

---

# 第三章：特征



# 引言

---

## The ingredients of machine learning

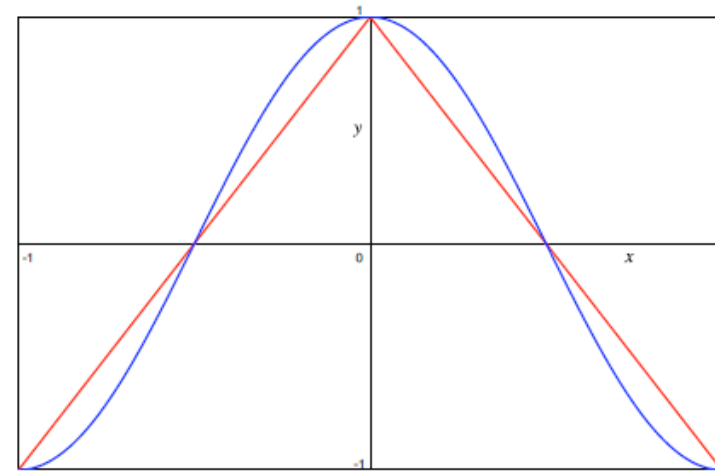
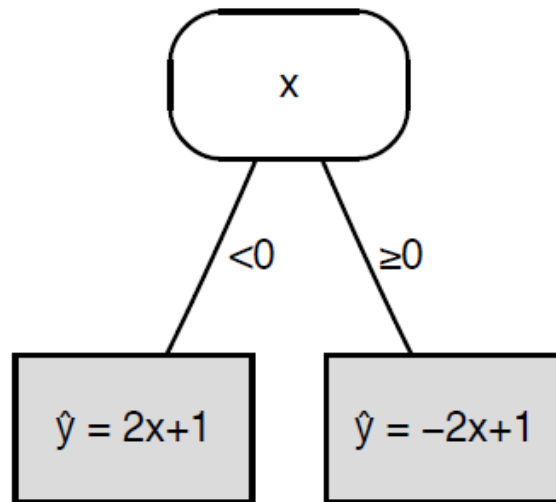
- Tasks: the problems that can be solved with machine learning
  - Looking for structure
- Models: the output of machine learning
  - Geometric models
  - Probabilistic models
  - Logical models
  - Grouping and grading
- Features: the workhorses of machine learning
  - Two uses of features
  - Feature construction and transformation

# 引言

---

- ❑ Feature: the workhorse of machine learning
- ❑ Features and models are intimately connected
  - Models are defined in terms of features
  - A single feature can be turned into a univariate model
  - Two uses of features
    - Features as splits
    - Features as predictors

# 引言



**(left)** A regression tree combining a one-split feature tree with linear regression models in the leaves. Notice how  $x$  is used as both a splitting feature and a regression variable.

**(right)** The function  $y = \cos \pi x$  on the interval  $-1 \leq x \leq 1$ , and the piecewise linear approximation achieved by the regression tree.

A small regression tree

# 引言

---

## □ 特征(属性): 描述事物或对象在某方面的表现或性质

- 我们可以根据特征域对不同特征进行区分: 常见的特征域包括实数和整数以及离散集 (如颜色、布尔型等)
- 我们也可以根据在特征上执行的有意义运算来区分特征。例如, 我们可以计算一个人群的平均年龄, 但无法计算他们的平均血型。因此, 求平均这个运算只适用于某些特征
- 虽然许多数据集都带有预先定义的特征, 但依然可以进行多种方式处理特征。例如, 我们可以通过变更尺度或离散化来改变某一特征的特征域; 也可以从一个规模较大的特征集中选择一些最优的特征, 且只使用这些特征; 还可以将两种或多种特征整合为一种新特征

# 大纲

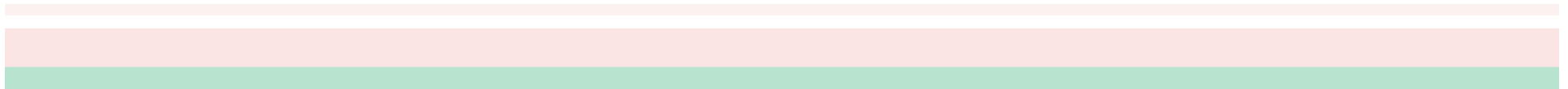
---

□ 特征类型

□ 特征变换

□ 特征构造

□ 特征选择



# 特征统计量

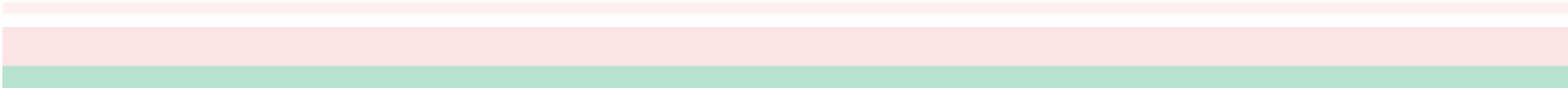
---

## □ 特征统计量: 特征上可以执行的运算

- 集中趋势统计量
- 离差统计量
- 形状统计量

每个类别都可以解释为某个未知总体的理论性质或某个给定抽样的具体性质。这里我们只关注样本统计量

## □ 集中趋势统计量

- 均值: 平均值
  - 中位数: 将样本按照特征值从低到高排序后, 排列在正中的特征值
  - 众数: 集合中出现频率最高的那个 (或那些) 特征值
- 

# 离差统计量

---

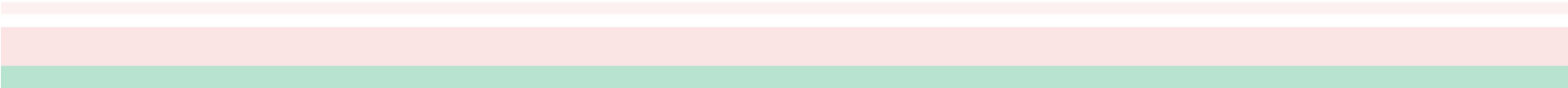
## □ 方差和标准差

方差是算术均值的偏差的方均值。标准差是方差的平方根，与特征具有相同的尺度

## □ 极差 (range) 和中列点 (midrange point)

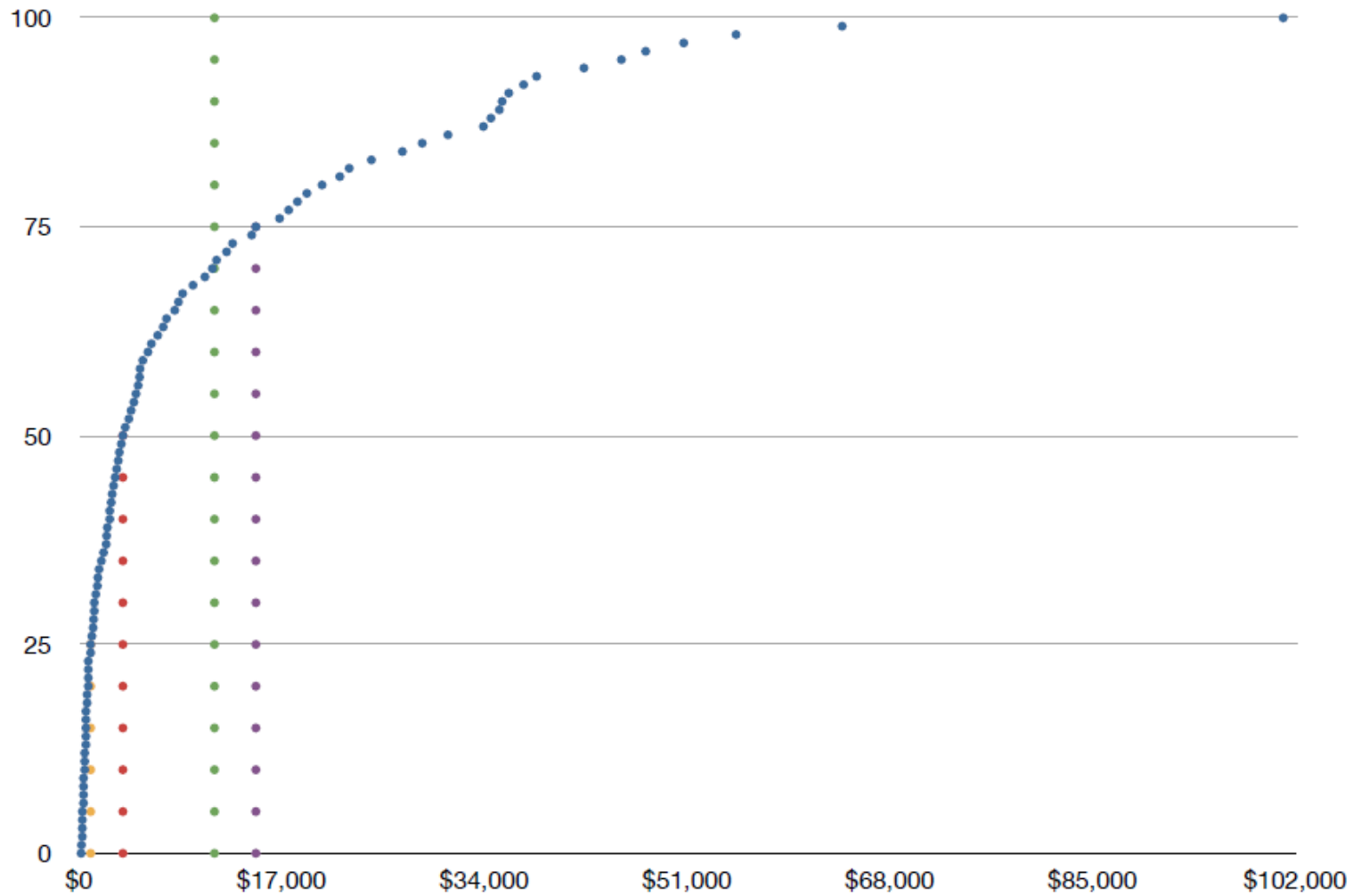
- 极差：最大值与最小值之差
- 中列点：最大值与最小值的均值

## □ 百分位数 (percentile)

- 百分位数：第 $p$ 个百分位数指有 $p\%$ 的样本均落在其后的那个值
  - 四分位数、十分位数
  - 四分位距：数据分布中第一个和第三个四分位数之间的距离
- 

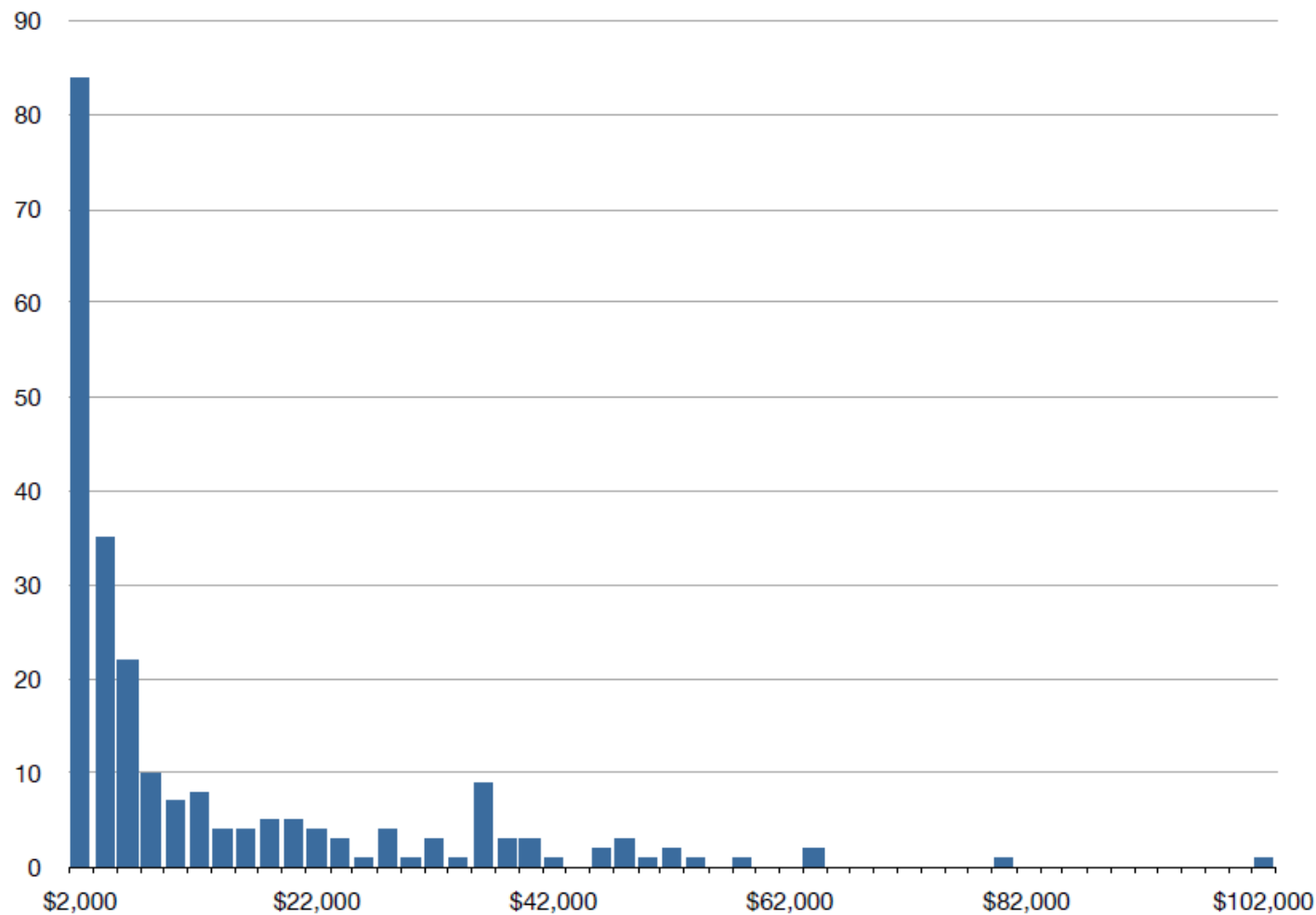


# 百分位图



231个国家的人均GDP百分位图

# 直方图



231个国家的人均GDP直方图

# 形状统计量

---

## □ 分布的偏斜程度和尖的程度

可以通过偏度 (skewness) 和峰度 (kurtosis) 等形状统计量刻画, 主要是计算样本的三阶和四阶中心矩

## □ $k$ 阶中心矩: $m_k = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^k$

- 一阶中心矩
- 二阶中心矩
- 三阶中心矩 $m_3$ : 可正可负

# 形状统计量

---

□ 偏度 (skewness) :  $m_3/\sigma^3$

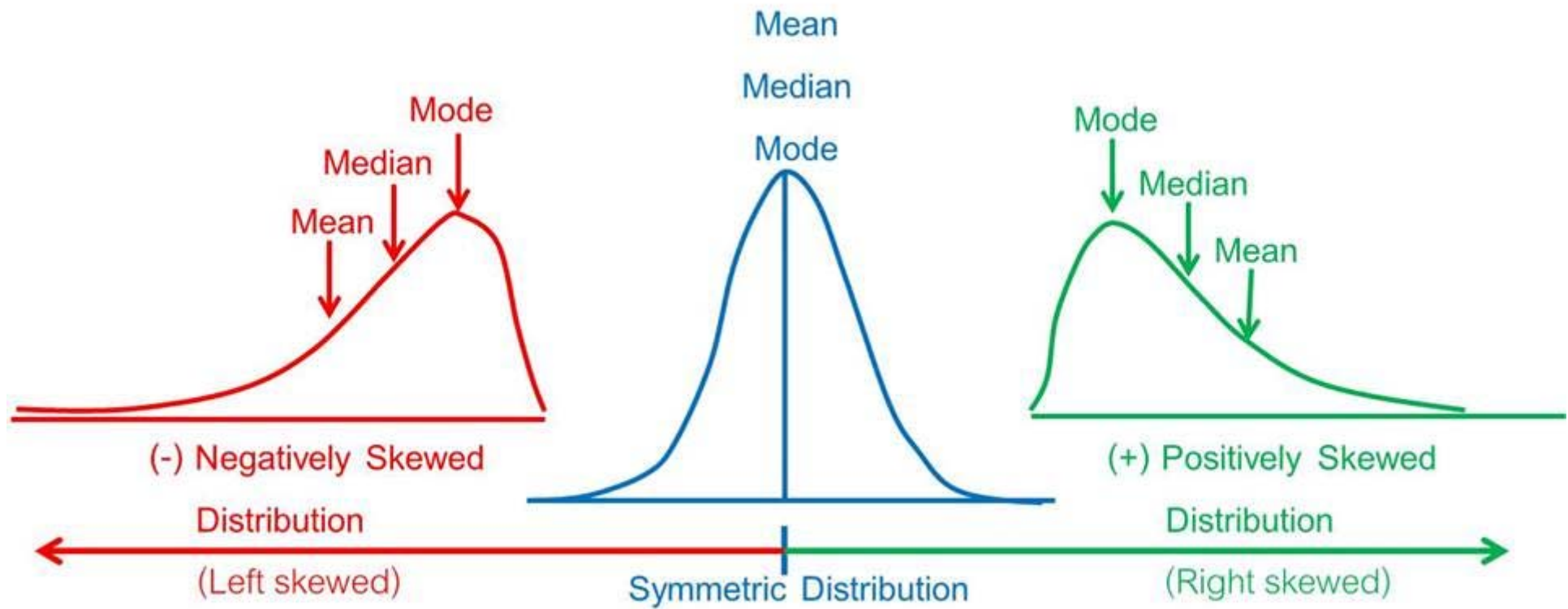
偏度为正，表示分布向右偏斜，即右侧尾部要长于左侧。反之，则分布向左偏斜

□ 峰度 (kurtosis) :  $m_4/\sigma^4$

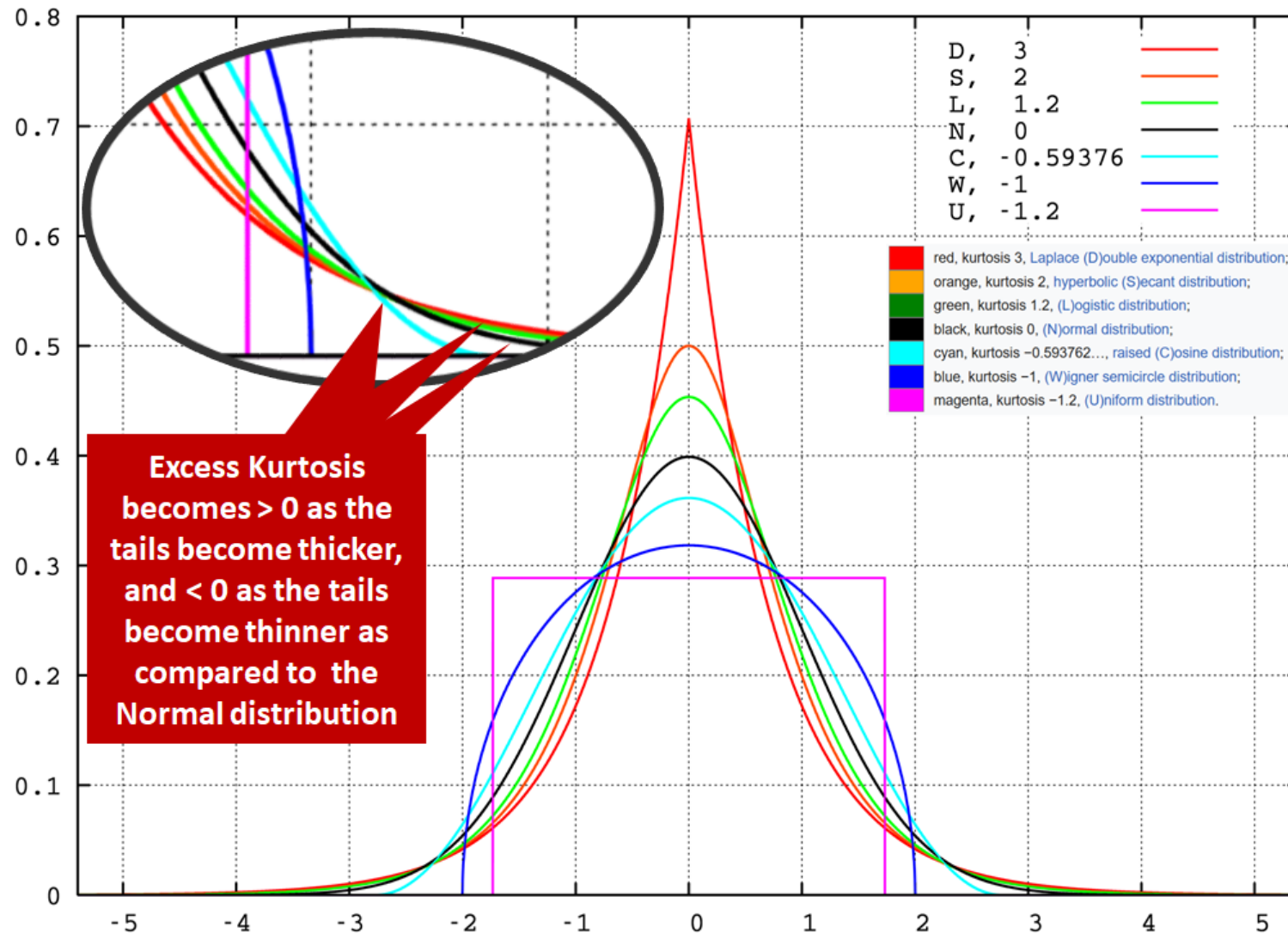
- 正态分布峰度为3
- 超峰度 (excess kurtosis)

□ 人均CDP例子中，偏度为2.12，超峰度为2.53

# 形状统计量



# 形状统计量



# 特征（作为分类用）的要求

---

## □ 很大的识别信息量

即所提供的特征应具有很好的可分性，使得学习器容易判别

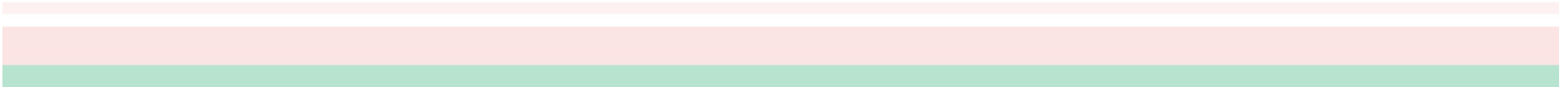
## □ 可靠性

对那些模棱两可，似是而非不易判别的特征应该去掉

## □ 尽可能的独立性

不含重复的特征，相关性强的特征只选一个，因为强相关性并没有增加更多的分类信息

## □ 数量尽可能少，同时损失的信息尽量少



# 特征形成

---

- ❑ 在设计一个具体的机器学习系统时，往往先接触一些训练样本，有领域专家和工程师联合研究对象所包含的特征信息，并给出相应的表述方法。这一阶段目标往往是获取尽可能多的表述特征
- ❑ 在这些特征中，有些可能满足前述要求，有的则可能不满足，不能作为分类的依据
- ❑ 根据样本分析得到一组表述观察对象的特征值，而不论特征是否实用，称这一步为特征形成，得到的特征称为原始特征
- ❑ 特征形成是机器学习与模式识别过程的重点与难点之一



# 特征的类别

---

## □ 数量特征 (quantitative feature)

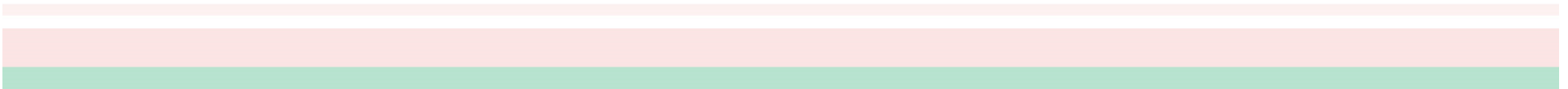
特征域上有无穷个可能的取值，涉及到实数的映射，也称为连续特征 (continuous feature) 或数值特征 (numerical feature)

## □ 有序特征 (ordinal feature)

具有位序但无数值尺度的特征

## □ 属性特征 (categorical feature)

不含位序和无数值尺度的特征，也称为列名特征 (nominal feature)

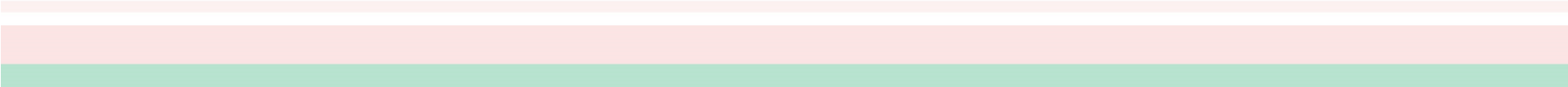


# 特征的类别

---

特征类型、属性以及可计算的统计量

类型	有序	有尺度	集中趋势	离差	形状
属性特征					
有序特征					
数量特征					



# 特征的类别

特征类型、属性以及可计算的统计量

类型	有序	有尺度	集中趋势	离差	形状
属性特征	否	否	众数	NA	NA
有序特征	是	否	中位数	分位数	NA
数量特征	是	是	均值	极差、四分位距、方差、标准差	偏度、峰度

每种特征类型都继承了该表中位于其上方的特征类型的统计量。  
例如，众数是一种适用于所有特征的集中趋势统计量

# 结构特征

---

结构特征的表达是先将观察对象分割成若干个基本构成要素，再确定基本要素之间的相互连接关系。例如人的指纹特征、人脸五官结构信息。指纹识别、汉字识别等领域的成功都离不开结构特征的选择。

- 可以较好地表达复杂的图像图形信息
- 结构信息对研究对象的尺寸一般不太敏感
- 属于比较容易感知的特征

# 结构特征

---

假设某封电子邮件被表示为一个单词序列。那么，除了常见的词频特征，还可以定义大量其他特征。包括：

- 短语 “machine learning” 是否在该邮件中出现；
- 该邮件是否包含至少8个连续的非英语单词；
- 该邮件是否为回文结构，例如 “Degas, are we not drawn onward, we freer few, drawn onward to new eras aged?”

进一步地，我们不必拘泥于单封邮件的属性，可以加入表达不同邮件之间的关系。如一封邮件是否引用了另一封邮件的内容，以及两封邮件正文是否存在一个或多个段落相同

# 大纲

---

- 特征类型
- 特征变换
- 特征构造
- 特征选择

# 特征变换

特征变换通过移除、修改或添加信息来提高特征的效用。

特征变换类型总结

↓目的, 来源→	数量特征	有序特征	属性特征	布尔特征
数量特征				
有序特征				
属性特征				
布尔特征				

# 特征变换

特征变换通过移除、修改或添加信息来提高特征的效用。

特征变换类型总结

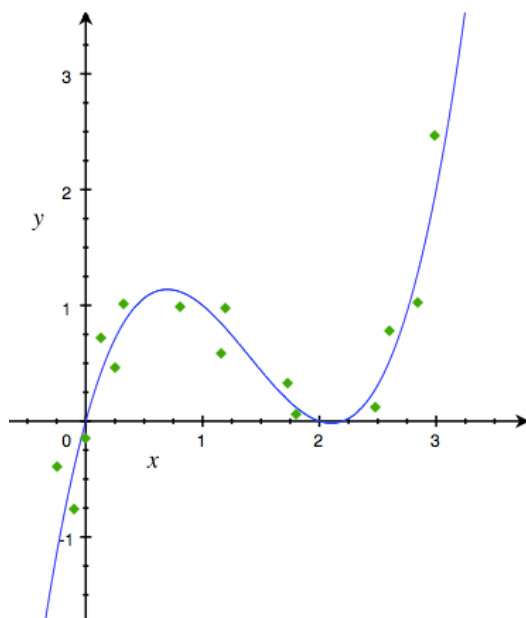
↓目的, 来源→	数量特征	有序特征	属性特征	布尔特征
数量特征	归一化	标定	标定	标定
有序特征	离散化	排序	排序	排序
属性特征	离散化	无序化	分组	
布尔特征	阈值化	阈值化	二值化	

归一化和标定会改变数量特征的尺度，或者赋予本来没有尺度的特征以某个尺度。排序会赋予特征某种次序或对原有次序进行修改。其他变换则通过抽象略去无关细节，以演绎方式（无序化、二值化）或通过引入新的信息（阈值化、离散化）



# 阈值化与离散化

- 阈值化：通过寻找用于划分的特征取值将数量特征或有序特征变换为布尔特征
- 离散化：多阈值情形。将原来的特征取值分段，转化为一个取值为1或0的向量。原始值落在某个段里，则向量中此段对应的元素为1，否则为0



# 阈值化与离散化

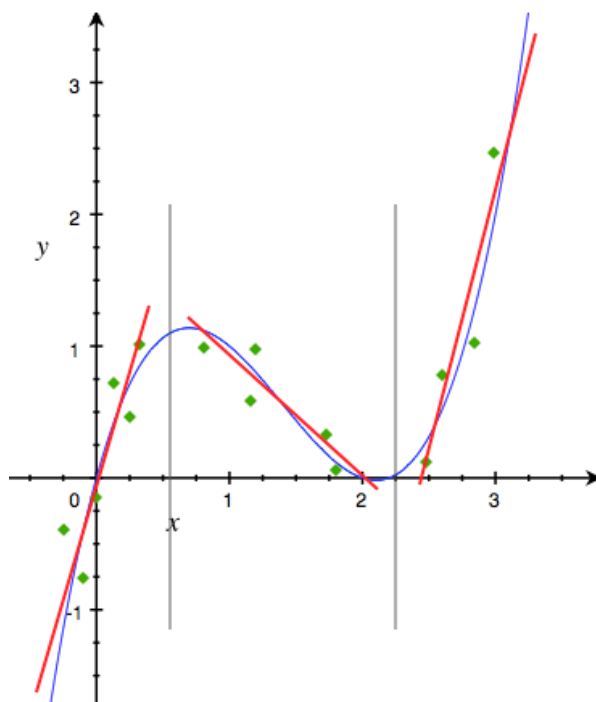
- 阈值化：通过寻找用于划分的特征取值将数量特征或有序特征变换为布尔特征
- 离散化：多阈值情形。将原来的特征取值分段，转化为一个取值为1或0的向量。原始值落在某个段里，则向量中此段对应的元素为1，否则为0

比如取离散点  $\{0.5, 1.5, 2.5\}$ ，通过判断原始数值属于  $(-\infty, 0.5)$ ， $[0.5, 1.5)$ ， $[1.5, 2.5)$ ， $[2.5, +\infty)$  中哪段 (bin) 来把原始数值离散化为4维的向量。下面是一些例子的离散结果：

原始数值	离散化后的值
0.1	(1,0,0,0)
1.3	(0,1,0,0)
3.2	(0,0,0,1)
5.8	(0,0,0,1)

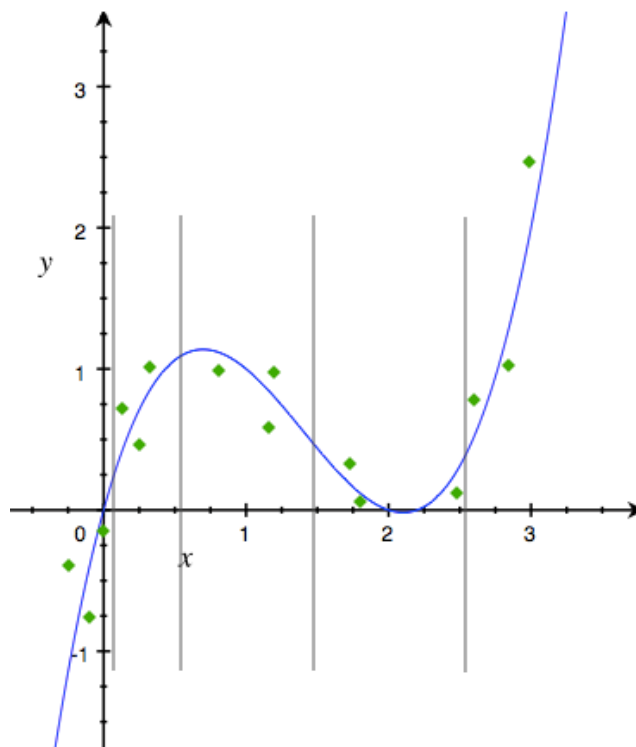
# 观测趋势离散化

- **观测趋势：**以 $x$ 为横坐标， $y$ 为纵坐标，画图，看曲线的趋势和拐点。通过观察下面的图我们发现可以利用3条直线（红色直线）来逐段近似原来的曲线。把离散点设为两条直线相交的各个点，我们就可以把 $x$ 离散化为长度为3的向量。



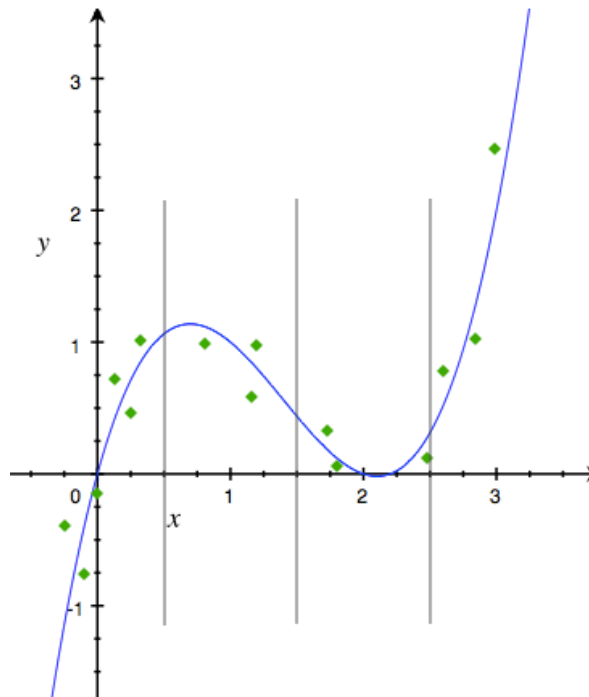
# 等频率离散化

- 等频率离散化 (equal-frequency discretization) : 选取的离散点保证落在每段里的样本数量大致相同。

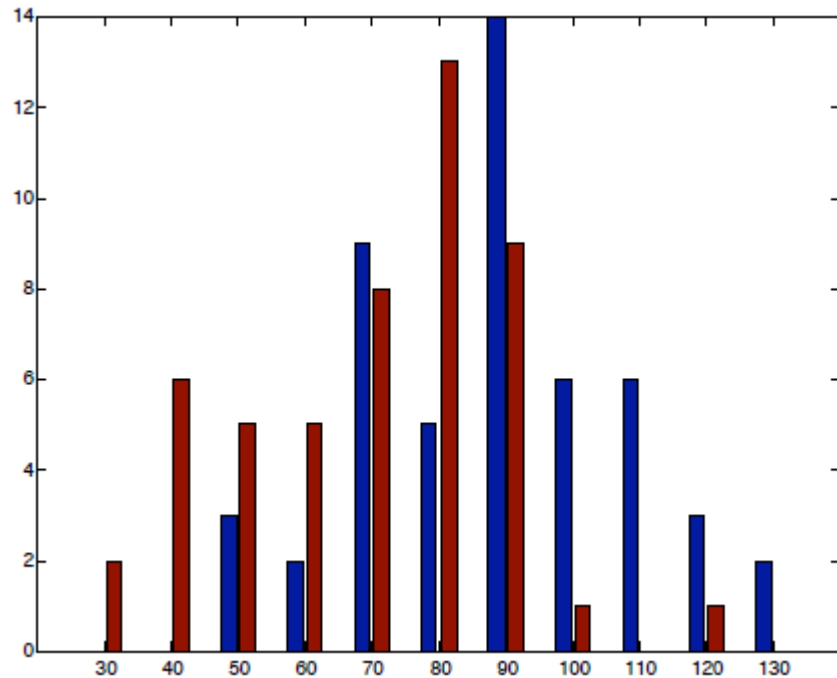


# 等宽离散化

- 等宽离散化 (equal-width discretization) : 离散点选取等距点。例如上述对原始数据取离散点  $\{0.5, 1.5, 2.5\}$  就是一种等距离散, 见下图。图中垂直的灰线代表离散点。

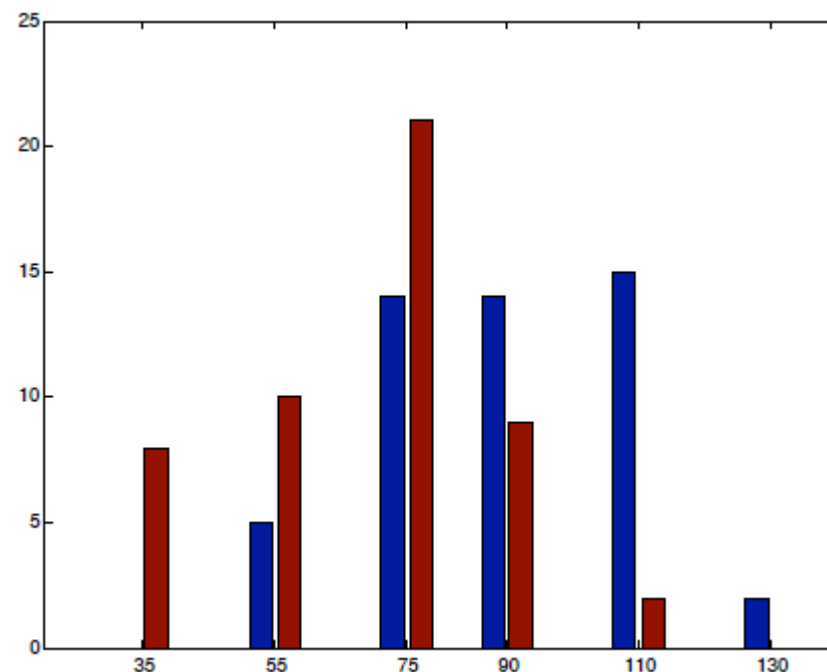
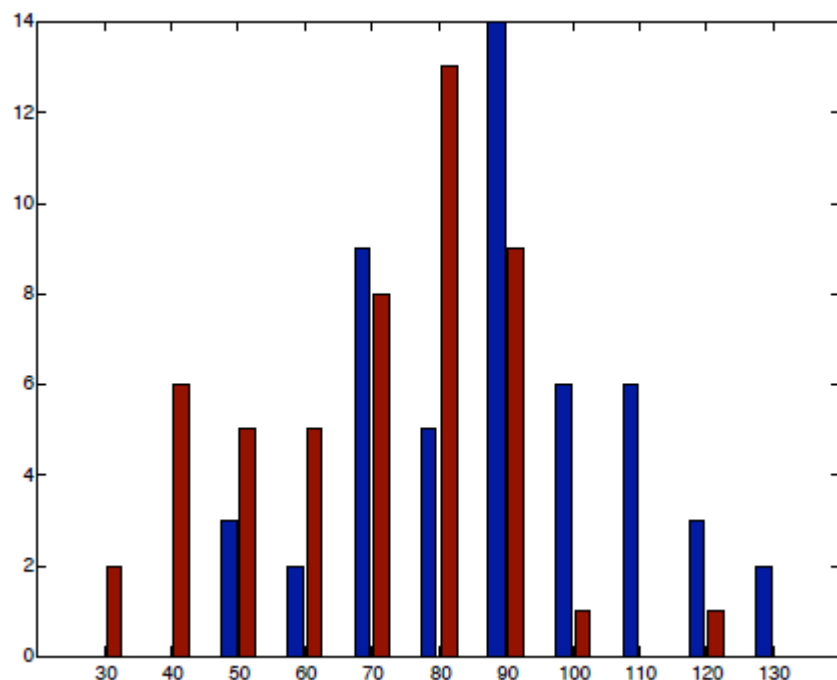


# 类别敏感离散化



(left) Artificial data depicting a histogram of body weight measurements of people with (blue) and without (red) diabetes, with eleven fixed intervals of 10 kilograms width each.

# 类别敏感离散化



**(left)** Artificial data depicting a histogram of body weight measurements of people with (blue) and without (red) diabetes, with eleven fixed intervals of 10 kilograms width each.

**(right)** By joining the first and second, third and fourth, fifth and sixth, and the eighth, ninth and tenth intervals, we obtain a discretisation such that the proportion of diabetes cases increases from left to right. This discretisation makes the feature more useful in predicting diabetes.

# 归一化

---

- 归一化 (normalization)：将数据按比例缩放，使之落入一个较小的特定区间。也称为规范化
- 可以提升模型收敛速度和准确率
- 不是所有的机器学习模型都需要数据归一化处理，例如决策树
- 常用的归一化方法：标准化 (standardization)、平均归一化和min-max归一化等



# 归一化

---

- 标准化 (standardization) : 也称为z-score归一化

$$x' = \frac{x - \bar{x}}{\sigma}$$

- 平均归一化 (mean normalization)

$$x' = \frac{x - \text{mean}(x)}{\text{max}(x) - \text{min}(x)}$$

- min-max归一化 (min-max normalization)

$$x' = \frac{x - \text{min}(x)}{\text{max}(x) - \text{min}(x)}$$

- 单位长度缩放 (scaling to unit length)

$$x' = \frac{x}{\|x\|}$$

- 非线性归一化: 例如  $y = \log_{10}(x)$ ,  $y = \text{atan}(x) * 2 / \pi$

# 特征缺失 (incomplete feature)

---

- ❑ 删除：如果某个特征或样本中占比过大，可以直接删除
- ❑ 替代：用平均值、中值、分位数、众数、随机值等替代
- ❑ 预测：用其他变量做预测模型来算出缺失变量
- ❑ 映射到高维空间：比如性别，有男、女、缺失三种情况，则映射成3个变量：是否男、是否女、是否缺失。完整保留了原始数据的全部信息、不用考虑缺失值、不用考虑线性不可分之类的问题。缺点是计算量大大提升。而且只有在样本量非常大的时候效果才好

# 特征缺失 (incomplete feature)

客户对书籍的喜好程度的评分

	《笑傲江湖》	《万历十五年》	《人间词话》	《云海玉弓缘》	《人类的故事》
赵大	5	?	?	3	2
钱二	?	5	3	?	5
孙三	5	3	?	?	?
李四	3	?	5	4	?

能否将表中已经通过读者评价得到的数据当作**部分信号**，基于压缩感知的思想**恢复完整信号**从而进行书籍推荐呢？从题材、作者、装帧等角度看（相似题材的书籍有相似的读者），表中反映的信号是**稀疏**的，能通过类似压缩感知的思想加以处理。

**“矩阵补全” 技术解决此类问题**

# 大纲

---

- 特征类型
- 特征变换
- 特征构造
- 特征选择

# 特征构造

---

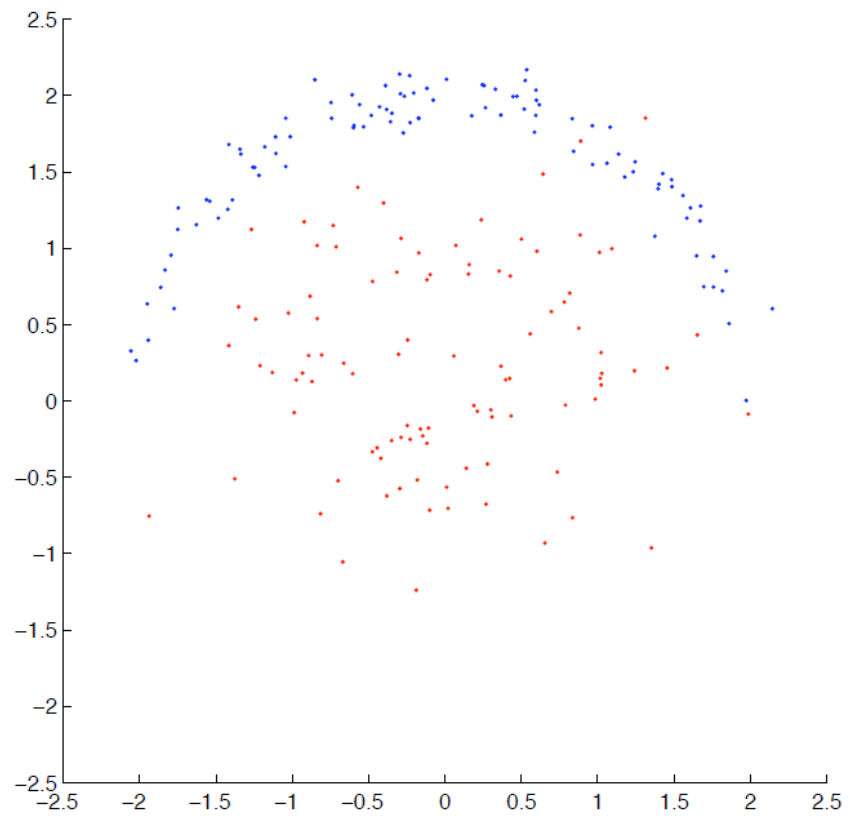
- 对数量特征取算术组合或多项式组合
- 基于笛卡尔乘积构造新特征

可以使用笛卡尔乘积的方式来组合2个或更多个特征。例如有两个特征color和shape，他们的取值分别为red、green、blue和circle、triangle、square。对它们做笛卡尔乘积转化，就可以组合出长度为9的特征。

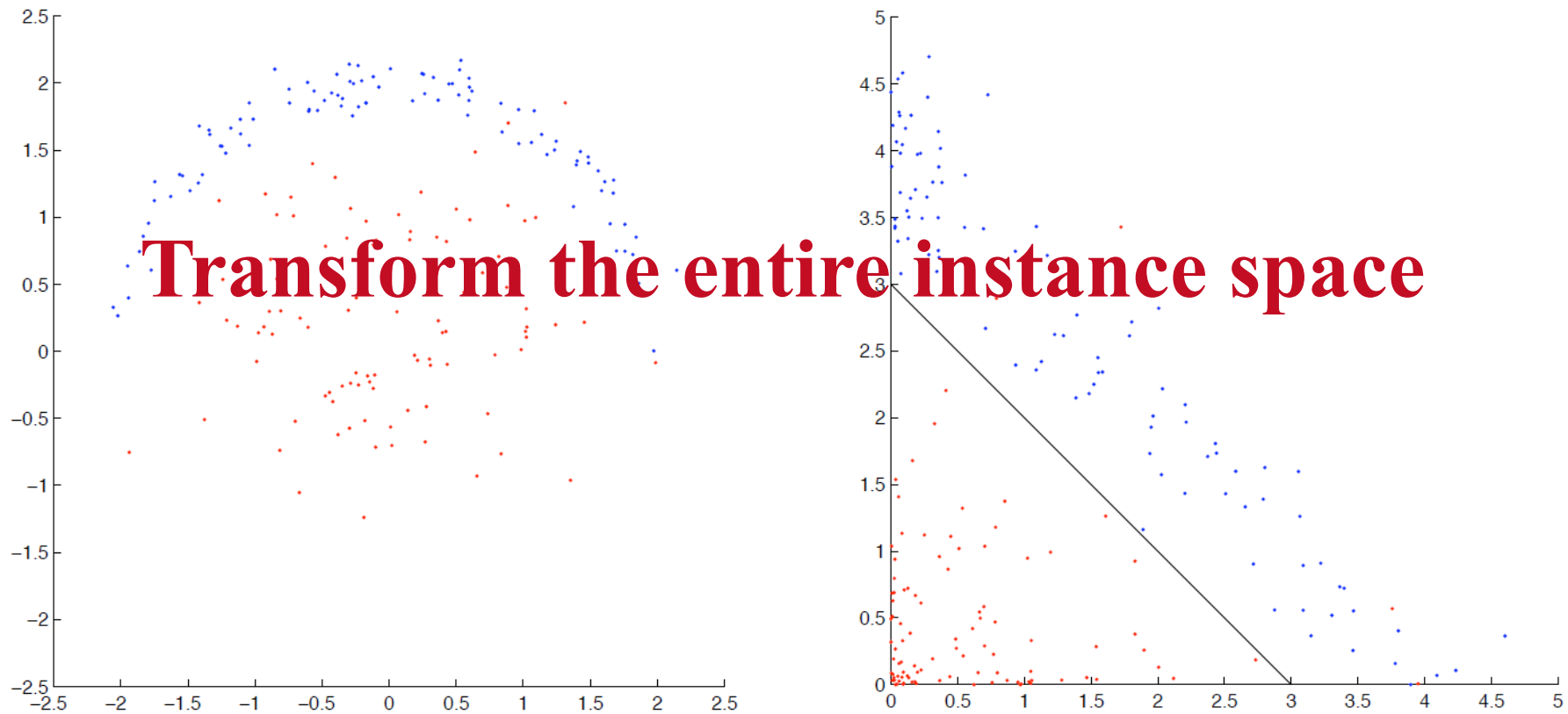
特征取值	circle	triangle	square
red			
green			
blue			

# 特征构造

---



# 特征构造



**(left)** A linear classifier would perform poorly on this data. **(right)** By transforming the original  $(x, y)$  data into  $(x', y') = (x^2, y^2)$ , the data becomes more 'linear', and a linear decision boundary  $x' + y' = 3$  separates the data fairly well. In the original space this corresponds to a circle with radius  $\sqrt{3}$  around the origin.

# 特征构造

---

Let  $\mathbf{x}_1 = (x_1, y_1)$  and  $\mathbf{x}_2 = (x_2, y_2)$  be two data points, and consider the mapping  $(x, y) \mapsto (x^2, y^2, \sqrt{2}xy)$  to a three-dimensional feature space. The points in feature space corresponding to  $\mathbf{x}_1$  and  $\mathbf{x}_2$  are  $\mathbf{x}'_1 = (x_1^2, y_1^2, \sqrt{2}x_1y_1)$  and  $\mathbf{x}'_2 = (x_2^2, y_2^2, \sqrt{2}x_2y_2)$ . The dot product of these two feature vectors is

$$\mathbf{x}'_1 \cdot \mathbf{x}'_2 = x_1^2 x_2^2 + y_1^2 y_2^2 + 2x_1 y_1 x_2 y_2 = (x_1 x_2 + y_1 y_2)^2 = (\mathbf{x}_1 \cdot \mathbf{x}_2)^2$$

That is, by squaring the dot product in the original space we obtain the dot product in the new space *without actually constructing the feature vectors*! A function that calculates the dot product in feature space directly from the vectors in the original space is called a *kernel* – here the kernel is  $\kappa(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1 \cdot \mathbf{x}_2)^2$ .

**Build the feature space classifier without constructing the feature space**

---

---

---



# 大纲

---

- 特征类型
- 特征变换
- 特征构造
- 特征选择

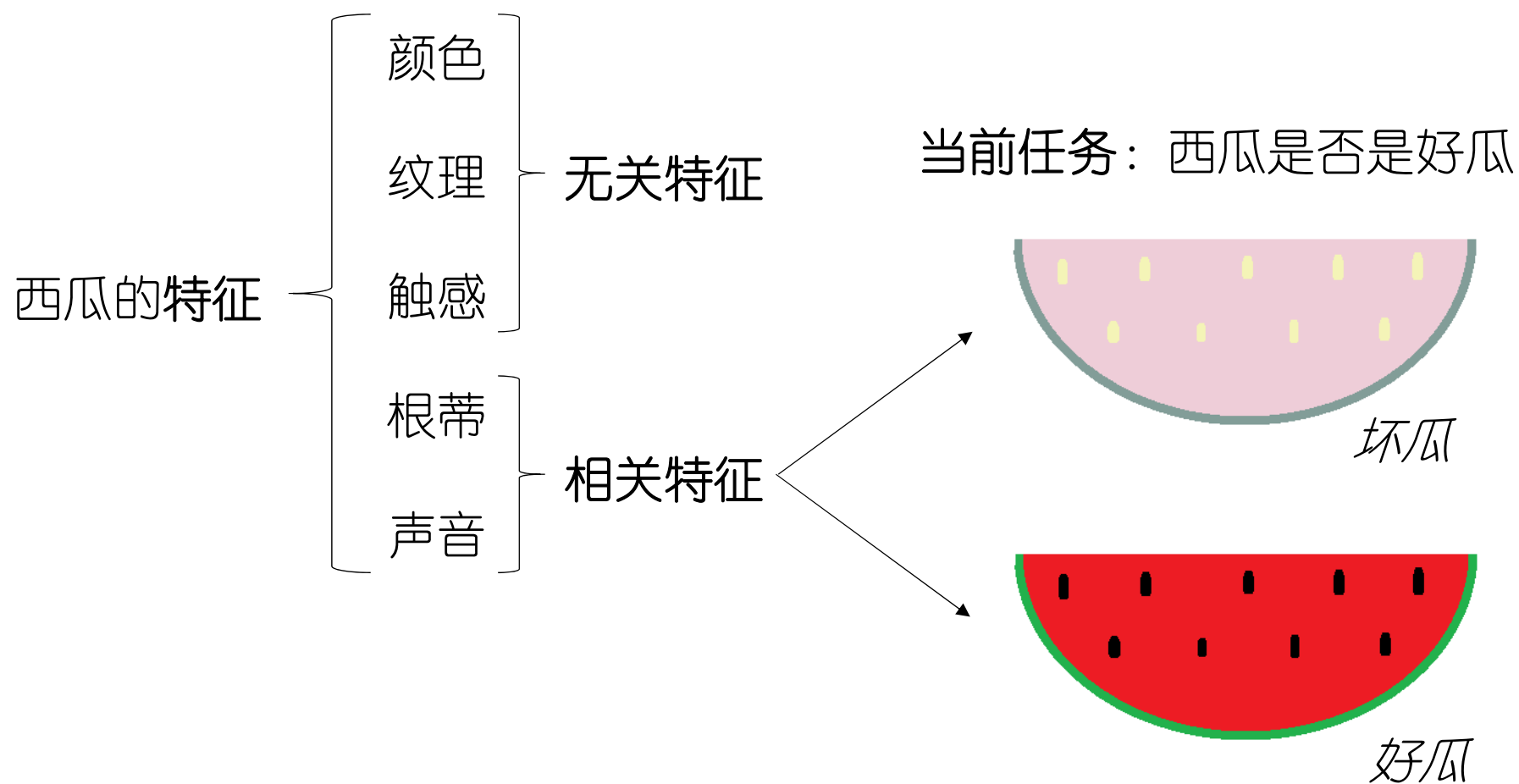
# 相关特征和无关特征

---

- 对于一个学习任务来说，给定特征集，其中某些特征可能很关键，另一些特征可能没有什么用
- 特征的分类：根据与学习任务的关系
  - 相关特征：对当前学习任务有用的属性
  - 无关特征：与当前学习任务无关的属性
  - 冗余特征\*：其所包含信息能由其他特征推演出来

\*为简化讨论，本节内容不涉及冗余特征

# 例子：西瓜的特征



# 特征选择

---

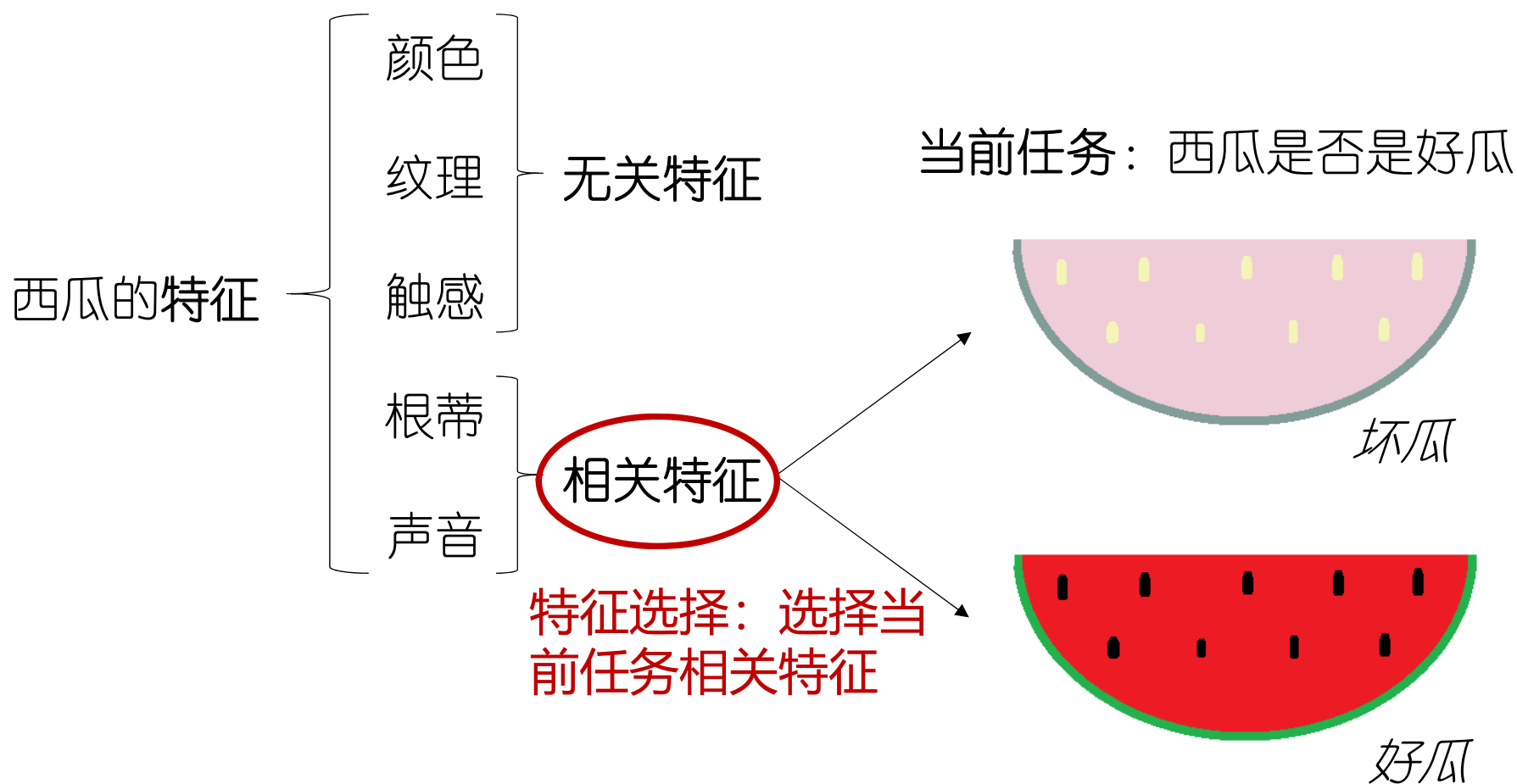
## □ 特征选择

- 从给定的特征集合中选出任务相关特征子集
- 必须确保不丢失重要特征

## □ 原因

- 减轻维度灾难：在少量属性上构建模型
- 降低学习难度：留下关键信息

# 例子：判断是否好瓜时的特征选择



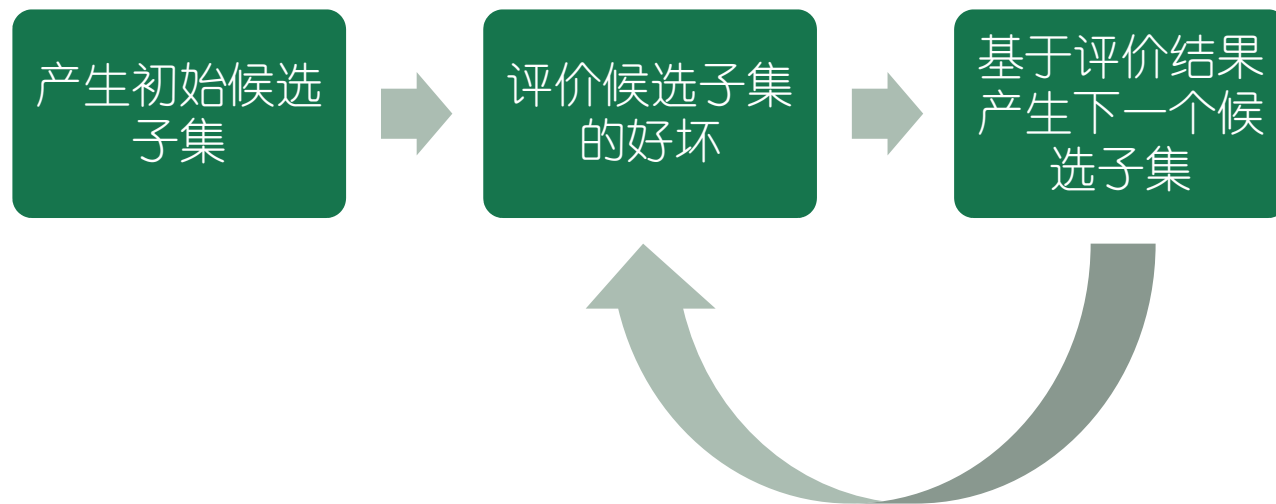
# 特征选择的一般方法

---

❑ 遍历所有可能的子集（穷尽搜索）

- 计算上遭遇组合爆炸，不可行

❑ 可行方法



两个关键环节：子集搜索和子集评价

# 子集搜索

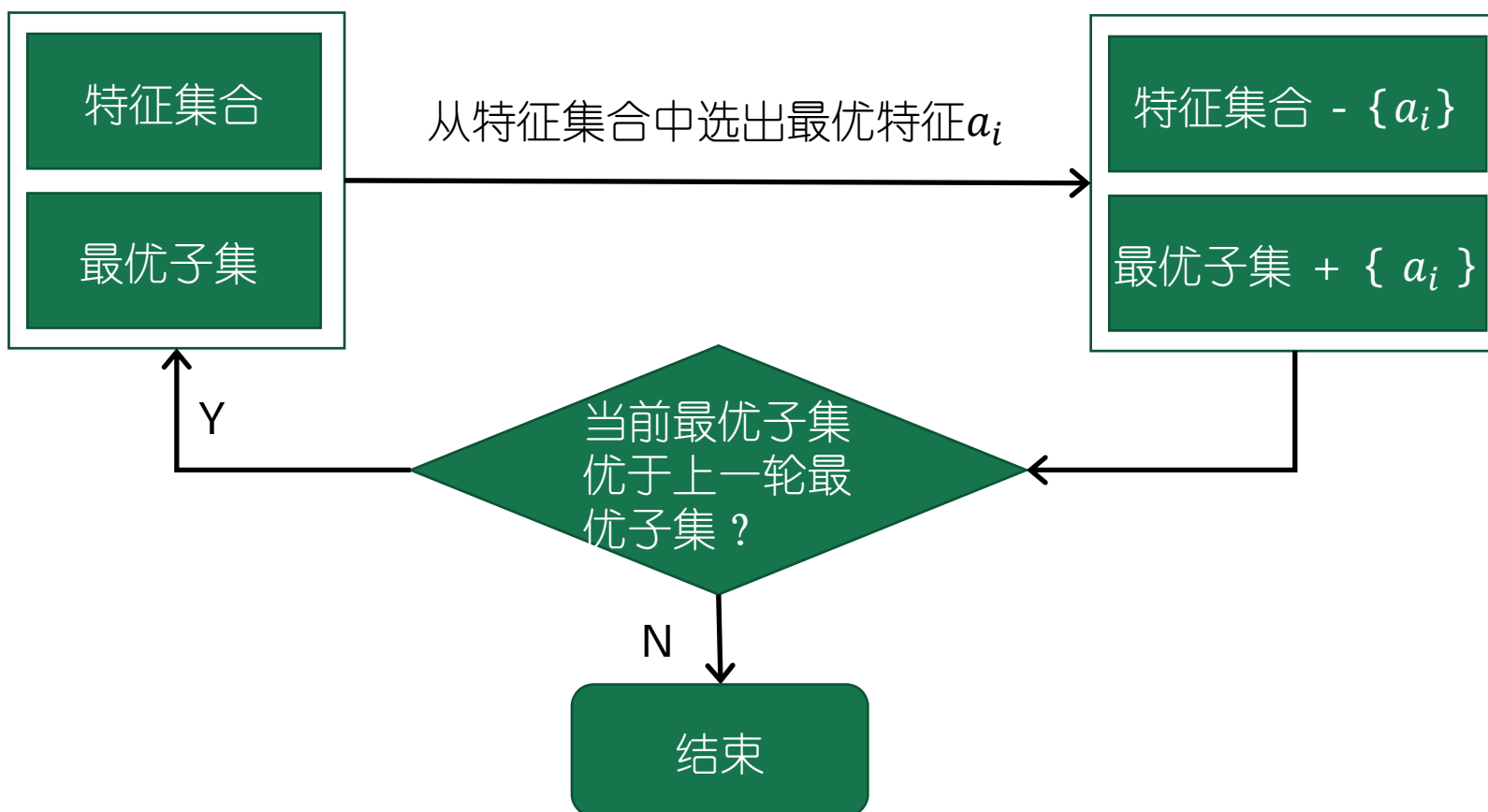
---

用贪心策略选择包含重要信息的特征子集

- 前向搜索：逐渐增加相关特征
- 后向搜索：从完整的特征集合开始，逐渐减少特征
- 双向搜索：每一轮逐渐增加相关特征，同时减少无关特征

# 前向搜索

- 最优子集初始为空集，特征集合初始时包括所有给定特征





# 子集评价

---

- 特征子集确定了对数据集的一个划分
  - 每个划分区域对应着特征子集的某种取值
- 样本标记对应着对数据集的真实划分

通过估算这两个划分的差异，就能对特征子集进行评价；与样本标记对应的划分的差异越小，则说明当前特征子集越好

# 用信息熵进行子集评价

- 特征子集 $A$ 确定了对数据集 $D$ 的一个划分
  - $A$ 上的取值将数据集 $D$ 分为 $V$ 份，每一份用 $D^v$ 表示
  - $\text{Ent}(D^v)$ 表示 $D^v$ 上的信息熵
- 样本标记 $Y$ 对应着对数据集 $D$ 的真实划分
  - $\text{Ent}(D)$ 表示 $D$ 上的信息熵

$D$ 上的信息熵定义为

$$\text{Ent}(D) = - \sum_{i=1}^{|Y|} p_k \log_2 p_k$$

第 $i$ 类样本所占比例为 $p_i$

特征子集 $A$ 的信息增益为：

$$\text{Gain}(A) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v)$$

# 常见的特征选择方法

---

将特征子集搜索机制与子集评价机制相结合，即可得到特征选择方法

常见的特征选择方法大致分为如下三类：

- ❑ 过滤式 (filter)
- ❑ 包裹式 (wrapper)
- ❑ 嵌入式 (embedding)

# 过滤式选择

---

先用特征选择过程过滤原始数据，再用过滤后的特征来训练模型；特征选择过程与后续学习器无关

## □ Relief (Relevant Features) 方法 [Kira and Rendell, 1992]

- 为每个初始特征赋予一个“相关统计量”，度量特征的重要性
- 特征子集的重要性由子集中每个特征所对应的相关统计量分量之和决定
- 设计一个阈值，然后选择比阈值大的相关统计量分量所对应的特征
- 或者指定欲选取的特征个数，然后选择相关统计量分量最大的指定个数特征

如何确定相关统计量？



# Relief方法中相关统计量的确定

□ 猜中近邻 (near-hit) :  $\mathbf{x}_i$  的同类样本中的最近邻  $\mathbf{x}_{i,nh}$

□ 猜错近邻 (near-miss) :  $\mathbf{x}_i$  的异类样本中的最近邻  $\mathbf{x}_{i,nm}$

□ 相关统计量对应于属性  $j$  的分量为

$$\delta^j = \sum_i -\text{diff}(x_i^j, x_{i,nh}^j)^2 + \text{diff}(x_i^j, x_{i,nm}^j)^2$$

若  $j$  为离散型, 则  $x_a^j = x_b^j$  时  
 $\text{diff}(x_a^j, x_b^j) = 0$ , 否则为 1 ;  
若  $j$  为连续型, 则  
 $\text{diff}(x_a^j, x_b^j) = |x_a^j - x_b^j|$ , 注  
意  $x_a^j, x_b^j$  已规范化到  $[0,1]$  区间

□ 相关统计量越大, 属性  $j$  上, 猜对近邻比猜错近邻越近, 即属性  $j$  对区分对错越有用

□ Relief方法的时间开销随采样次数以及原始特征数线性增长, 运行效率很高

# Relief方法的多类拓展

Relief方法是为二分类问题设计的，其扩展变体  
Relief-F [Kononenko, 1994] 能处理多分类问题

- 数据集中的样本来自 $|\mathcal{Y}|$ 个类别，其中 $\mathbf{x}_i$ 属于第 $k$ 类
- 猜中近邻：第 $k$ 类中 $\mathbf{x}_i$ 的最近邻 $\mathbf{x}_{i,nh}$
- 猜错近邻：第 $k$ 类之外的每个类中找到一个 $\mathbf{x}_i$ 的最近邻作为猜错近邻，记为 $\mathbf{x}_{i,l,nm} (l = 1, 2, \dots, |\mathcal{Y}|; l \neq k)$
- 相关统计量对应于属性的分量为

$$\delta^j = \sum_i -\text{diff}(x_i^j, x_{i,nh}^j)^2 + \sum_{l \neq k} \left( p_l \times \text{diff}(x_i^j, x_{i,l,nm}^j)^2 \right)$$

$p_l$ 为第 $l$ 类样本  
在数据集 $D$ 中  
所占的比例

# 包裹式选择

---

包裹式选择直接把最终将要使用的学习器的性能作为特征子集的评价准则

- 包裹式特征选择的目的是为给定学习器选择最有利于其性能、“量身定做”的特征子集
- 包裹式选择方法直接针对给定学习器进行优化，因此从最终学习器性能来看，包裹式特征选择比过滤式特征选择更好
- 包裹式特征选择过程中需多次训练学习器，计算开销通常比过滤式特征选择大得多

# LVW包裹式特征选择方法

---

LVW (Las Vegas Wrapper) [Liu and Setiono, 1996] 在拉斯维加斯方法框架下使用随机策略来进行子集搜索，并以最终分类器的误差作为特征子集评价准则

## 基本步骤

- 在循环的每一轮随机产生一个特征子集
- 在随机产生的特征子集上通过交叉验证推断当前特征子集的误差
- 进行多次循环，在多个随机产生的特征子集中选择误差最小的特征子集作为最终解\*

\*若有运行时间限制，则该算法有可能给不出解



# LVW包裹式特征选择方法

---

输入: 数据集  $D$ ;  
特征集  $A$ ;  
学习算法  $\mathfrak{L}$ ;  
停止条件控制参数  $T$ .

过程:

```
1:  $E = \infty$ ;  
2:  $d = |A|$ ;  
3:  $A^* = A$ ;  
4:  $t = 0$ ;  
5: while  $t < T$  do  
6:   随机产生特征子集  $A'$ ;  
7:    $d' = |A'|$ ;  
8:    $E' = \text{CrossValidation}(\mathfrak{L}(D^{A'}))$ ;  
9:   if  $(E' < E) \vee ((E' = E) \wedge (d' < d))$  then  
10:     $t = 0$ ;  
11:     $E = E'$ ;  
12:     $d = d'$ ;  
13:     $A^* = A'$   
14:   else  
15:     $t = t + 1$   
16:   end if  
17: end while
```

输出: 特征子集  $A^*$ .

---

# 嵌入式选择

嵌入式特征选择是将特征选择过程与学习器训练过程融为一体，两者在同一个优化过程中完成，在学习器训练过程中自动地进行特征选择

- 考虑最简单的线性回归模型，以平方误差为损失函数，并引入 $L_2$ 范数正则化项防止过拟合，则有

$$\min_{\mathbf{w}} \sum_{i=1}^m (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_2^2$$

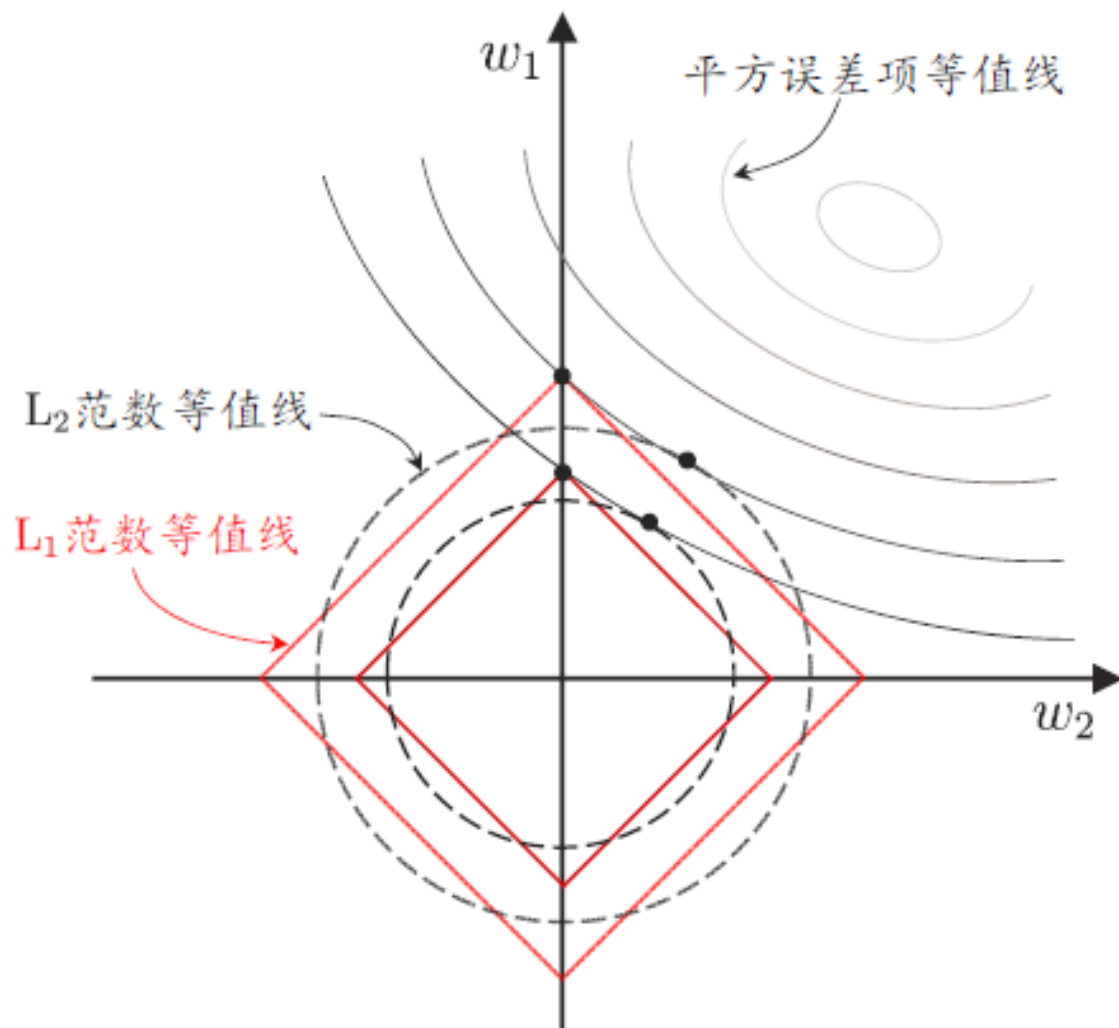
岭回归 (ridge regression)  
[Tikhonov and Arsenin, 1977]

- 将 $L_2$ 范数替换为 $L_1$ 范数，则有**LASSO** [Tibshirani, 1996]

$$\min_{\mathbf{w}} \sum_{i=1}^m (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|_1$$

易获得稀疏解，是一种嵌入式特征选择方法

# 使用 $L_1$ 范数正则化易获得稀疏解



假设 $\mathbf{x}$ 仅有两个属性，那么 $\mathbf{w}$ 有两个分量 $w_1$ 和 $w_2$ 。那么目标优化的解要在平方误差项与正则化项之间折中，即出现在图中平方误差项等值线与正则化等值线相交处。

从图中看出，采用 $L_1$ 范数时交点常出现在坐标轴上，即产生 $w_1$ 或者 $w_2$ 为0的稀疏解。

等值线即取值相同的点的连线

# L1正则化问题的求解(1)

---

$$\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$$

近端梯度下降 (Proximal Gradient Descend, 简称PGD) 解法  
[Boyd and Vandenberghe, 2004]

□ 写出 $f(\mathbf{x})$ 的二阶泰勒展式

$$f(\mathbf{x}) = f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^\top \frac{\delta^2 f(\mathbf{x}_k)}{\delta \mathbf{x}_k^2} (\mathbf{x} - \mathbf{x}_k)$$

□ 假设 $f(\mathbf{x})$ 满足L-Lipschitz条件, 即存在常数 $L > 0$ , 使得

$$\|\nabla f(\mathbf{x}') - \nabla f(\mathbf{x})\| \leq L \|\mathbf{x}' - \mathbf{x}\|_2$$

# L1正则化问题的求解(2)

---

□ L-Lipschitz条件代入泰勒展式, 可得

$$\begin{aligned} f(\mathbf{x}) &\cong f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{L}{2} \|\mathbf{x} - \mathbf{x}_k\|^2 \\ &= \frac{L}{2} \|\mathbf{x} - (\mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k))\|^2 + \text{const} \end{aligned}$$

□ 将上式关于 $f(\mathbf{x})$ 的近似代入到原优化问题中, 得

$$\min_{\mathbf{x}} \sum_{i=1}^m \frac{L}{2} \|\mathbf{x} - (\mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k))\|^2 + \lambda \|\mathbf{x}\|_1$$

# L1正则化问题的求解(3)

---

- 每次在 $\mathbf{x}_k$ 的附近寻找最优点，不断迭代，即寻找

$$\mathbf{x}_{k+1} = \min_{\mathbf{x}} \sum_{i=1}^m \frac{L}{2} \|\mathbf{x} - (\mathbf{x}_k - \frac{1}{L} \nabla f(\mathbf{x}_k))\|^2 + \lambda \|\mathbf{x}\|_1$$

- 假设 $\mathbf{z} = \mathbf{x}_k - 1/L \nabla f(\mathbf{x}_k)$ ，上式有闭式解

$$x_{k+1}^i = \begin{cases} z^i - \lambda/L, & \lambda/L < z^i ; \\ 0, & |z^i| \leq \lambda/L ; \\ z^i + \lambda/L, & z^i < -\lambda/L , \end{cases}$$