



# 深度学习：从理论到实践

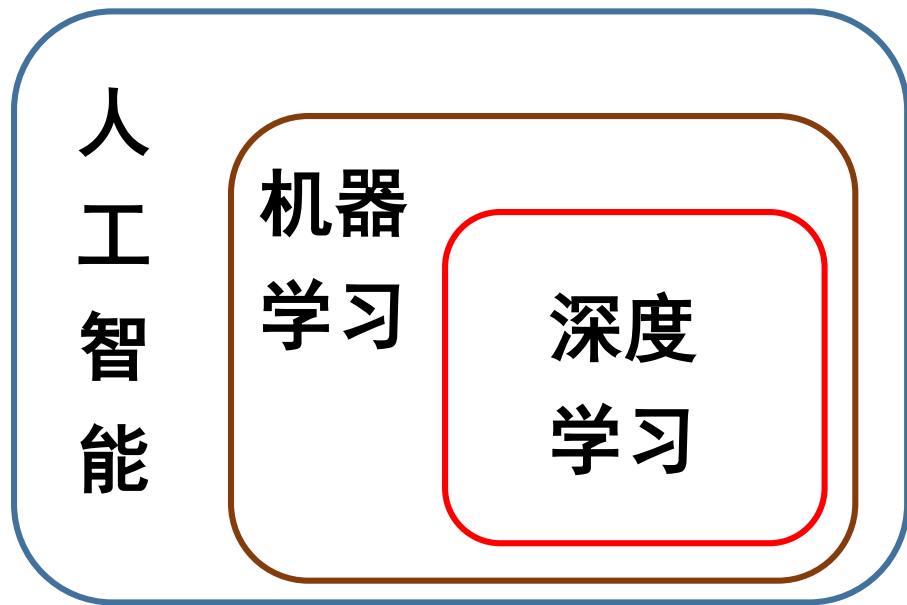
## 第一章：深度学习基础（上）



- ✓ 机器学习与深度学习
- ✓ 贝叶斯决策论
- ✓ 密度估计
  - ◆ 参数估计：极大似然估计与EM算法
  - ◆ 非参数估计：KNN与Parzen窗估计

- ✓ 机器学习与深度学习
- ✓ 贝叶斯决策论
- ✓ 密度估计
  - ◆ 参数估计：极大似然估计与EM算法
  - ◆ 非参数估计：KNN与Parzen窗估计

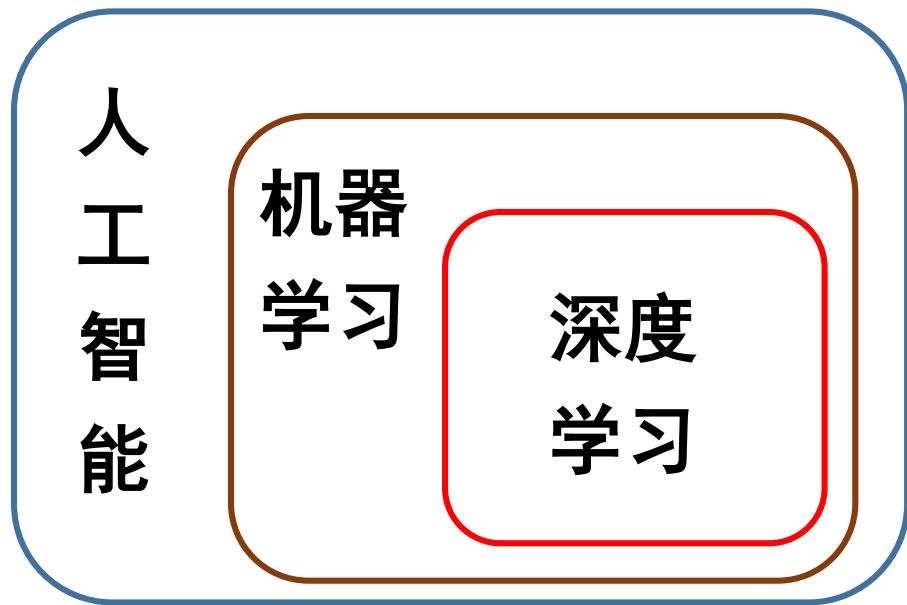
# 机器学习与深度学习



## 分类器构造

1. 回归
2. 支持向量机
3. 神经网络
4. 决策树、随机森林
5. 独立于算法的机器学习

# 机器学习与深度学习

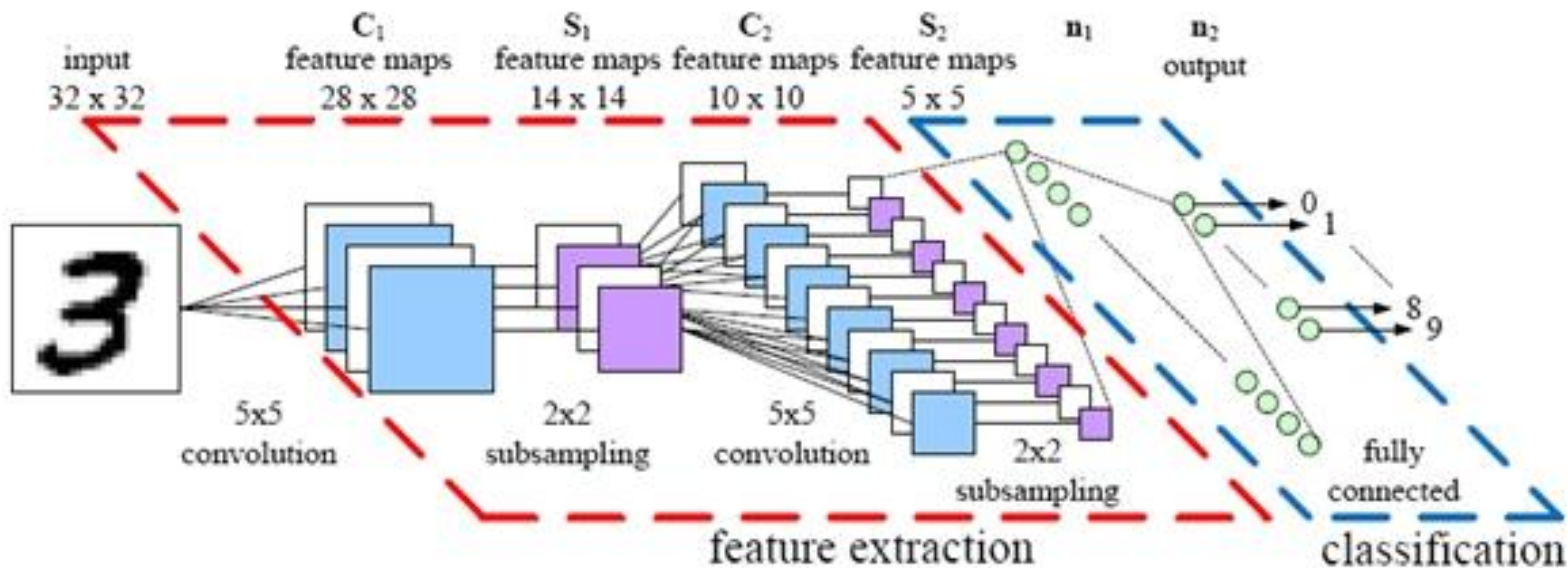


## 机器学习任务

1. 监督学习
2. 无监督学习
3. 半监督学习
4. 强化学习

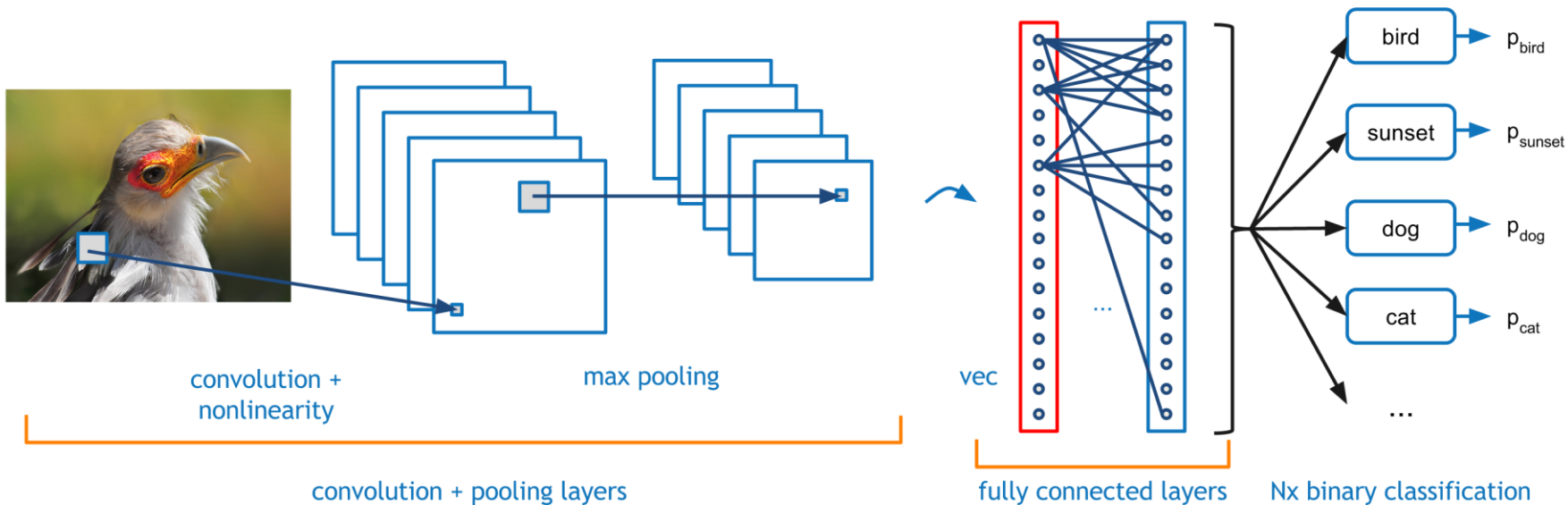
# 机器学习与深度学习

## 文字分类



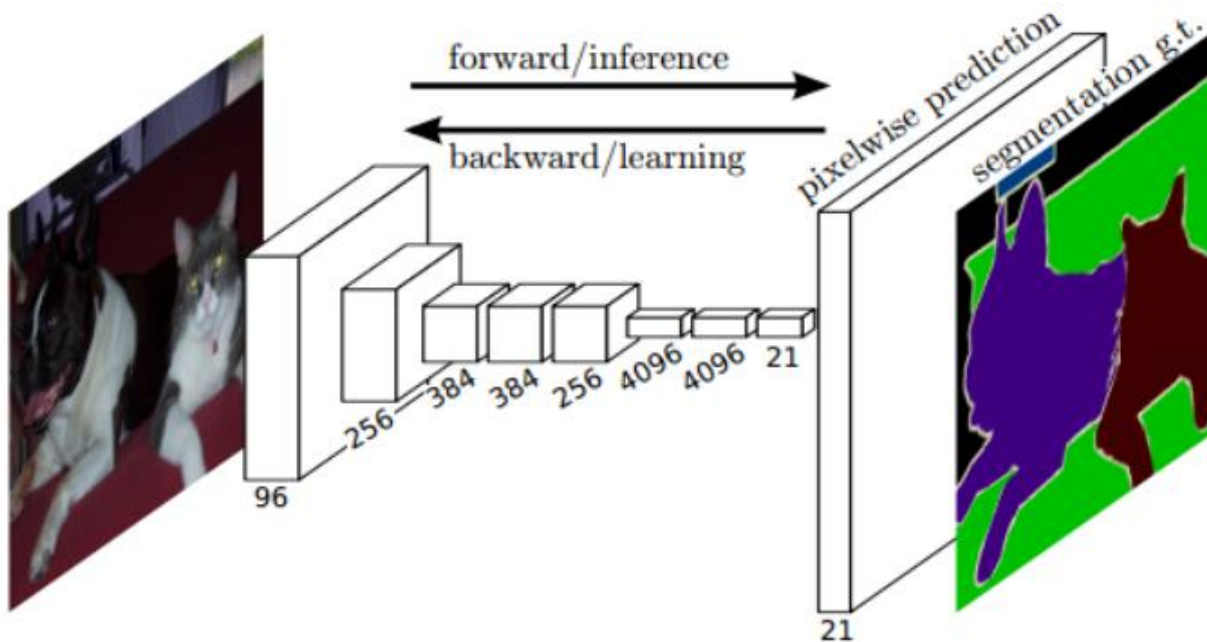
# 机器学习与深度学习

## 动物分类



# 机器学习与深度学习

## 图像分割

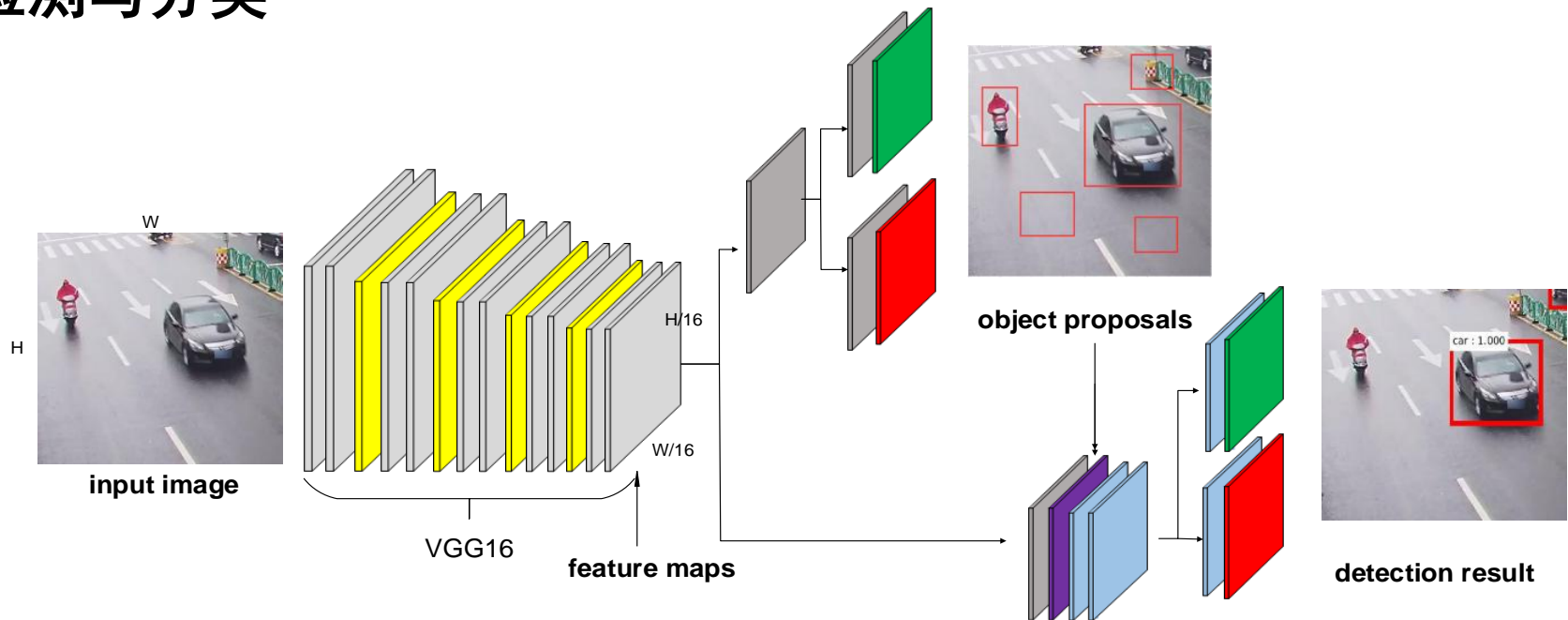


卷积神经网络  
目标分割



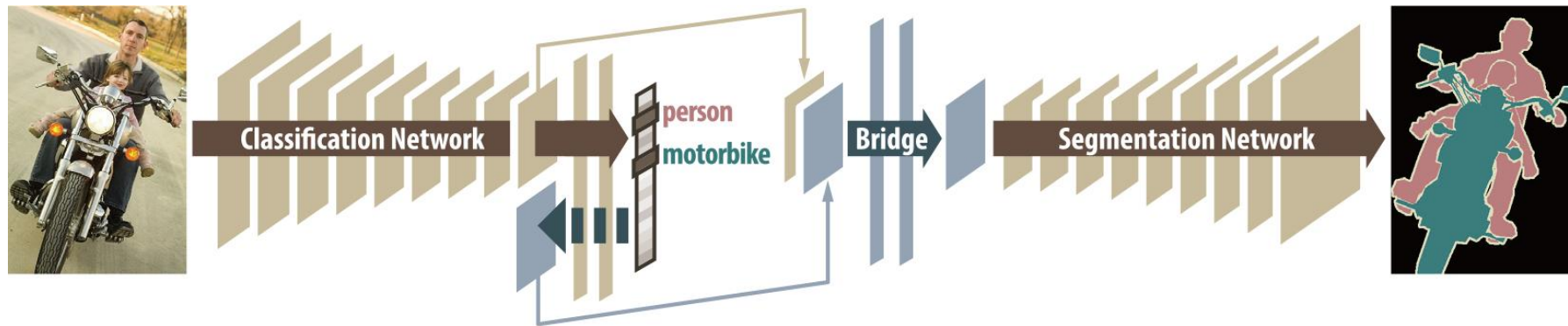
# 机器学习与深度学习

## 检测与分类



## 检测与分割

### 卷积神经网络目标检测与分割



## 课程目标

1. 能从机器学习、模式识别等角度看深度学习
2. 深入理解深度学习理论知识
3. 熟悉常见深度学习模型以及实现方法
4. 熟悉Caffe深度学习开发框架

- ✓ 机器学习与深度学习
- ✓ 贝叶斯决策论
- ✓ 密度估计
  - ◆ 参数估计：极大似然估计与EM算法
  - ◆ 非参数估计：KNN与Parzen窗估计

# 贝叶斯决策—引言

问 题：警察判断驾驶员是否酒驾。

判别方法：观察驾驶员的脸部颜色（发红的程度）。

**驾驶员的脸色红到什么程度，可以认为是酒驾呢？**

## 问题形式化：

$\omega_1$  表示酒驾事件， $\omega_2$  表示未酒驾事件

$x$  表示驾驶员脸红的程度

为了判断是否酒驾，可计算两个后验概率  $P(\omega_1 | x)$  和  $P(\omega_2 | x)$

判断方法：如果  $P(\omega_1 | x) > P(\omega_2 | x)$ ，则判断该驾驶员酒驾了。

# 贝叶斯决策—引言

## 贝叶斯决策的核心概念：

类别：  $\omega_i, i = 1, \dots, c$

特征矢量：  $\mathbf{x} = [x_1, \dots, x_d] \in R^d$

先验概率：  $P(\omega_i) \quad \sum_{i=1}^c P(\omega_i) = 1$

概率密度函数(条件概率、似然)：  $p(\mathbf{x} | \omega_i)$

后验概率： 
$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j)P(\omega_j)}$$

**问题1：后验概率是否满足归一化条件。**

# 贝叶斯决策—最小错误率

分类错误率(两类问题)：

$$P(e) = \int P(e | x) p(x) dx$$

$$\text{其中, } P(e | x) = \begin{cases} P(\omega_2 | x), & \text{如果决策 } x \in \omega_1 \\ P(\omega_1 | x), & \text{如果决策 } x \in \omega_2 \end{cases}$$

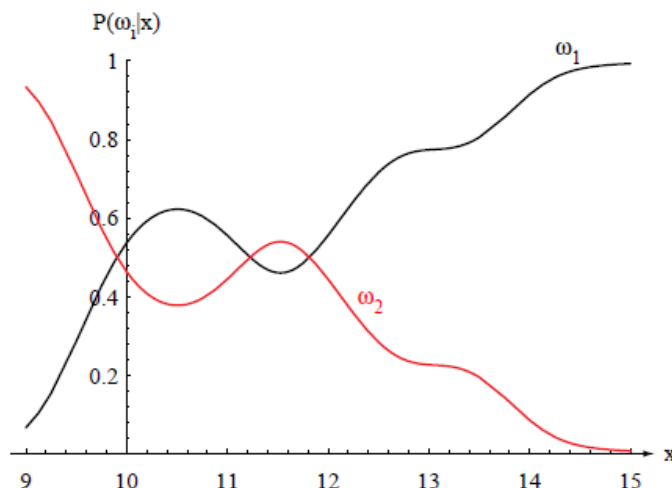
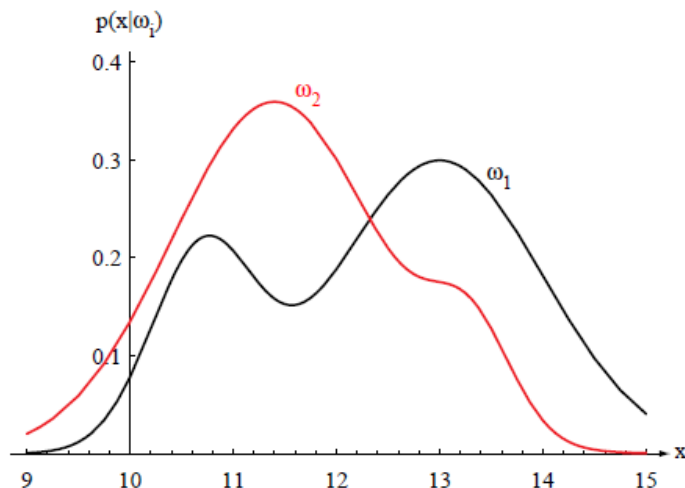
最小错误率决策： $\min P(e)$

判断方法：如果  $P(\omega_1 | x) > P(\omega_2 | x)$ ，则  $x \in \omega_1$ ，否则  $x \in \omega_2$ 。



# 贝叶斯决策—最小错误率

举例说明：



$$P(\omega_1) = 2/3$$

$$P(\omega_2) = 1/3$$

$$P(\omega_i | \mathbf{x}) = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{p(\mathbf{x})} = \frac{p(\mathbf{x} | \omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x} | \omega_j)P(\omega_j)}$$

# 贝叶斯决策—简要总结

## 贝叶斯问题：

- ✓ 是否可以直接使用先验进行判断？（可以，但是准度不高。）
- ✓ 先验和后验的关系？（通过似然或者观察，将先验转换为后验。）
- ✓ 如果似然分布一样，会怎么样？（后验和先验一致，相当于用先验）
- ✓ 引入新的观察（特征），后验是否可转换为先验？（理论上可以，通常将所有特征一起考虑，即一起用于计算似然。）

# 贝叶斯决策—简要总结

## 贝叶斯问题：

- ✓ 除了最小错误率贝叶斯决策，还有其它决策吗？（有，例如：最小风险贝叶斯决策，将风险因子考虑在内。）
- ✓ 后验计算的核心是先验和似然（类条件概率密度），如何估计？（先验通常利用已知的知识得到，似然一般通过数据估计得到。）

- ✓ 机器学习与深度学习
- ✓ 贝叶斯决策论
- ✓ 密度估计
  - ◆ 参数估计：极大似然估计与EM算法
  - ◆ 非参数估计：KNN与Parzen窗估计

# 参数估计——最大似然估计

## 问题描述

给定样本  $D = \{x_1, x_2, \dots, x_n\}$ ，估计分布的参数  $\theta$

## 要求

- ◆ 分布：分布的类型确定，参数确定但未知
- ◆ 样本：所有样本要求独立同分布

# 参数估计——最大似然估计

**最大似然估计**期望所有样本的似然最大，即

$$\max p(D|\theta) = p(x_1, x_2, \dots, x_n|\theta) = \prod_{k=1}^n p(x_k|\theta)$$

**样本独立**

对数似然函数  $l(\theta) \equiv \ln p(D|\theta) = \sum_{k=1}^n \ln p(x_k|\theta)$

求解  $\theta = \arg \max_{\theta} l(\theta)$       **最优的必要条件，极值点导数为0**

$$0 = \nabla_{\theta} l(\theta) = \sum_{k=1}^n \nabla_{\theta} \ln p(x_k|\theta)$$

# 参数估计——最大似然估计

例子1：假设所有样本服从方差已知而均值未知的高斯分布，确定均值 $\mu$

第一步：计算每个样本的对数似然函数  $\ln p(\mathbf{x}_k|\mu) = -\frac{1}{2}\ln[(2\pi)^d|\Sigma|] - \frac{1}{2}(\mathbf{x}_k - \mu)^t \Sigma^{-1}(\mathbf{x}_k - \mu)$

第二步：计算每个样本的参数 $\theta$  (均值 $\mu$ ) 的导数  $\nabla_{\theta} \ln p(\mathbf{x}_k|\mu) = \Sigma^{-1}(\mathbf{x}_k - \mu)$

第三步：所有样本导数求和为0  $\sum_{k=1}^n \Sigma^{-1}(\mathbf{x}_k - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k$

多维高斯分布 
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2}(\mathbf{x} - \mu)^t \Sigma^{-1}(\mathbf{x} - \mu) \right]$$

# 参数估计——最大似然估计

例子2：假设所有样本服从均值方差都未知的一维高斯分布，确定均值和方差

第○步：确定参数形式  $\theta_1 = \mu$  and  $\theta_2 = \sigma^2$

第一步：计算每个样本的对数似然函数  $\ln p(x_k|\theta) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2}(x_k - \theta_1)^2$

第二步：计算每个样本的参数 $\theta$ 的导数  $\nabla_{\theta} l = \nabla_{\theta} \ln p(x_k|\theta) = \begin{bmatrix} \frac{1}{\theta_2}(x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$

第三步：所有样本导数求和为0

$$\sum_{k=1}^n \frac{1}{\hat{\theta}_2}(x_k - \hat{\theta}_1) = 0 \Rightarrow \hat{\mu} = \frac{1}{n} \sum_{k=1}^n x_k \Rightarrow -\sum_{k=1}^n \frac{1}{\hat{\theta}_2} + \sum_{k=1}^n \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \Rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \hat{\mu})^2$$

一维高斯分布

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



# 参数估计—EM算法

## 问题描述：

**数据缺失**情况下的参数估计，即：给定输入数据

$$\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} = \mathcal{D}_g \cup \mathcal{D}_b$$

其中，部分样本  $\mathbf{x}_k = \{\mathbf{x}_{kg}, \mathbf{x}_{kb}\}$ 。估计参数  $\theta$ 。

## 要求

和最大似然估计的要求一致。

# 参数估计—EM算法

期望最大化算法(EM算法)的核心思路：迭代优化。即：给定上次迭代参数 $\theta^i$ ，估计新的参数 $\theta^{i+1}$ 。因此，EM算法的整体框架如下：

1. 初始化 $\theta^0$ 以及 $i = 0$
2. *do*  $i \leftarrow i + 1$
3.     利用 $\theta^i$ 更新 $\theta^{i+1}$ ;
4. *until* 终止条件
5. 返回 $\theta = \theta^{i+1}$

EM算法的两个核心问题

1. 如何利用 $\theta^i$ 更新 $\theta^{i+1}$
2. 如何确认定终止条件

# 参数估计—EM算法

如何利用参数 $\theta^i$ 估计新参数 $\theta$

1. 利用参数 $\theta^i$ ，对缺失函数求期望（期望步）

$$Q(\theta; \theta^i) = \mathcal{E}_{\mathcal{D}_b}[\ln p(\mathcal{D}_g, \mathcal{D}_b; \theta) | \mathcal{D}_g; \theta^i]$$

2. 最大化期望（最大化步）

$$\theta^{i+1} \leftarrow \arg \max_{\theta} Q(\theta; \theta^i)$$

终止条件（参数变化很小）  $Q(\theta^{i+1}; \theta^i) - Q(\theta^i; \theta^{i-1}) \leq T$

# 参数估计—EM算法

例子1：假设数据由二维空间的4个点组成，如下：

$$\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} = \left\{ \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} * \\ 4 \end{pmatrix} \right\}$$

其中\*表示样本4的第一个特征值未知（或缺失）。目的，估计二维高斯分布（协方差矩阵为对角阵），

即估计  $\theta = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix}$ ，已知的初始化为  $\theta^0 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$

# 参数估计—EM算法

期望步（对缺失数据计算期望）：

$$\begin{aligned} Q(\theta; \theta^0) &= \mathcal{E}_{x_{41}} [\ln p(\mathbf{x}_g, \mathbf{x}_b; \theta | \theta^0; \mathcal{D}_g)] \\ &= \int_{-\infty}^{\infty} \left[ \sum_{k=1}^3 \ln p(\mathbf{x}_k | \theta) + \ln p(\mathbf{x}_4 | \theta) \right] p(x_{41} | \theta^0; x_{42} = 4) dx_{41} \\ &= \sum_{k=1}^3 [\ln p(\mathbf{x}_k | \theta)] + \underbrace{\int_{-\infty}^{\infty} \ln p \left( \binom{x_{41}}{4} \middle| \theta \right) \frac{p \left( \binom{x_{41}}{4} | \theta^0 \right)}{\left( \int_{-\infty}^{\infty} p \left( \binom{x'_{41}}{4} \middle| \theta^0 \right) dx'_{41} \right)} dx_{41}}_{\equiv K} \end{aligned}$$

# 参数估计—EM算法

期望步（对缺失数据计算期望）：

$$\begin{aligned} Q(\theta; \theta^0) &= \sum_{k=1}^3 [\ln p(\mathbf{x}_k | \theta)] + \frac{1}{K} \int_{-\infty}^{\infty} \ln p \left( \begin{pmatrix} x_{41} \\ 4 \end{pmatrix} \middle| \theta \right) \frac{1}{2\pi \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix}} \exp \left[ -\frac{1}{2} (x_{41}^2 + 4^2) \right] dx_{41} \\ &= \sum_{k=1}^3 [\ln p(\mathbf{x}_k | \theta)] - \frac{1 + \mu_1^2}{2\sigma_1^2} - \frac{(4 - \mu_2)^2}{2\sigma_2^2} - \ln (2\pi\sigma_1\sigma_2). \end{aligned}$$

# 参数估计—EM算法

最大化步（期望最大化）：

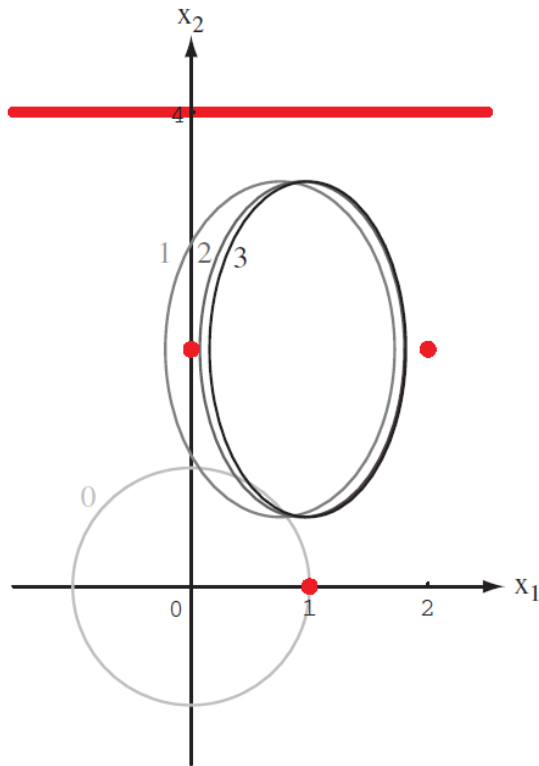
$$\max \sum_{k=1}^3 [\ln p(\mathbf{x}_k | \boldsymbol{\theta})] - \frac{1 + \mu_1^2}{2\sigma_1^2} - \frac{(4 - \mu_2)^2}{2\sigma_2^2} - \ln (2\pi\sigma_1\sigma_2)$$

求解结果：

$$\boldsymbol{\theta}^1 = \begin{pmatrix} 0.75 \\ 2.0 \\ 0.938 \\ 2.0 \end{pmatrix}$$

# 参数估计—EM算法

迭代过程:



3次迭代后的结果

$$\mu = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix}, \text{ and } \Sigma = \begin{pmatrix} 0.667 & 0 \\ 0 & 2.0 \end{pmatrix}$$



# 参数估计—EM算法

例子2：多高斯参数估计。多高斯分布定义如下：

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k p(\mathbf{x} | \theta_k)$$
$$\text{subject to } \sum_{k=1}^K \pi_k = 1$$

其中， $p(\mathbf{x} | \theta_k) = \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$  为单高斯分布。

# 参数估计—EM算法

最大似然估计（1. 计算对数似然，2. 对参数求导为0）

$$\max LL = \log \prod_{n=1}^N p(\mathbf{x}_n) = \sum_{n=1}^N \log \sum_{k=1}^K \pi_k p(\mathbf{x}_n | \theta_k)$$

$$\nabla_{\pi_k} LL = 0, \quad \nabla_{\mu_k} LL = 0, \quad \nabla_{\Sigma_k} LL = 0$$

无法解析求解！

# 参数估计—EM算法

EM估计：将数据看作是不完整数据，每个数据包含隐含的类别指示变量

$$z_{nk} \in \{0, 1\}, \quad k = 1, \dots, K$$

期望步，定义期望： $Q(\Theta, \Theta^{old}) = \sum_{\mathbf{Z}} \log p(\mathbf{X}, \mathbf{Z} | \Theta) p(\mathbf{Z} | \mathbf{X}, \Theta^{old})$

期望计算如下： $Q(\Theta, \Theta^{old}) = E_{\mathbf{Z}}[\log p(\mathbf{X}, \mathbf{Z} | \Theta)] = \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \{ \log \pi_k + \log \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \}$

$$p(\mathbf{X}, \mathbf{Z} | \Theta) = \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)^{z_{nk}} \quad \gamma(z_{nk}) = P(z_{nk} = 1 | \mathbf{x}_n) = \frac{\pi_k N(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j N(\mathbf{x}_n | \mu_j, \Sigma_j)}$$

# 参数估计—EM算法

EM估计：将数据看作是不完整数据，每个数据包含隐含的类别指示变量

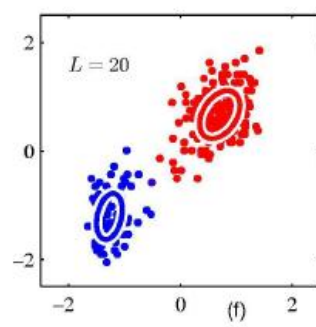
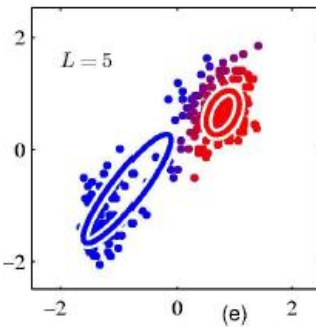
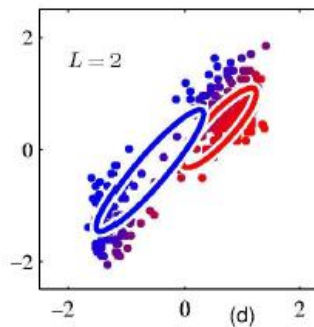
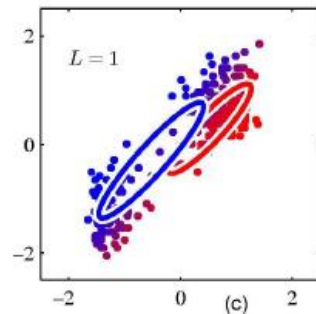
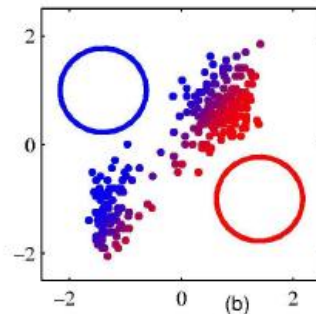
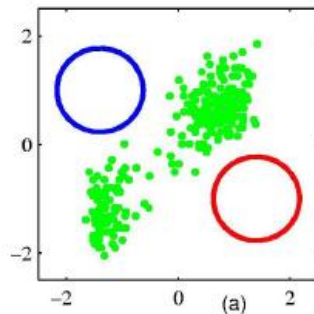
$$z_{nk} \in \{0, 1\}, \quad k = 1, \dots, K$$

最大化步（求导为0）  $\nabla_{\pi_k} Q = 0, \quad \nabla_{\mu_k} Q = 0, \quad \nabla_{\Sigma_k} Q = 0$

$$\left\{ \begin{aligned} \mu_k^{new} &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \\ \Sigma_k^{new} &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{new})(\mathbf{x}_n - \mu_k^{new})^T \\ \pi_k^{new} &= \frac{N_k}{N} \end{aligned} \right. \quad N_k = \sum_{n=1}^N \gamma(z_{nk})$$

# 参数估计—EM算法

## EM估计（示例图）



# 非参数估计——密度估计

密度估计问题：假定 $n$ 个样本 $x_1, \dots, x_n$ 独立同分布，服从于概率密度函数 $p(x)$ 。现在要估计 $p(x)$ 。

首先，样本 $x$ 落入区域 $\mathcal{R}$ 中的概率为 $P = \int_{\mathcal{R}} p(t) dt$ ，则

$$P \approx p(x)V.$$

其中， $V$ 是区域 $\mathcal{R}$ 所包含的体积。

同时，概率 $P$ 也可由落入区域 $\mathcal{R}$ 中的样本比例估计得到，即：

$$P \approx k/n$$

其中 $k$ 为落入区域 $\mathcal{R}$ 的样本数目， $n$ 为总样本数目。



$$P \approx p(x)V \approx k/n$$



$$p(x) \approx \frac{k/n}{V}$$

# 非参数估计—密度估计

密度估计问题：假定 $n$ 个样本 $x_1, \dots, x_n$ 独立同分布，服从于概率密度函数 $p(x)$ 。现在要估计 $p(x)$ 。

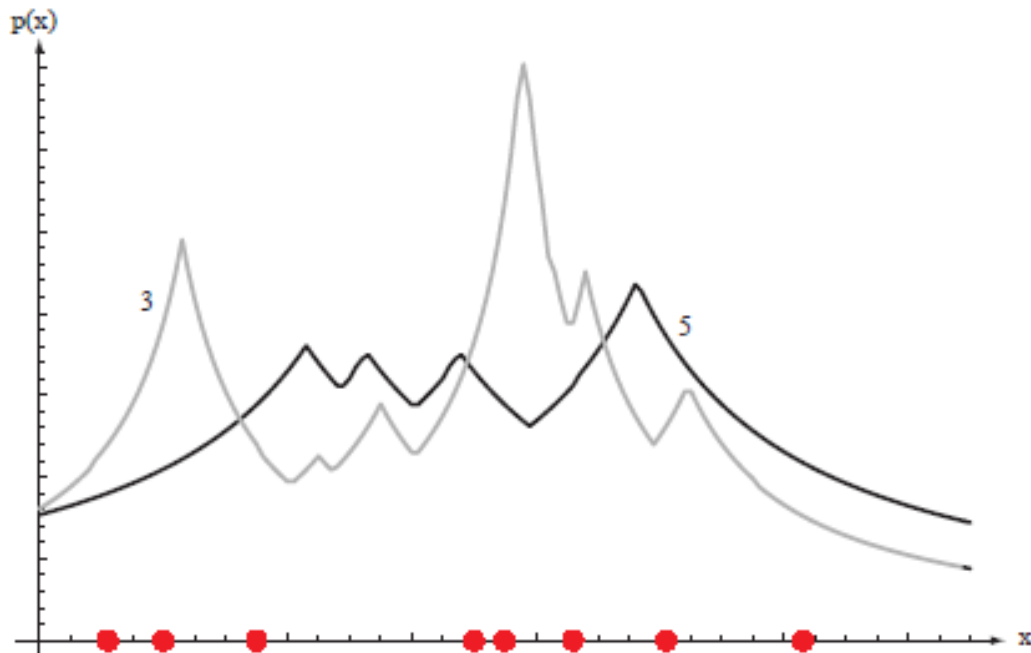
$p(x) \approx \frac{k/n}{V}$   为估计密度函数，需要计算局部样本数 $k$  和体积 $V$   


KNN(K-近邻)：固定局部样本数 $k$ ，体积 $V$ 变化

Parzen窗：固定体积 $V$ ，局部样本数 $k$  变化

# 非参数估计—KNN

KNN例子（1维）：





# 非参数估计—KNN

KNN例子：

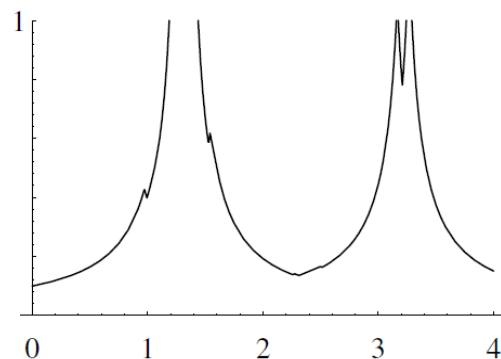
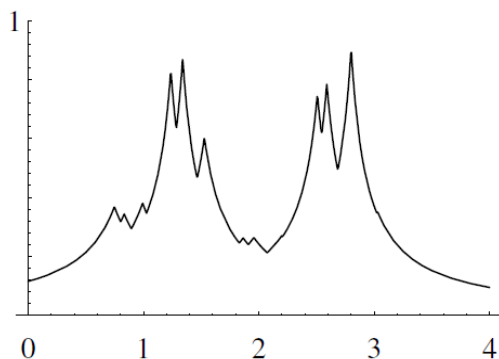
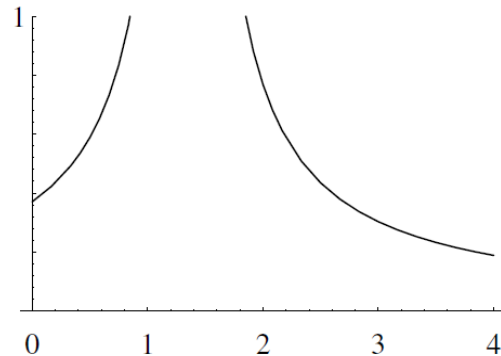
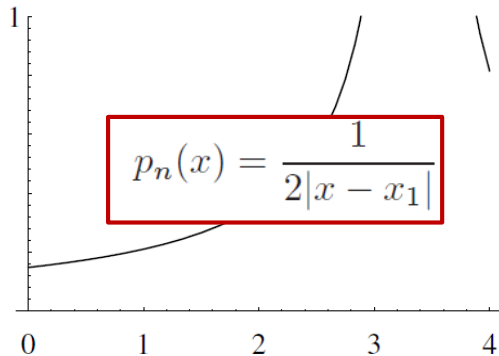
$$k_n = \sqrt{n}$$

$$\begin{matrix} n = 1 \\ k_n = 1 \end{matrix}$$

$$p_n(x) = \frac{1}{2|x - x_1|}$$

注意：当 $n$ 有限时，  
估计会非常崎岖不平

$$\begin{matrix} n = 16 \\ k_n = 4 \end{matrix}$$



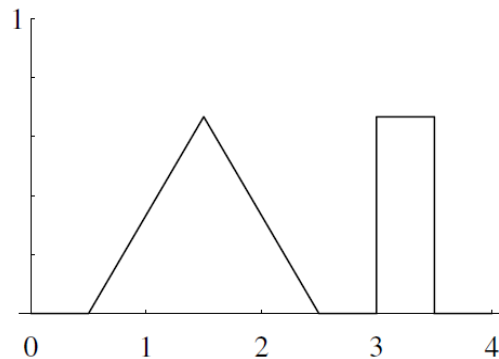
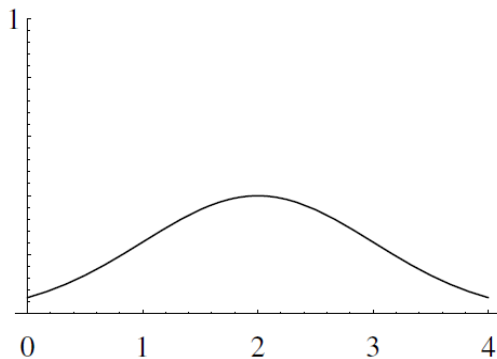
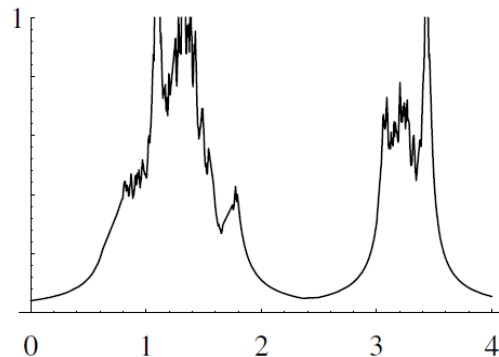
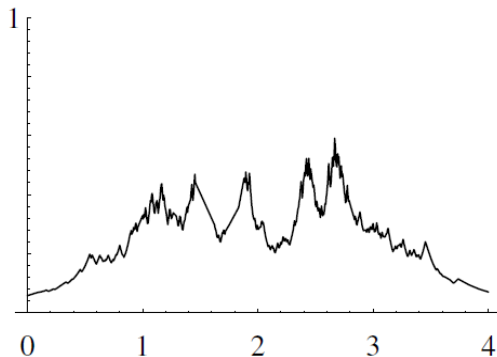
# 非参数估计—KNN

KNN例子：

$$k_n = \sqrt{n}$$

$$n = 256$$
$$k_n = 16$$

注意：当 $n$ 有限时，  
估计会非常崎岖不平



# 非参数估计—Parzen窗估计

Parzen窗估计（固定体积，变化局部样本数）

窗函数（方窗函数） 
$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq 1/2 \\ 0 & \text{otherwise.} \end{cases} \quad j = 1, \dots, d$$

该方窗函数带宽 $h_n$ 看作是1

体积为  $V_n = h_n^d$  的局部区域内样本数 
$$k_n = \sum_{i=1}^n \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

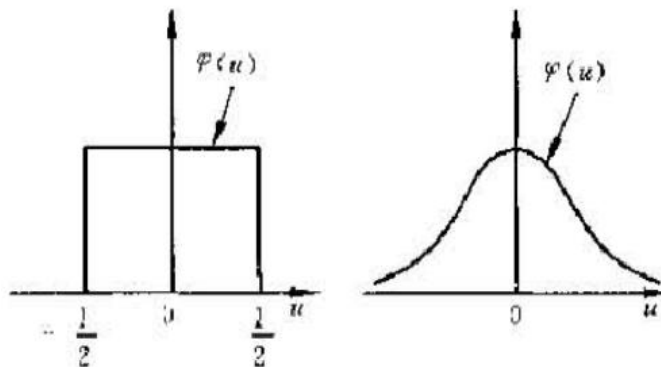
Parzen窗密度估计 
$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_i}{h_n}\right)$$

# 非参数估计—Parzen窗估计

核函数（方窗函数函数的推广）——直观上，可看作样本的权重

方窗核：
$$\varphi(u) = \begin{cases} 1 & |u_j| \leq 1/2 \\ 0 & \text{otherwise.} \end{cases} \quad j = 1, \dots, d$$

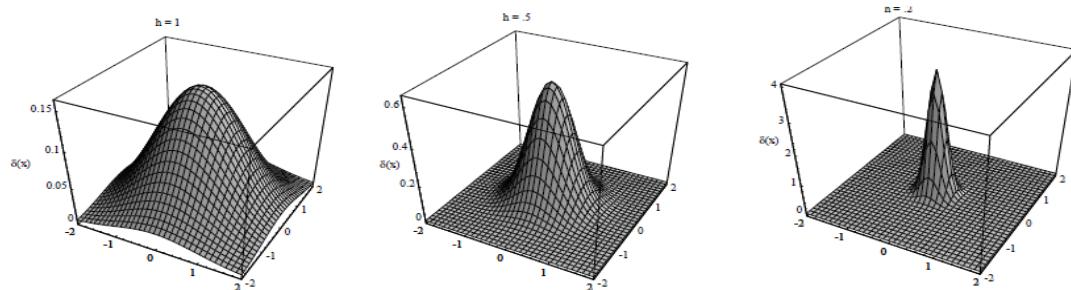
高斯核：
$$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^T u}$$



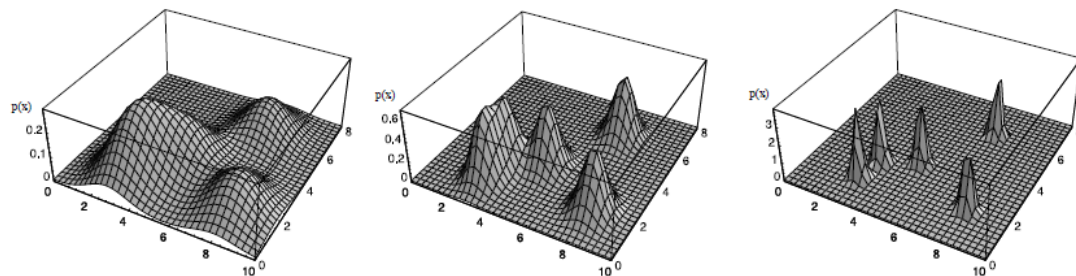
# 非参数估计—Parzen窗估计

Parzen窗估计例子：高斯核函数， $h = 1, 0.5, 0.2$

核函数



密度估计



$h$  (带宽) 越大，密度估计越平滑，反之亦然。 $h$  越小，容易过拟合。

# 总结

---

- ✓ 简述了机器学习与深度学习
- ✓ 简要讲解了贝叶斯决策理论，同时讲解一种常用的贝叶斯决策方法：最小误差贝叶斯决策
- ✓ 通过案例详细讲解了两种密度估计方法

# 参考文献

---

- ✓ 《模式分类》
- ✓ 《模式识别与机器学习》
- ✓ 深度学习相关论文（图片使用）







**感谢各位聆听 !**  
Thanks for Listening

