机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

信息与通信工程学院 网络搜索教研中心

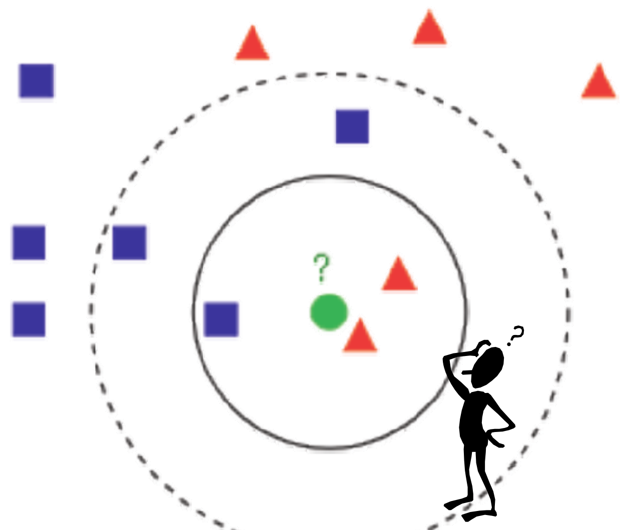
北京邮电大学

专题 一：基于实例的学习

- **内容提要**
  - 引言
  - 最近邻规则 **(Nearest Neighbor Rule)**
    - 非线性回归模型
  - 帕森窗**(Parzen Windows)**
    - Kernels
    - 瓦森-纳达拉亚估计器(Watson-Nadaraya Estimator)
  - 应用问题举例**:**
    - MNIST数据集 / VOC 与 BoW模型

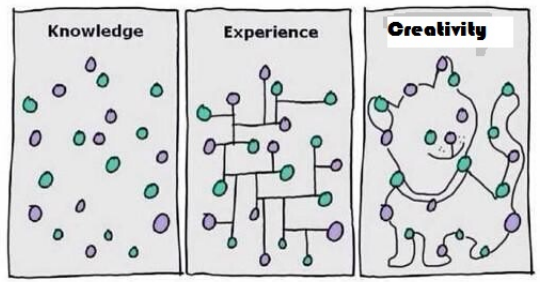
基于记忆的学习

- 查表法(Table lookup)
  - 数据库查询
  - 手机黑白名单



- 不具备泛化能力
- 不是学习的过程
  - 没有学习能力！

- 基于记忆的学习
  - 在记住的基础上，还要“学习”
    - 具备泛化能力
      - Lazy learning

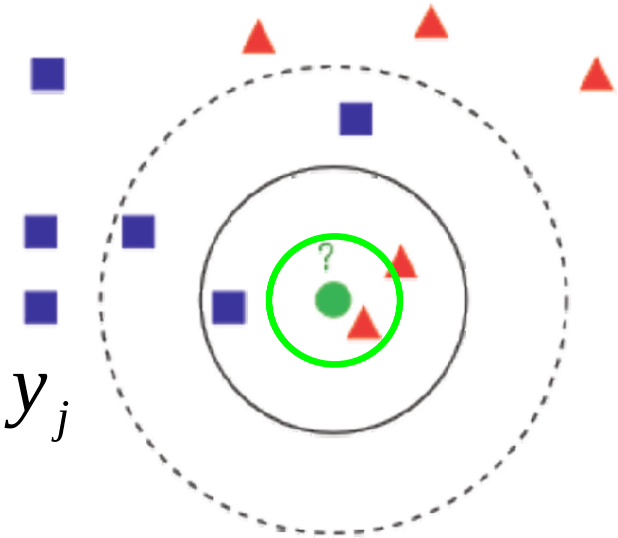


最近邻(1-NN:Nearest Neighbor)规则

- 最近邻规则
  - 局部邻域定义为与测试向量x最邻近的训练样本，即 $N_1(\mathbf{x}) = \arg\min_{\mathbf{x}_i} \mathrm{d}(\mathbf{x}, \mathbf{x}_i)$

    其中 $\mathbf{x}_i \in X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$
  - 把x的响应y定义为 $y = F(\mathbf{x}) = y_j$

    其中 $j : \mathbf{x}_j \in N_1(\mathbf{x})$



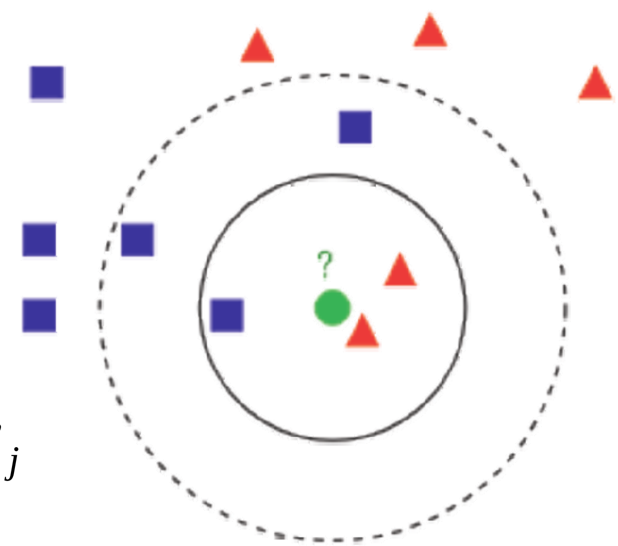
训练样本: $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

k近邻(k-NN)规则

- k近邻规则
  - 局部邻域定义为与测试向量**x**最邻近的**k**个训练样本，即 $N_k(\mathbf{x}) = \arg\min_{\mathbf{x}_i} \mathrm{d}(\mathbf{x}, \mathbf{x}_i)$

    其中 $\mathbf{x}_i \in X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

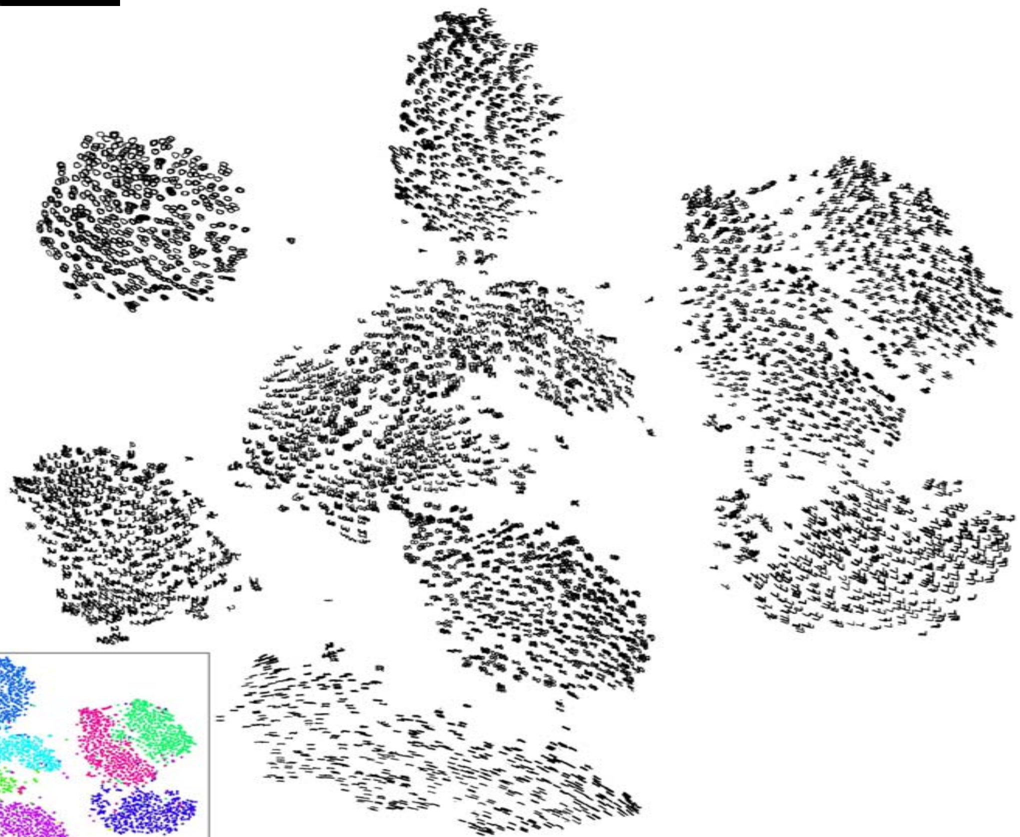


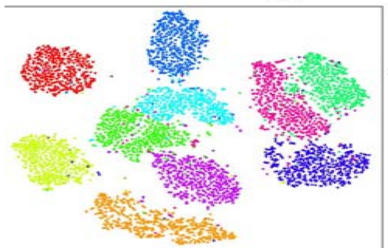
  - 把*x*的响应**y**定义为
    - 回归问题: $y = F(\mathbf{x}) = \frac{1}{k} \sum_{j \in N_k(\mathbf{x})} y_j$
    - 分类问题:
      - 使用多数表决规则，使用表决获胜的类别来定义**x**的类别

训练样本: $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$

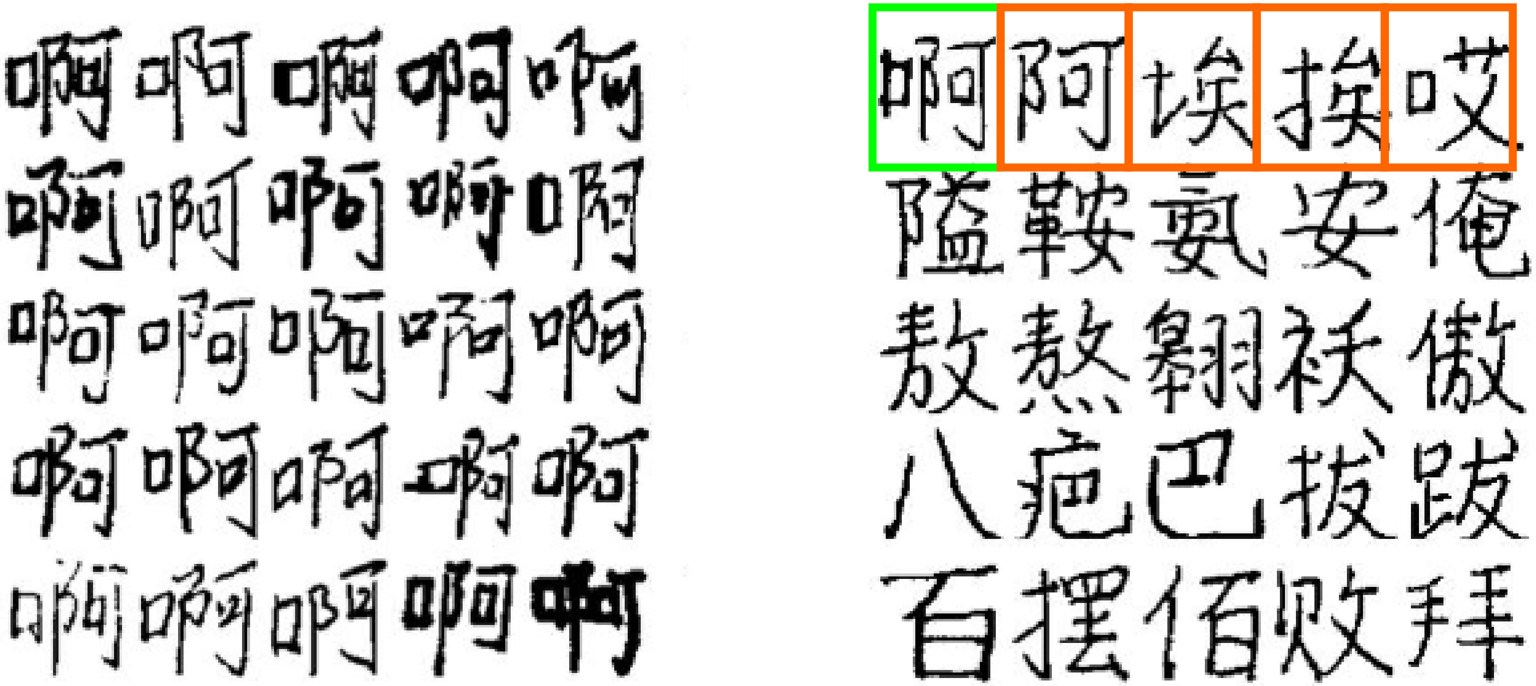
$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$

应用 1：手写数字图像识别





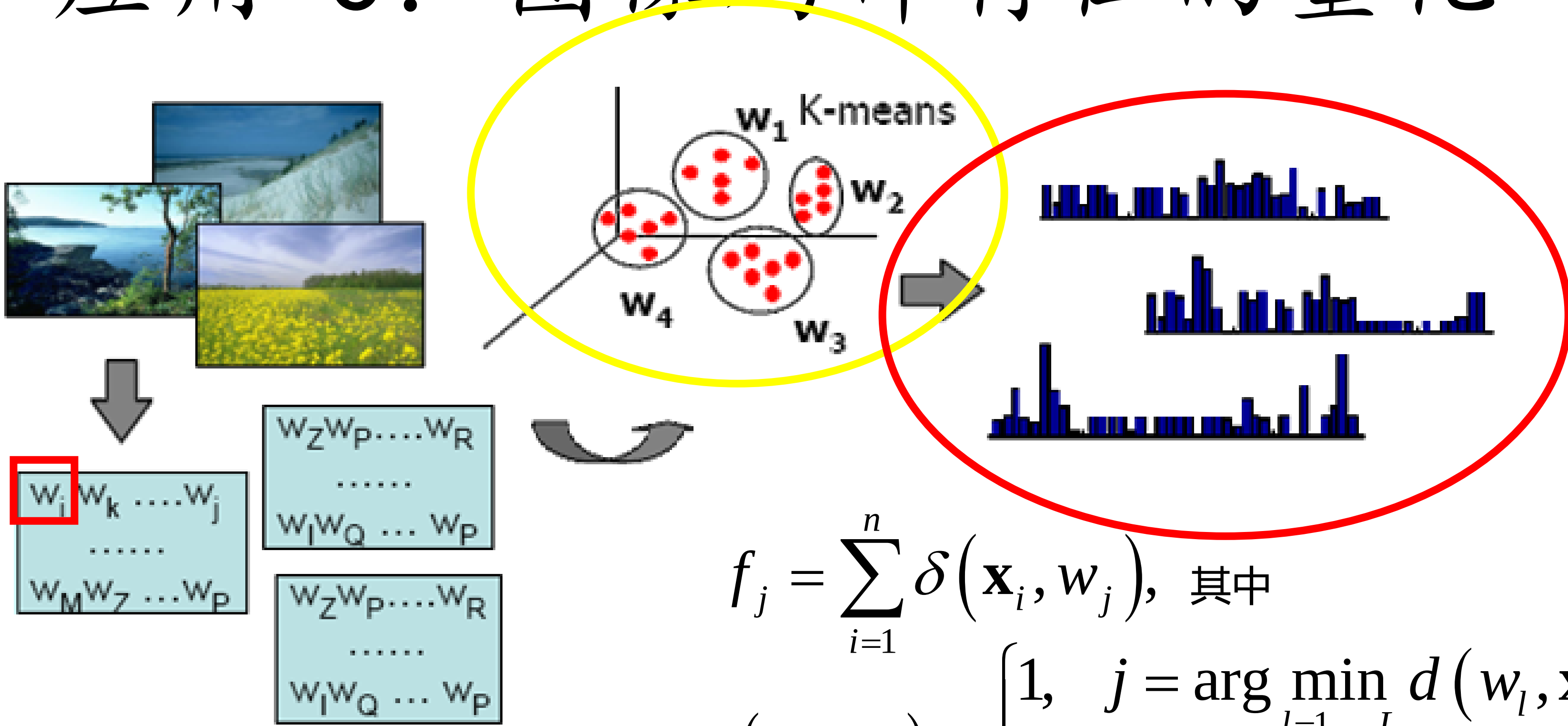
应用 2：手写汉字图像识别

- 

(a) (b)

- 以HCL2000数据库为例：
  - 手写汉字识别任务的类别数: **3755 !**
  - 一般的策略: 粗分类 + 细分类

应用 3：图像局部特征的量化

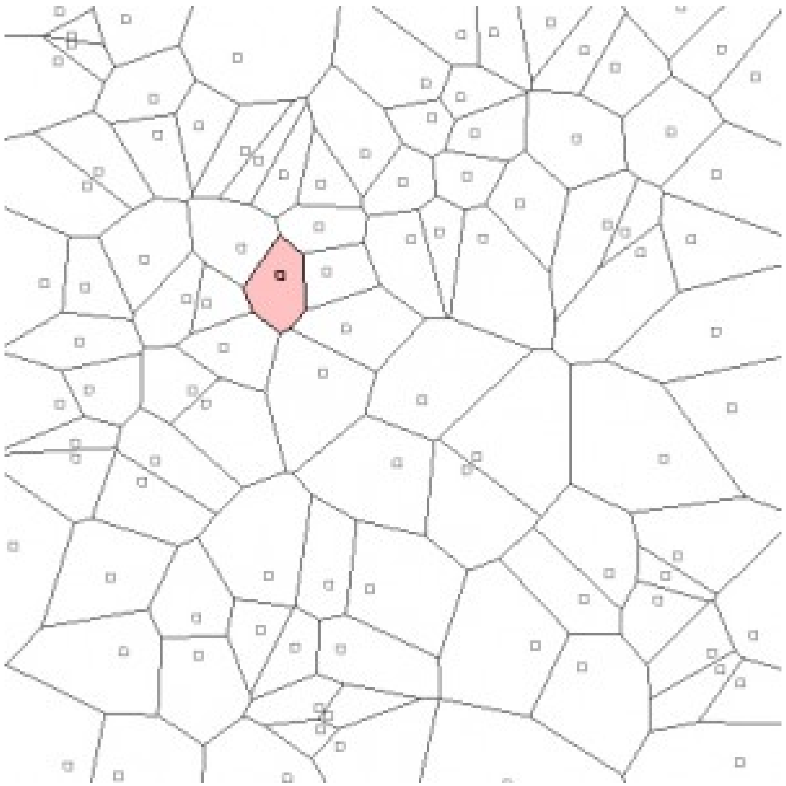


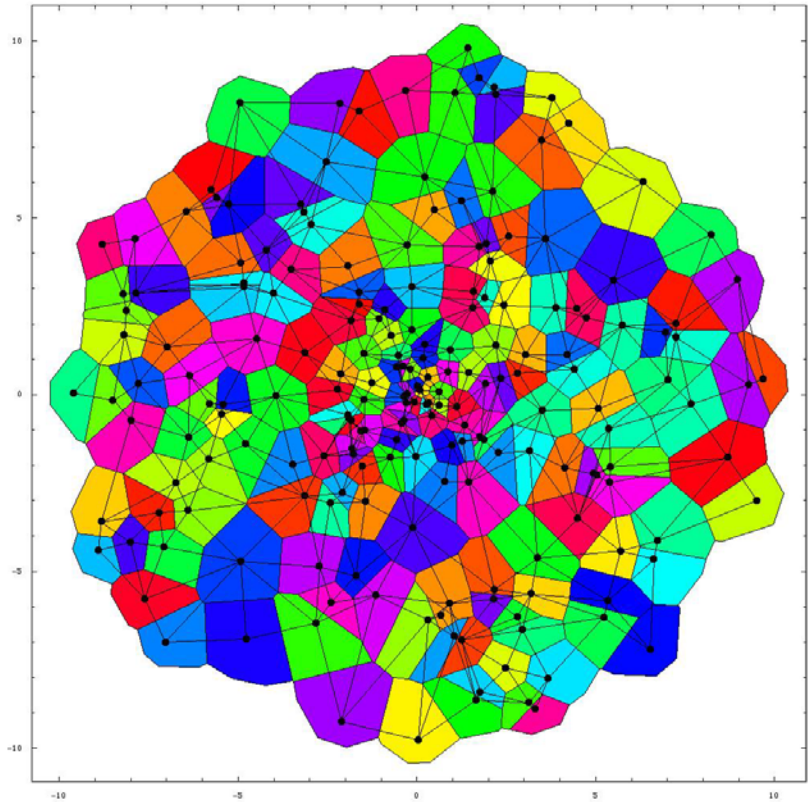
$$f_j = \sum_{i=1}^{n} \delta\left(\mathbf{x}_i, w_j\right),$$ 其中

$$\delta\left(\mathbf{x}_i, w_j\right) = \begin{cases} 1, & j = \arg\min\limits_{l=1,\dots,L} d\left(w_l, \mathbf{x}_i\right) \\ 0, & j \neq \arg\min\limits_{l=1,\dots,L} d\left(w_l, \mathbf{x}_i\right) \end{cases}$$

- 使用1-NN量化局部特征
  - 码字的硬指派(codeword hard assignment)

码字量化所得的边界图





– 码字(硬)指派的过程实质上把特征空间进行了(硬)划分

– 构造BoV直方图的过程是一个密度估计过程, 即统计落入各个cell/cube中的数据点的个数的过程

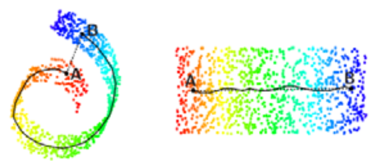
应用 4：k-NN回归

- 给定统计数据是每月15号的价格
  - 左图的蓝色折线给出的是**k-NN**回归方法的结果
    - 其中，k=2

    - 1-NN的回归结果呢？
      - 右图线段所示阶梯状函数

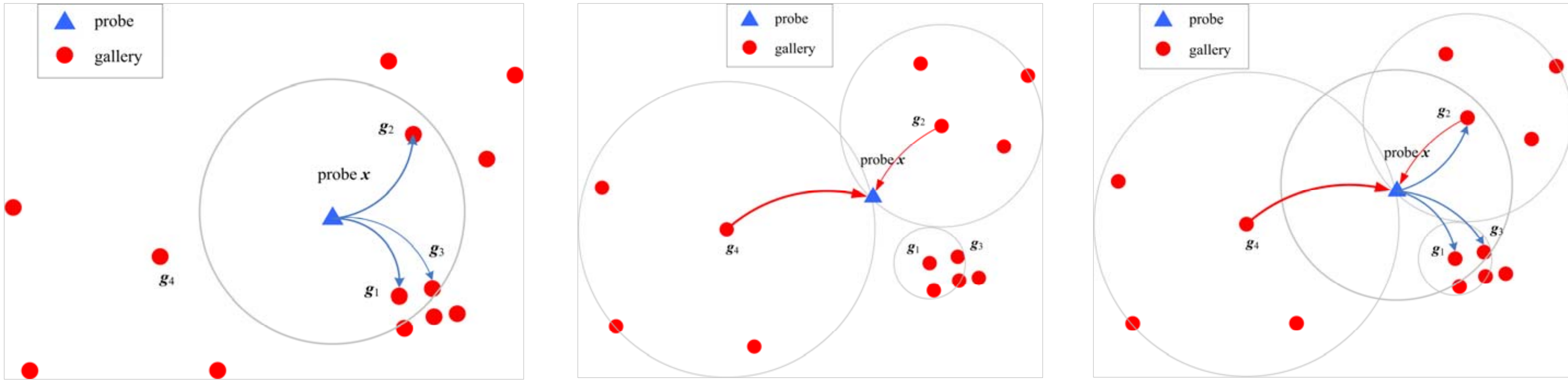
k-近邻......too simple？

- YES! But it’s too useful…
  - k近邻是一项应用广泛的技术，除了直接用于分类和回归之外，还与各种算法相结合
    - k-nn + LDA (Hastie & Tibshirani, PAMI1996)
    - k-nn + SVM (Support Vector Machine) (CVPR2006)
      - » knn + large margin = LMNN, (NIPS2006)
    - k-nn for collaborative filtering\recommendation
    - k-nn +SRC = Local SRC (ICPR2010)
    - k-nn +MC = high rank matrix completion (AISTAT2012)
    - k-nn + ***X*** = algorithms in Manifold Learning, e.g., LLE, Isomap, LaplaceEigenmap....



kNN: 换一个观察方向会如何？

- “我认识TA，TA不认识我”
  - 前半句：
    - 我的k-nearest neighbors包含TA
  - 后半句：
    - TA的k-nearest neighbors不包含我



(a) k-NN (b) k-INN (Inverse NN) (c) k-RNN (Reciprocal NN)

[1] Ruopei Guo, C.-G. Li, et al. "Density-Adaptive Kernel based Re-Ranking for Person Re-Identification", submitted to ICPR 2018.

Q / A

- Any Question? …

专题一：基于实例的学习

- **内容提要**
  - 引言
  - 最近邻规则 **(Nearest Neighbor Rule)**
    - 从非线性回归模型看k-近邻回归
  - 帕森窗**(Parzen Windows)**
    - Kernels
    - 瓦森-纳达拉亚估计器(Watson-Nadaraya Estimator)
  - 应用问题举例**:**
    - MNIST数据集 / VOC 与 BoW模型

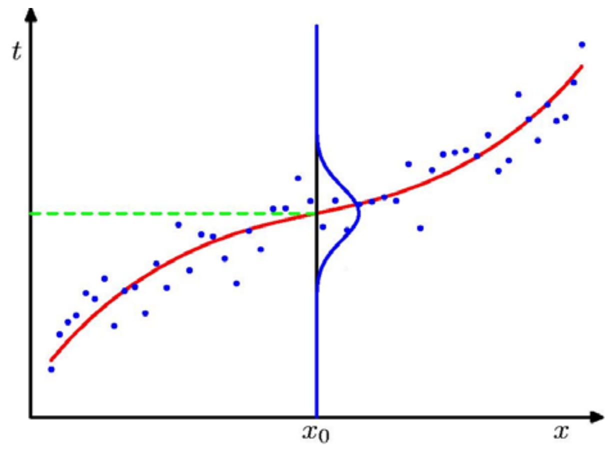
非线性回归模型

- **考虑一个非线性回归模型**
  - 设X是随机输入向量，Y是实数值随机标量，联合分布密度p(x,y)，寻找一个确定性函数f(·)，使得用f(x)可以很好地近似与输入向量x相对应的y，即

$$y \approx f\left(\mathbf{x}\right)$$

- **当使用平方误差损失函数** $\left(y - f\left(\mathbf{x}\right)\right)^2$ **时，回归模型的解为：**

$$f\left(\mathbf{x}\right) = \mathbf{E}\left(Y \mid X = \mathbf{x}\right)$$

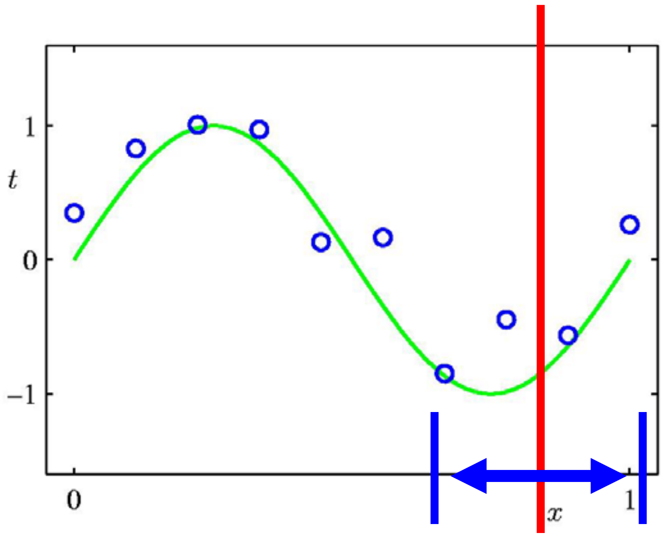


非线性回归模型到k近邻回归

- k-近邻回归可以看作条件期望的样本估计

$$f(\mathbf{x}) = \mathbf{E}[Y \mid X = \mathbf{x}] \quad \Rightarrow \quad F(\mathbf{x}) = \frac{1}{k}\sum_{i \in N_k(\mathbf{x})} y_i$$

- 从左式到右式，经过两次近似：
  1. **使用在样本数据上求平均值近似期望**
  2. **把在点x上取条件放宽为在靠近测试点x的某邻域上取条件**



思考: k-近邻分类规则呢?

专题一：基于实例的学习

- **内容提要**
  - 引言
  - 最近邻规则 **(Nearest Neighbor Rule)**
    - 非线性回归模型
  - 帕森窗**(Parzen Windows)**
    - 密度估计问题的引出
    - Kernels
    - 瓦森-纳达拉亚估计器(Watson-Nadaraya Estimator)
  - 应用问题举例**:**
    - MNIST数据集 / VOC 与 BoW模型

非线性回归模型

- **考虑一个非线性回归模型**
  – 设X是随机输入向量, Y是实数值随机标量, 联合分布密度p(x,y), 寻找一个函数f(x), 实现通过X预测Y

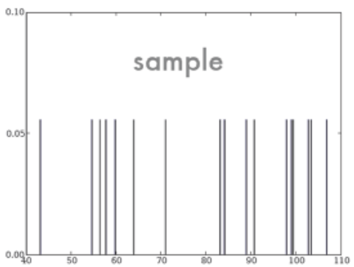
$$y \approx f\left(\mathbf{x}\right)$$

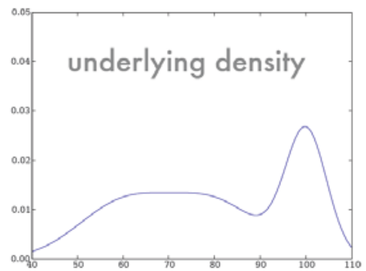
- **回归模型的解**

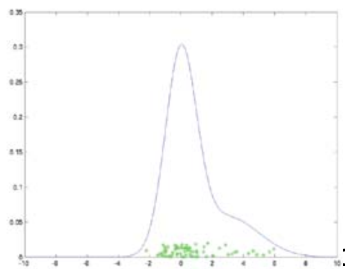
$$f\left(\mathbf{x}\right) = \mathbf{E}\left(Y \mid X = \mathbf{x}\right) = \frac{\int_{-\infty}^{\infty} y \cdot p_{X,Y}\left(\mathbf{x}, y\right) dy}{p_X\left(\mathbf{x}\right)}$$

  – 需要估计$p_{X,Y}\left(\mathbf{x}, y\right)$和 $p_X\left(\mathbf{x}\right)$

  - **这是密度估计问题**







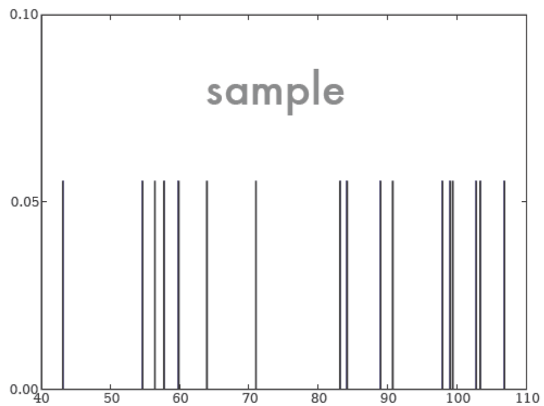
密度估计问题的引出

- 定义delta函数如下:

$$\delta(\mathbf{x},\mathbf{x}_i)=\begin{cases}1 & \text{if} \quad \mathbf{x}=\mathbf{x}_i \\ 0 & \text{else} \quad \mathbf{x}\neq\mathbf{x}_i\end{cases}$$

  – 在样本点上取值为**1**，其它位置为**0**

- 给定一个样本集{ x_i }，则相当于给定一个朴素(naïve)的经验分布直方图



$$p(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{x},\mathbf{x}_i)$$

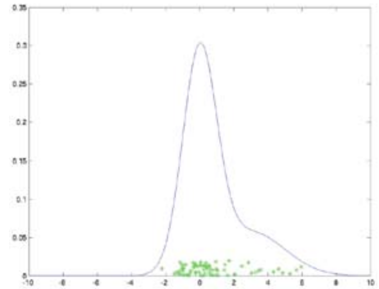
  – 特点：处处不连续
  – 缺点：处处不连续，则没有任何泛化能力(稍偏离样本即为0)

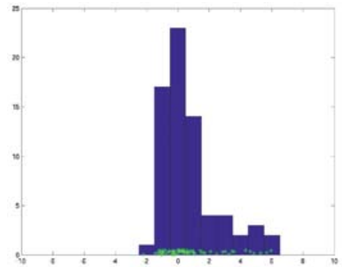
密度估计的直方图法

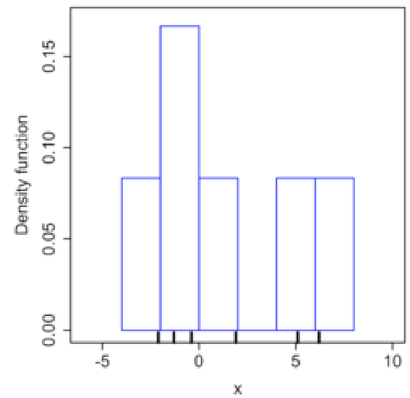
- 概率分布的直方图估计
  - 离散形式：概率分布 $\hat{P}_i = \dfrac{n_i}{n}$
    - 理论依据：大数定律
      - » Hoeffding不等式

$$\Pr\left\{\left|\nu_n - \mu\right| > \varepsilon\right\} \le 2\exp\left(-2\varepsilon^2 n\right)$$

  - 连续形式：分布密度
    - 1维空间 $\hat{p}_i = \dfrac{P_i}{\Delta_i} = \dfrac{n_i}{n \cdot \Delta_i}$
    - 高维空间 $\hat{p}_i = \dfrac{\hat{P}_i}{\Delta_i} = \dfrac{n_i}{n \cdot \Delta_i}$
  - 直接采用直方图法将遭遇维数灾难！







密度估计基本公式

- 密度估计基本公式的导出
  - 两个假设
    - 如果样本数 n 足够大, 则落入以 x 为中心的体积 V 的邻域内的样本点个数 K 近似为 $P\cdot n$, 其中 P 为样本落入 x 的邻域内的概率
    - 如果包含 x 的邻域足够小，那么概率密度函数 $p(x)$ 可以近似为常函数，即 $P \approx p(x)\ V$，其中 V 为邻域的体积
  - 两者合起来即得

$$K = p(\mathbf{x})Vn \quad \Rightarrow \quad p(\mathbf{x}) = \frac{K}{n \cdot V}$$

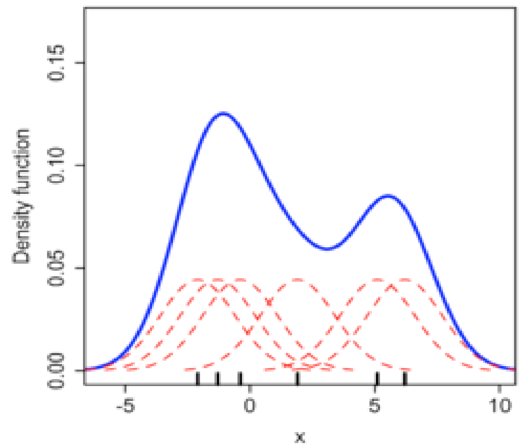
- 注意到：K 与 V 存在函数关系

当n→无穷大, V随n收缩且K随n增长, 则密度估计收敛于真实密度(Duda&Hart 1973)

密度估计的两种思路

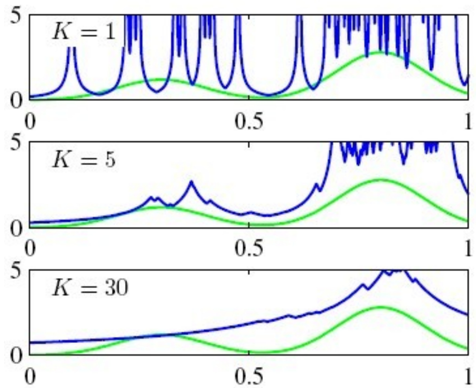
- 密度估计基本公式

$$p(\mathbf{x}) = \frac{K}{n \cdot V}$$



- 两种思路
  - 固定**V**，根据数据确定**K**
    - **Kernel密度估计技术**

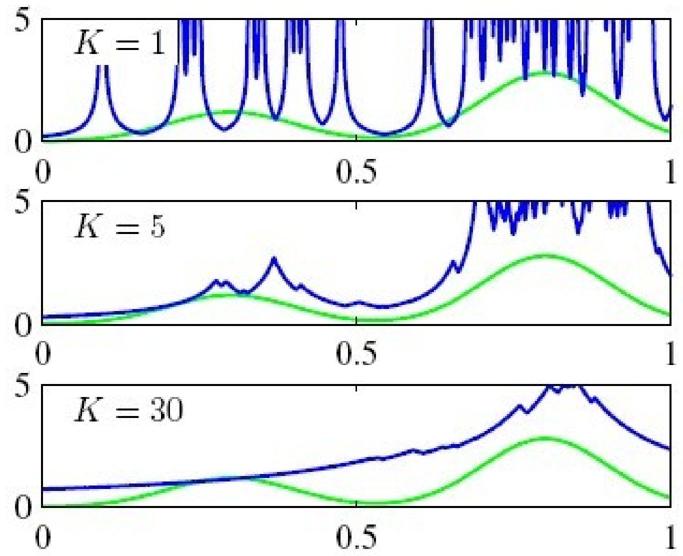
      其中 $V = h^m$ ，h为邻域半径, 即带宽参数
  - 固定**K**，根据数据确定**V**
    - **K-近邻密度估计技术**



K近邻密度估计技术

- 密度估计基本公式

$$p(\mathbf{x}) = \frac{K}{n \cdot V}$$



  - 固定K，根据数据确定V
    - 邻域半径(带宽参数)可变

  其中 $V = h_k^{\ m}$，h_k为到第k个近邻点的距离(即邻域半径)

  - **举例:**
    - 考虑1维数据的情况

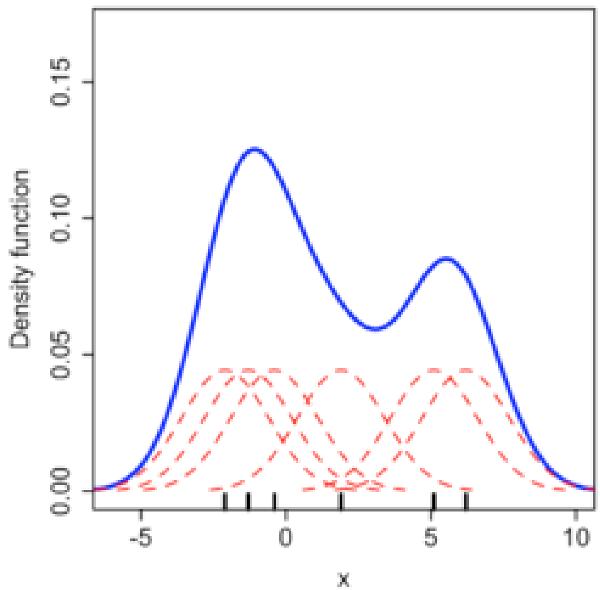
  k-NN密度估计公式： $p(x) = \dfrac{1}{2n} \dfrac{k}{\left| x - x^{(k)} \right|}$

  其中 $x^{(k)}$ 表示 x 的第k个近邻

核密度估计技术

- 密度估计基本公式

$$p(\mathbf{x}) = \frac{K}{n \cdot V}$$



– 固定V，根据数据确定K

- 邻域半径(带宽参数)固定

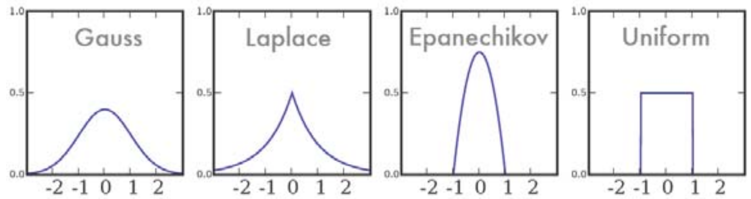
其中 $V = h^m$

– 举例:

- Parzen-Rosenblatt 密度估计器

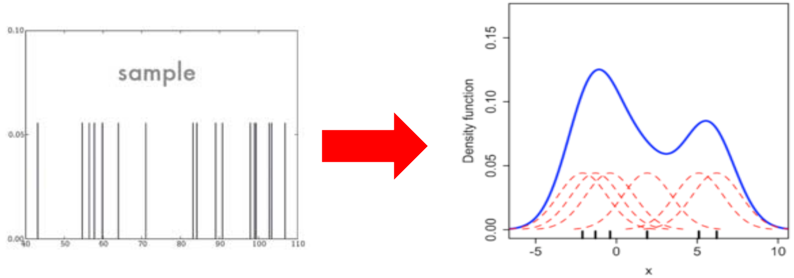
  – 典型的核函数图像:

$$p(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^m} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h} \right),$$



$$k(t) = \frac{3}{4}(1 - t^2), |t| \le 1$$

帕森窗(Parzen Window)

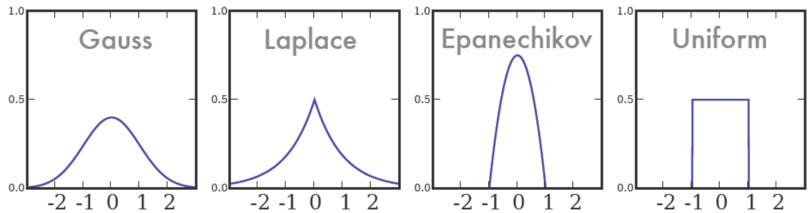


- 核密度估计法
  - 使用非负的光滑核函数

$$p(\mathbf{x})=\frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{x},\mathbf{x}_i) \quad \Rightarrow \quad p(\mathbf{x})=\frac{1}{nh^m}\sum_{i=1}^{n}k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right),$$

  - 核函数$k(x)$与概率密度函数性质相同
    - **k(x)是关于x的连续有界实函数，偶函数，且在原点取得最大值**
    - **在核$k(x)$的曲面下的总体积等于1，即对于m维向量x有**

$$\int_{R^m}k(\mathbf{x})\,d\mathbf{x}=1$$



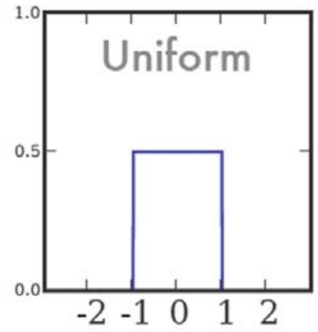
例1: k 近邻分类规则的导出

- k-近邻分类规则
  - 密度估计: $p(\mathbf{x}) = \dfrac{K}{n \cdot V}$
  - 类别 ω_j 的先验概率估计: $P(\omega_j) = \dfrac{n_j}{n}$,
  - 类别 ω_j 的概率密度估计: $p(\mathbf{x} \mid \omega_j) = \dfrac{K_j}{n_j V}$,
  - 计算后验概率: $p(\omega_j \mid \mathbf{x}) = \dfrac{p(\mathbf{x} \mid \omega_j) P(\omega_j)}{p(\mathbf{x})} = \dfrac{K_j}{K}$
    - 分类规则:

$$\boxed{j^* = \arg\max_{j=1,\ldots,C} p(\omega_j \mid \mathbf{x}) = \frac{K_j}{K}}$$

多数表决

Kernel密度估计，核函数如下:



例2：基于核密度估计的回归模型(1/2)

- 计算条件期望E(y|x)

$$f(\mathbf{x}) = E(y \mid \mathbf{x}) = \int_{-\infty}^{\infty} y \cdot p_{Y|X}(y \mid \mathbf{x})\, dy$$

– 其中

$$\hat{p}_X(\mathbf{x}) = \frac{1}{n \cdot h^m} \sum_{i=1}^{n} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right),$$

$$\hat{p}_{X,Y}(\mathbf{x}, y) = \frac{1}{n \cdot h^{m+1}} \sum_{i=1}^{n} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) k\left(\frac{y - y_i}{h}\right)$$

– 计算

$$\int_{-\infty}^{\infty} y \cdot \hat{p}_{X,Y}(\mathbf{x}, y)\, dy = \frac{1}{n \cdot h^{m+1}} \int_{-\infty}^{\infty} \sum_{i=1}^{n} k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \cdot y \cdot k\left(\frac{y - y_i}{h}\right) \cdot dy$$

$$= \frac{1}{n \cdot h^m} \sum_{i=1}^{n} y_i k\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right)$$

例2：基于核密度估计的回归模型(2/2)

- 计算条件期望E(y|x)

$$f(\mathbf{x}) = E(y \mid \mathbf{x}) = \int_{-\infty}^{\infty} y \cdot p_{Y|X}(y \mid \mathbf{x})\, dy$$

– 其中

$$\hat{p}_X(\mathbf{x}) = \frac{1}{n \cdot h^m} \sum_{i=1}^{n} k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right),$$

$$\int_{-\infty}^{\infty} y \cdot \hat{p}_{X,Y}(\mathbf{x}, y)\, dy = \frac{1}{n \cdot h^m} \sum_{i=1}^{n} y_i k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)$$

– 回归函数的核密度估计

$$\hat{f}(\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{j=1}^{n} k\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right)} = \sum_{i=1}^{n} y_i \frac{k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{j=1}^{n} k\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right)} = \sum_{i=1}^{n} y_i W_{n,i}(\mathbf{x})$$

例3: Nadaraya-Watson核回归估计器(1/2)

- Nadaraya-Watson回归估计器
  - 定义归一化加权函数 $W_{n,i}(\mathbf{x}) = \dfrac{k\left(\dfrac{\mathbf{x}-\mathbf{x}_i}{h}\right)}{\sum_{i=1}^{n} k\left(\dfrac{\mathbf{x}-\mathbf{x}_i}{h}\right)}$
    - 其中 $\sum_{i=1}^{n} W_{n,i}(\mathbf{x}) = 1$
  - 由此，得到**N-W**核回归估计：

$$\hat{f}(\mathbf{x}) = \sum_{i=1}^{n} y_i W_{n,i}(\mathbf{x})$$

例3: Nadaraya-Watson核回归估计器(2/2)

- 特例: 径向基函数(Radial Basis Function)
  - **假设核函数K(x)球对称**

$$k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)=k\left(\frac{\left\|\mathbf{x}-\mathbf{x}_i\right\|}{h}\right)$$

举例:

$$k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right)=\frac{1}{\left(2\pi\sigma^2\right)^{m/2}}\exp\left(\frac{\left\|\mathbf{x}-\mathbf{x}_i\right\|^2}{2\sigma^2}\right)$$

  - **规范化RBF:** $\Psi_n\left(\mathbf{x}-\mathbf{x}_i\right)=k\left(\frac{\left\|\mathbf{x}-\mathbf{x}_i\right\|}{h}\right)/\sum_{i=1}^{n}k\left(\frac{\left\|\mathbf{x}-\mathbf{x}_i\right\|}{h}\right)$

其中$\sum_{i=1}^{n}\Psi_n\left(\mathbf{x}-\mathbf{x}_i\right)=1$

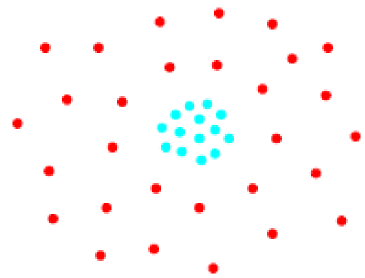
得出核回归估计:

$$\hat{f}\left(\mathbf{x}\right)=\sum_{i=1}^{n}y_i\Psi_n\left(\mathbf{x}-\mathbf{x}_i\right)$$

两种思路的融合: 自适应密度估计

- 在Kernel密度估计技术中参数不容易选定
  - 采用**K-**近邻所提供的局部邻域尺度信息定义带宽参数

$$p(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^m}k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right),$$



$$p(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i^m}k\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right),$$

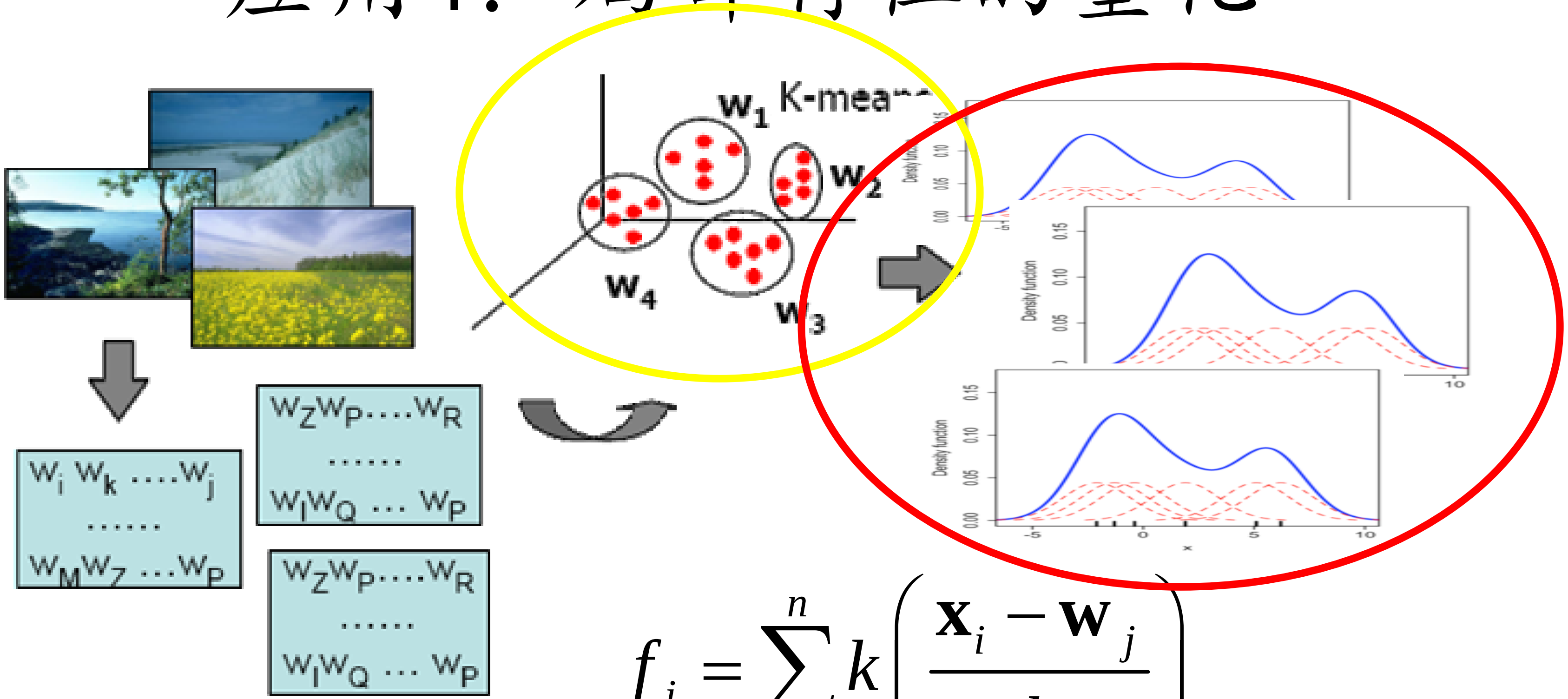
- 其中 $h_i = \left\|\mathbf{x}_i^{(k)} - \mathbf{x}_i\right\|_2$，$\mathbf{x}_i^{(k)}$是 $\mathbf{x}_i$ 的第k个近邻

[1] Leo Breiman , William Meisel, and Edward Purcell, "Variable Kernel Estimates of Multivariate Densities", Technometrics, Vol.19, No.2, pp.135-144.
[2] L. Zelnik-manor and P. Perona, "Self-tuning spectral clustering," Advances in Neural Information Processing Systems (NIPS) 17, 2004, pp. 1601-1608.

专题 一：基于实例的学习

- **内容提要**
  – 引言
  – 最近邻规则 **(Nearest Neighbor Rule)**
    - 非线性回归模型
  – 帕森窗**(Parzen Windows)**
    - 密度估计问题的引出
    - Kernels
    - 瓦森-纳达拉亚估计器(Watson-Nadaraya Estimator)
  – 应用问题举例**:**
    - MNIST数据集 / VOC 与 BoW模型

应用4：局部特征的量化



$$f_j = \sum_{i=1}^{n} k\left(\frac{\mathbf{x}_i - \mathbf{w}_j}{h}\right)$$

- 使用核密度估计器量化局部特征
  - 码字的软指派(**codeword soft assignment)**

[1] Gemert and Geusebroek: “Kernel codebooks for scene categorization”, ECCV 2008; IEEE Trans. PAMI, 2010.

专题 一：基于实例的学习

- **内容提要**
  - 引言
  - 最近邻规则 **(Nearest Neighbor Rule)**
    - 非线性回归模型
  - 帕森窗**(Parzen Windows)**
    - 密度估计问题的引出
    - Kernels
    - 瓦森-纳达拉亚估计器(Watson-Nadaraya Estimator)
  - 应用问题举例**:**
    - MNIST数据集 / VOC 与 BoW模型

参考阅读

- A.R. Webb, Statistical Pattern Recognition （统计模式识别） Chpt-3

- R. Duda, P. Hart, D. Stork, Pattern Classification（模式分类） Chpt-4

Q / A

- Any Question? …

经典问题的拓展与延伸

- 对于分类和回归问题，不能局限于经典问题的模式，要结合新的问题设定，比如:
  - 当训练数据和测试数据之间并不一致时，如何解决学习问题？这是领域自适应**(Domain Adaptation)**问题
  - 多视图数据
  - 多模态数据
  - 知识的迁移问题

补充材料

- **内容提要**
  - 最近邻分类器错误率的界
  - 变分法求非线性回归模型
  - 集中不等式(Concentration inequality)及其应用

最近邻的性能保证

- 最近邻分类的错误率分析

$$\varepsilon_* \le \varepsilon \le \varepsilon_* \left(2 - \frac{C}{C-1} \varepsilon_* \right)$$

  – 其中 ε_* 为贝叶斯分类错误率
    - 一般情况下，ε_* 较小，因此，我们有: $\varepsilon_* \le \varepsilon \le 2\varepsilon_*$

    - 在数据独立同分布，样本数趋于无穷时，最近邻法的误差率不会高于贝叶斯分类错误率的2倍
      – Cover &Hart (1967)
    - k-近邻法的错误率要低于最近邻法

补充材料

- **内容提要**
  - 最近邻分类器错误率的界
  - 变分法求非线性回归模型
  - 集中不等式**(Concentration inequality)**及其应用

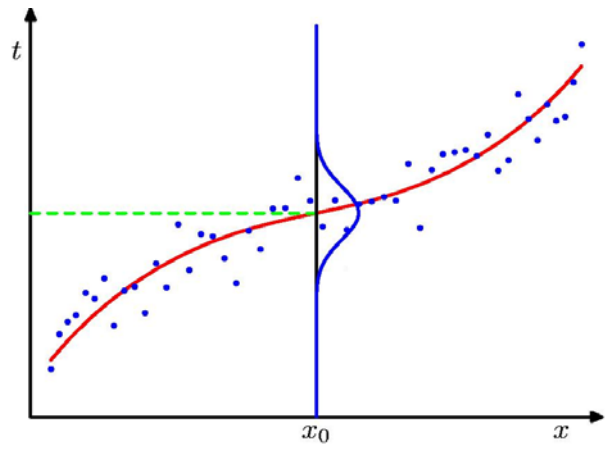
非线性回归模型

- **考虑一个非线性回归模型**
  – 设X是随机输入向量，Y是实数值随机标量，联合分布密度p(x,y)，寻找一个确定性函数f(·)，使得用f(x)可以很好地近似与输入向量x相对应的y，即

$$y \approx f(\mathbf{x})$$

- **当使用平方误差损失函数 $\left(y - f(\mathbf{x})\right)^2$ 时，回归模型的解为：**

$$f(\mathbf{x}) = \mathbf{E}(Y \mid X = \mathbf{x})$$



Fréchet微分与泛函极值

- 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：

$$d\varepsilon(f,h)=\left[\frac{d}{d\beta}\varepsilon(f+\beta\cdot h)\right]_{\beta=0}$$

- 泛函极值条件:
  – 函数$\boldsymbol{F(x)}$为泛函的一个相对极值的必要条件是：

$$d\varepsilon(f,h)=0,\quad \forall h(x)$$

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

求解回归模型

- **考虑一个非线性回归模型**
  - 设X是随机输入向量，Y是实数值随机标量，联合分布密度$p(x,y)$，寻找一个函数f(x)，给定输入X的值预测Y

$$y = f(\mathbf{x}) + \varepsilon$$

- 优化问题建模与求解
  - 基于平方误差损失函数，定义期望误差

$$\varepsilon(f) = \mathbf{E}\left[Y - f(X)\right]^2 = \int\int \left\{y - f(\mathbf{x})\right\}^2 p(\mathbf{x}, y)\, d\mathbf{x}dy$$

  - 对目标泛函求极值点，即为$f(x)$：

令 $d\varepsilon(f, h) = 0$ 得出 $f(x) = \mathbf{E}\left[Y \mid X = x\right]$

    - 如果选择平方误差作为损失函数，则回归模型的解为条件期望(条件均值)

补充材料

- **内容提要**
  - 最近邻分类器错误率的界
  - 变分法求非线性回归模型
  - 集中不等式(Concentration inequality)及其应用

大数定律与集中不等式

说说“靠谱儿”

- 大数定律:
  - 独立随机变量的算术平均值以很大概率趋近于其数学期望

$$\lim_{n\to\infty} \mathbf{P}\left(\left|\bar{\mu}_n - \mu\right| \geq \varepsilon\right) = 0,\ \forall \varepsilon > 0$$

- 集中不等式 (Concentration inequality)
  - 提供随机变量偏离某值(比如期望)的概率界
    - Markov / Chebyshev / Chernoff / Hoeffding / Bennett / Bernstein / McDiarmid 不等式

$$\mathbf{P}\left(\left|\bar{\mu}_n - \mu\right| < \varepsilon\right) > 1 - \delta$$

  - “以概率1-δ保证偏差小于ε”

常用的集中不等式

- Hoeffding不等式
  - 如果 $X_1,\ldots,X_n$ 是n个独立随机变量，假设 X_i 有界，即 $X_i \in [a_i,b_i]$ ，

令 $$S_n = \frac{1}{n}\left(X_1 + \ldots + X_n\right)$$

则有：

$$\mathbf{P}\left(\left|S_n - \mathbf{E}\left(S_n\right)\right| \geq \varepsilon\right) \leq 2\exp\left(-\frac{2\varepsilon^2 n^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right)$$

集中不等式的应用举例

- 例1: 大数定律的收敛速度

– Hoeffding不等式

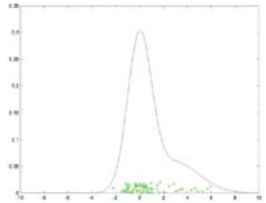
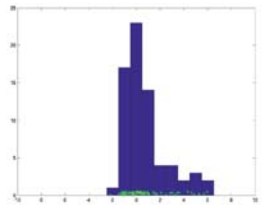
$$\mathbf{P}\left(\left|\bar{\mu}_n-\mu\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{2\varepsilon^2 n}{\frac{1}{n}\sum_{i=1}^{n}\left(b_i-a_i\right)^2}\right)=2\exp\left(-\frac{2\varepsilon^2 n}{c^2}\right)$$

– 令 $\delta=2\exp\left(-\frac{2\varepsilon^2 n}{c^2}\right)$

$$\Rightarrow\quad \varepsilon=c\sqrt{\frac{1}{2n}\log\frac{2}{\delta}}$$ 其中 $$c^2=\frac{1}{n}\sum_{i=1}^{n}\left(b_i-a_i\right)^2$$

例2: 直方图估计分布的收敛速度

- 概率分布的直方图估计
  - 离散形式: 概率分布 $\widehat{P}_i = \dfrac{n_i}{n}$

  - 考虑其中一个bin，利用Hoeffding 不等式得:

$$\Pr\left\{\left|\mu - \nu_n\right| > \varepsilon\right\} \le 2\exp\left(-2n\varepsilon^2\right)$$

  - 若把横轴划分为|A|份，则要考虑|A|个bin:

$$\Pr\left\{\sup_{a\in A}\left|\hat{p}(a) - p(a)\right| > \varepsilon\right\} \le 2|A|\exp\left(-2n\varepsilon^2\right)$$

$$\Rightarrow \quad \varepsilon \le \sqrt{\frac{\log 2|A| - \log\delta}{2n}}, \quad \delta = 2|A|\exp\left(-2n\varepsilon^2\right)$$

Q / A

- Any Question? …