



# 机器学习

Machine Learning



主讲人：张敏 清华大学长聘副教授



机器学习

MACHINE LEARNING-MIN ZHANG

Unit.06

# 基于实例的学习方法

\*图片均来自网络或已发表刊物

# 动机

- 之前「三步走」的学习方法
  - 估计问题特性（如分布）
  - 作出模型假设
    - LSE, Decision Tree, MAP, MLE, Naïve Bayes, ...
  - 找到最优的参数



有没有一种学习方法不遵循「模型假设 + 参数估计」的思路？

# 动机

- 人们通过记忆和行动来推理学习
- 思考即回忆、进行类比 Thinking is reminding, making analogies
- One takes the behavior of one's company 「近朱者赤，近墨者黑」



# 动机

「找到和这张图片最相似的 10 张图片」



「找到两个基因组之间所有匹配的基因片段」



# Topic 6.1 基于实例的学习 (I)

## 一、基本概念

# 一些名词概念

- 参数化(Parametric) vs. 非参数化 (Non-parametric)
  - 参数化：
    - 设定一个特定的函数形式
    - 优点：**简单**，容易估计和解释
    - **可能存在很大的偏置**：实际的数据分布可能不遵循假设的分布
  - 非参数化：
    - 分布或密度的估计是数据驱动的 (**data-driven**)
    - 需要事先对函数形式作的估计相对更少

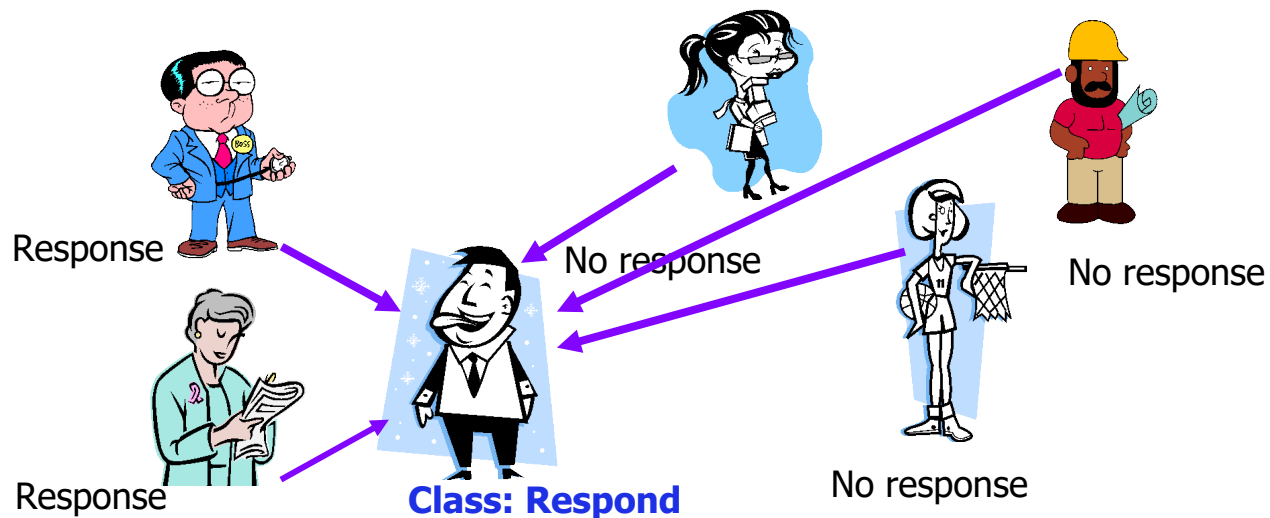
# 其他名词

- **Instance-Based Learning (IBL)** : 基于**实例**的学习  
or **Instance Based Methods (IBM)** : 基于**实例**的方法
- **Memory-Based Learning** : 基于**记忆**的学习
- **Case-Based Learning** : 基于**样例**的学习
- **Similarity-Based Learning** : 基于**相似度**的学习
  
- **Case-Based Reasoning** : 基于样例的推理
- **Memory-Based Reasoning** : 基于记忆的推理
- **Similarity-Based Reasoning** : 基于相似度的推理



# 基于实例的学习

- 无需构建模型 —— 仅存储所有训练样例
- 直到有新样例需要分类才开始进行处理

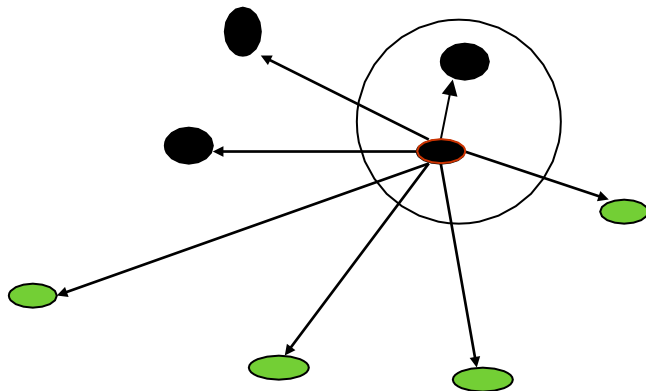


# 基于实例的概念表示

- 一个概念  $c_i$  可以表示为:
  - 样例的集合  $c_i = \{e_{i1}, e_{i2}, \dots\}$ ,
  - 一个相似度估计函数  $f$ , 以及
  - 一个阈值  $\theta$
- 一个实例 'a' 属于概念  $c_i$ , 当
  - 'a' 和  $c_i$  中的某些  $e_j$  相似, 并且
  - $f(e_j, a) > \theta$

# 1. 最近邻

- 相似度  $\leftrightarrow$  距离



# 最近邻的例子

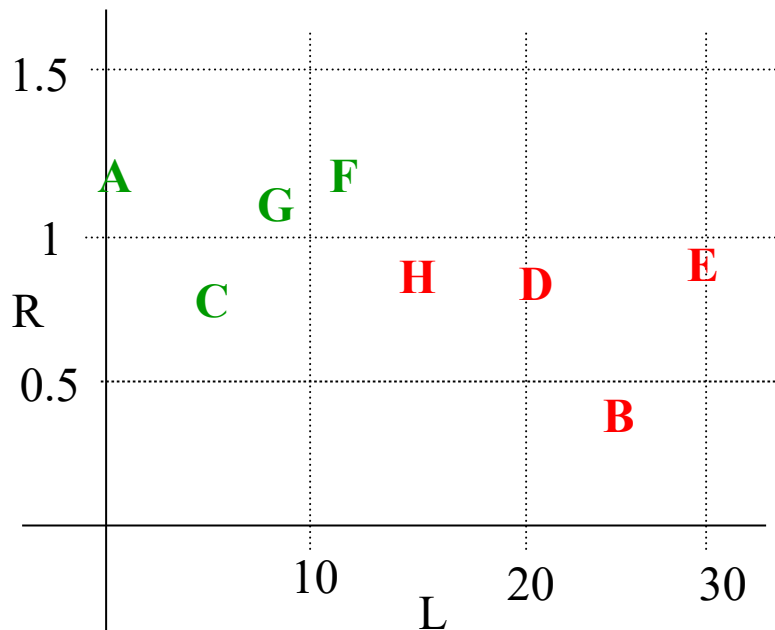
信用评分

分类：好 / 坏

特征：

- L = #延迟还款/年
- R = 收入/花销

name	L	R	G/P
A	0	1.2	G
B	25	0.4	P
C	5	0.7	G
D	20	0.8	P
E	30	0.85	P
F	11	1.2	G
G	7	1.15	G
H	15	0.8	P



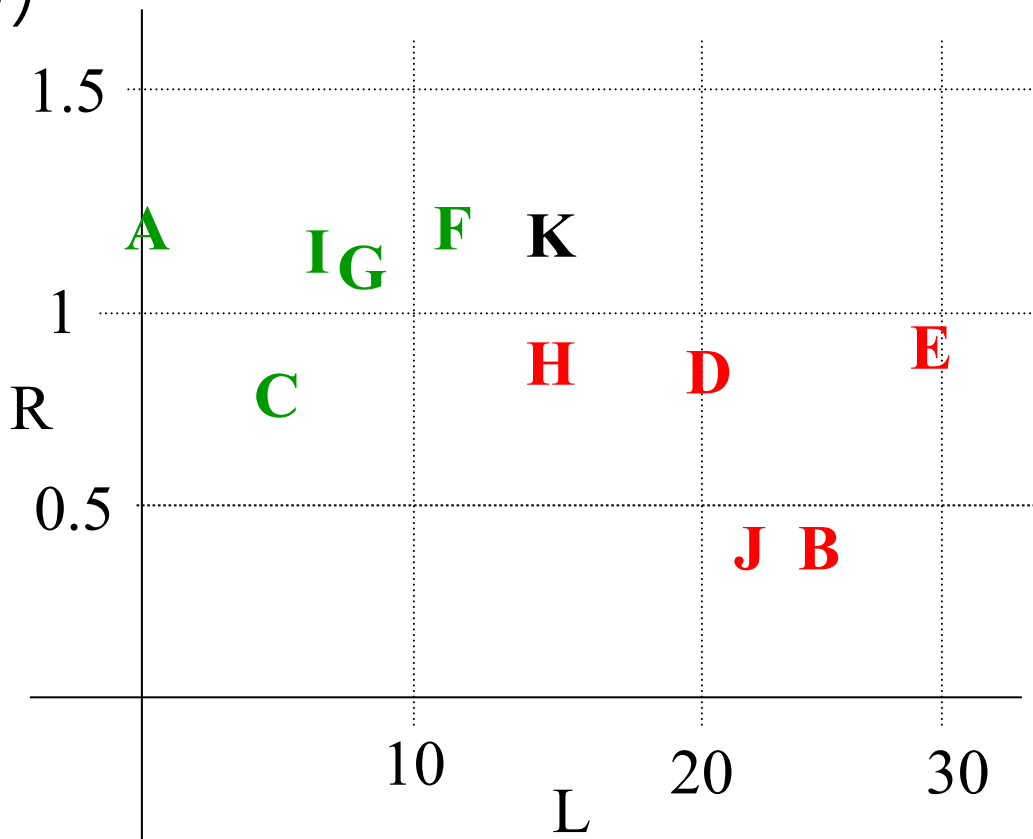
## 最近邻示例 (续)

name	L	R	G/P
I	6	1.15	?
J	22	0.45	?
K	15	1.2	?

距离度量：

- 缩放距离

$$\sqrt{(L_1 - L_2)^2 + (10R_1 - 10R_2)^2}$$



# 理论结果

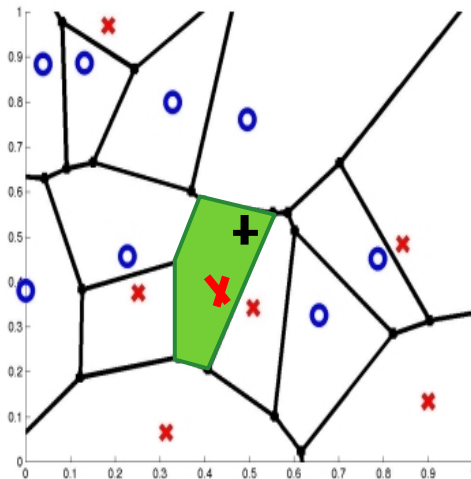
- 无限多训练样本下 1-NN 的错误率界限:

$$\text{Err}(\text{Bayes}) \leq \text{Err}(1\text{-NN}) \leq \text{Err}(\text{Bayes}) \left( 2 - \frac{K}{K-1} \text{Err}(\text{Bayes}) \right)$$

- 证明很长 (参照 Duda et al, 2000)
- 因此 1-NN 的错误率不大于 Bayes 方法错误率的 2 倍

# 最近邻 (1-NN) : 解释

- Voronoi Diagram
- Voronoi tessellation
  - 也称为 **Dirichlet tessellation**
- Voronoi decomposition
- 对于任意欧氏空间的离散点集合 $S$ ,以及几乎所有的点 $x$ ,  $S$ 中一定有一个和 $x$ 最接近的点
  - 没有说“所有的点”是因为有些点可能和两个或多个点距离相等（在边界上）



# 问题

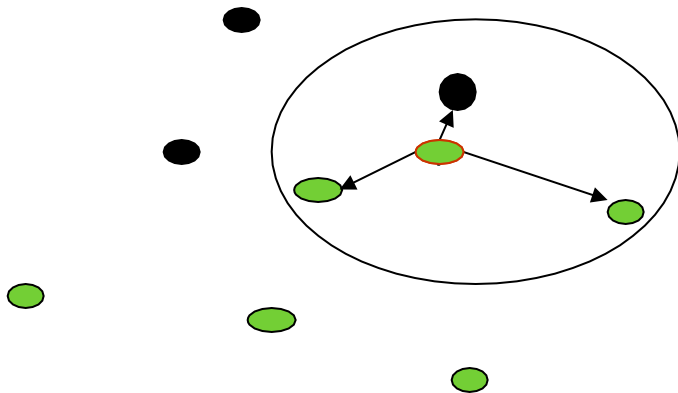
- 最近邻的点是噪音怎么办？

- 解决方法

- 用不止一个邻居
- 在邻居中进行投票









k-近邻 (KNN)





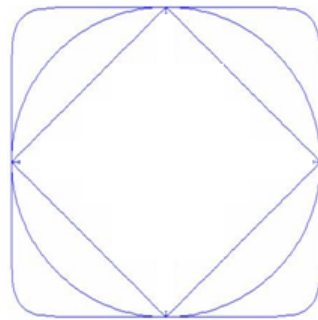
## 二、K-近邻 (KNN)

# KNN: 示例 (3-NN)

顾客	年龄	收入 (K)	卡片数	结果	距 David 的距离
John 	35	35	3	No	$\text{sqrt} [(35-37)^2+(35-50)^2+(3-2)^2]=15.16$
Mary 	22	50	2	Yes	$\text{sqrt} [(22-37)^2+(50-50)^2+(2-2)^2]=15$
Hannah 	63	200	1	No	$\text{sqrt} [(63-37)^2+(200-50)^2+(1-2)^2]=152.23$
Tom 	59	170	1	No	$\text{sqrt} [(59-37)^2+(170-50)^2+(1-2)^2]=122$
Nellie 	25	40	4	Yes	$\text{sqrt} [(25-37)^2+(40-50)^2+(4-2)^2]=15.74$
David 	37	50	2	Yes	

# KNN 讨论 1：距离度量

- Minkowski 或  $L_\lambda$  度量：
$$d(i, j) = \left( \sum_{k=1}^p |x_k(i) - x_k(j)|^\lambda \right)^{\frac{1}{\lambda}}$$



- 欧几里得距离 ( $\lambda = 2$ )

$$d_{ij} = \sqrt{\sum_{k=1}^p (x_{ik} - x_{jk})^2}$$

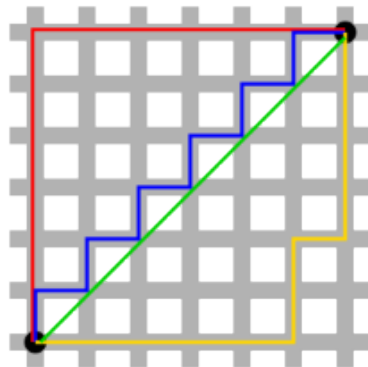
- 曼哈顿距离 Manhattan Distance

城市街区距离 City block Dis.

出租车距离 Taxi Distance

或  $L_1$  度量 ( $\lambda = 1$ ):

$$d(i, j) = \sum_{k=1}^p |x_k(i) - x_k(j)|$$



# KNN：距离度量

- 切比雪夫距离 (Chebyshev Distance)

棋盘距离 (Chessboard Dis.)

$L_\infty$

$$d(i, j) = \max_k |x_k(i) - x_k(j)|$$

- 加权欧氏距离

Mean Censored Euclidean

Weighted Euclidean Distance


$$\sqrt{\sum_k (x_{ik} - x_{jk})^2 / n}$$

- Bray-Curtis Dist.

$$\sum_k |x_{ik} - x_{jk}| / \sum_k (x_{ik} + x_{jk})$$

- 堪培拉距离 Canberra Dist.

$$\frac{\sum_k |x_{ik} - x_{jk}|}{k} / (x_{ik} + x_{jk})$$

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1		1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

# KNN 总览

- 基本算法
- 讨论
  - 更多距离度量
  - 属性

## KNN 讨论 2：属性



**John:**

Age=35

Income=95K

No. of credit cards=3



**Rachel:**







Age=41

Income=215K

No. of credit cards=2

- 邻居间的距离可能被某些取值特别大的属性所支配
  - e.g. 收入  $\text{Dis}(\text{John}, \text{Rachel}) = \sqrt{(35-41)^2 + (95,000-215,000)^2 + (3-2)^2}$
- 对特征进行归一化是非常重要的 (e.g., 把数值归一化到  $[0-1]$  )
  - Log, Min-Max, Sum, ...

# KNN: 属性归一化

Customer	Age	Income (K)	#cards	Response
John 	$55/63 = 0.55$	$35/200 = 0.175$	$3/4 = 0.75$	No
Rachel 	$22/63 = 0.34$	$50/200 = 0.25$	$2/4 = 0.5$	Yes
Hannah 	$63/63 = 1$	$200/200 = 1$	$1/4 = 0.25$	No
Tom 	$59/63 = 0.93$	$170/200 = 0.85$	$1/4 = 0.25$	No
Nellie 	$25/63 = 0.39$	$40/200 = 0.2$	$4/4 = 1$	Yes
David 	$37/63 = 0.58$	$50/200 = 0.25$	$2/4 = 0.5$	Yes

# KNN: 属性加权

- 一个样例的分类是基于所有属性的

➤ 与属性的相关性无关 —— 无关的属性也会被使用进来

- 根据每个属性的相关性进行加权 e.g.  $d_{WE}(i, j) = \left( \sum_{k=1}^p w_k (x_k(i) - x_k(j))^2 \right)^{\frac{1}{2}}$
- 在距离空间对维度进行缩放

➤  $w_k = 0 \rightarrow$  消除对应维度 (特征选择)

- 一个可能的加权方法：

使用 互信息  $I(\text{属性}, \text{类别})$

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \quad H: \text{熵 (entropy)}$$

$$H(X, Y) = - \sum p(x, y) \log p(x, y) \quad \text{联合熵 (Joint entropy)}$$



# KNN 总览

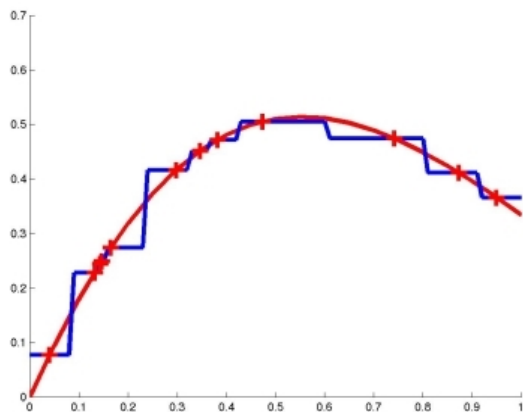
- 基本算法
- 讨论
  - 更多距离度量
  - 属性
    - 归一化、加权
  - 连续取值目标函数

## KNN 讨论 3: 连续取值目标函数

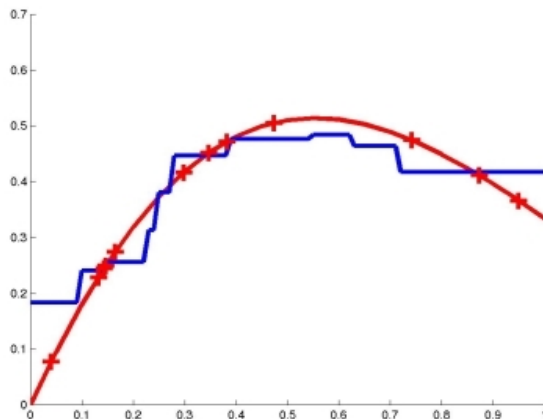
- 离散输出 - 投票
- 连续取值目标函数
  - $k$  个近邻训练样例的均值

# 连续取值目标函数

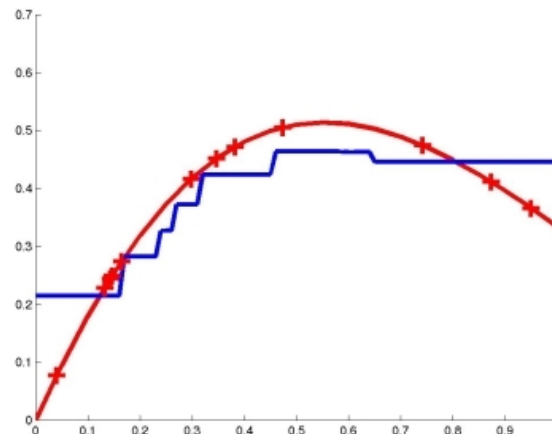
红色：实例的真实值      蓝色：估计值



1-nearest neighbor



3-nearest neighbor



5-nearest neighbor

# KNN 总览

- 基本算法
- 讨论
  - 更多距离度量
  - 属性
    - 归一化、加权
  - 连续取值目标函数
  - $k$  的选择

## KNN 讨论 4 : k 的选择

- 多数情况下  $k=3$
- 取决于训练样例的数目
  - 更大的  $k$  不一定带来更好的效果
- 交叉验证
  - Leave-one-out (Throw-one-out, Hold-one-out)
    - 每次：拿一个样例作为测试，所有其他的作为训练样例
- KNN 是稳定的
  - 样例中小的混乱不会对结果有非常大的影响

# KNN 总览

- 基本算法
- 讨论
  - 更多距离度量
  - 属性
    - 归一化、加权
  - 连续取值目标函数
  - $k$  的选择
  - 打破平局 (break ties)

## KNN 讨论 5：打破平局

- 如果  $k=3$  并且每个近邻都属于不同的类？
  - $P(w|X)=1/3$
  - 或者找一个新的邻居 (4<sup>th</sup>)
  - 或者取最近的邻居所属类
  - 或者随机选一个
  - 或者 ...



「之后会讨论一个更好的解决方案」

# KNN 总览

- 基本算法
- 讨论
  - 更多距离度量
  - 属性：归一化、加权
  - 连续取值目标函数
  - $k$  的选择
  - 打破平局
  - 关于效率（待续）