathbf{y} d\mathbf{x}$$

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}]$$

— 联合熵

- $H[x,y]$



# K-L散度(Kullback-Leibler Divergence)

- 定义:

$$\begin{aligned}\text{KL}(p\|q) &= - \int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left( - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x}\end{aligned}$$

- 意义: 度量密度之间的“距离”

- 性质:  $\text{KL}(p\|q) \geq 0$        $\text{KL}(p\|q) \neq \text{KL}(q\|p)$

- 样本近似计算

$$\text{KL}(p\|q) \simeq \frac{1}{N} \sum_{n=1}^N \{ -\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n) \}$$

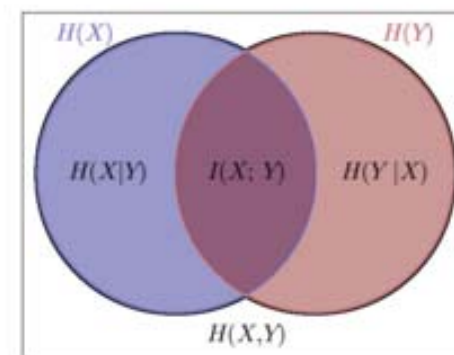
# 互信息 (Mutual Information)

- 定义:

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv \text{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= - \iint p(\mathbf{x}, \mathbf{y}) \ln \left( \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \end{aligned}$$

- 意义:

- 两个随机变量之间的相关性



- 与熵和条件熵之间的关系

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &= H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}] \\ &= H(\mathbf{x}) + H(\mathbf{y}) - H(\mathbf{x}, \mathbf{y}) \end{aligned}$$

# Q / A

- Any Questions...



# 小结：部分数学基础

- 概率与统计

- 先验概率、后验概率、贝叶斯定理
- 函数的数学期望、高斯分布
- 最大似然估计
  - 从贝叶斯定理到MLE / MAP 估计

- 决策论

- 推理与决策
- 生成式 vs. 鉴别式
- 回归模型

- 信息论

- 熵
- KL散度
- 最大熵



不是吧?



“Again, you can't connect the dots looking forward; you can only connect them looking backwards. So you have to trust that the dots will somehow connect in your future. You have to trust in something - your gut, destiny, life, karma, whatever. This approach has never let me down, and it has made all the difference in my life.” ---- Stephen Jobs, 2005.

**“你在向前展望的时候不可能将这些片断串连起来；你只能在回顾的时候将点点滴滴串连起来。所以你必须相信这些片断会在你未来的某一天串连起来。你必须要相信某些东西：你的勇气、目的、生命、因缘.....这个过程从来没有令我失望,只是让我的生命更加地与众不同。” -**  
--- Stephen Jobs, 2005.