

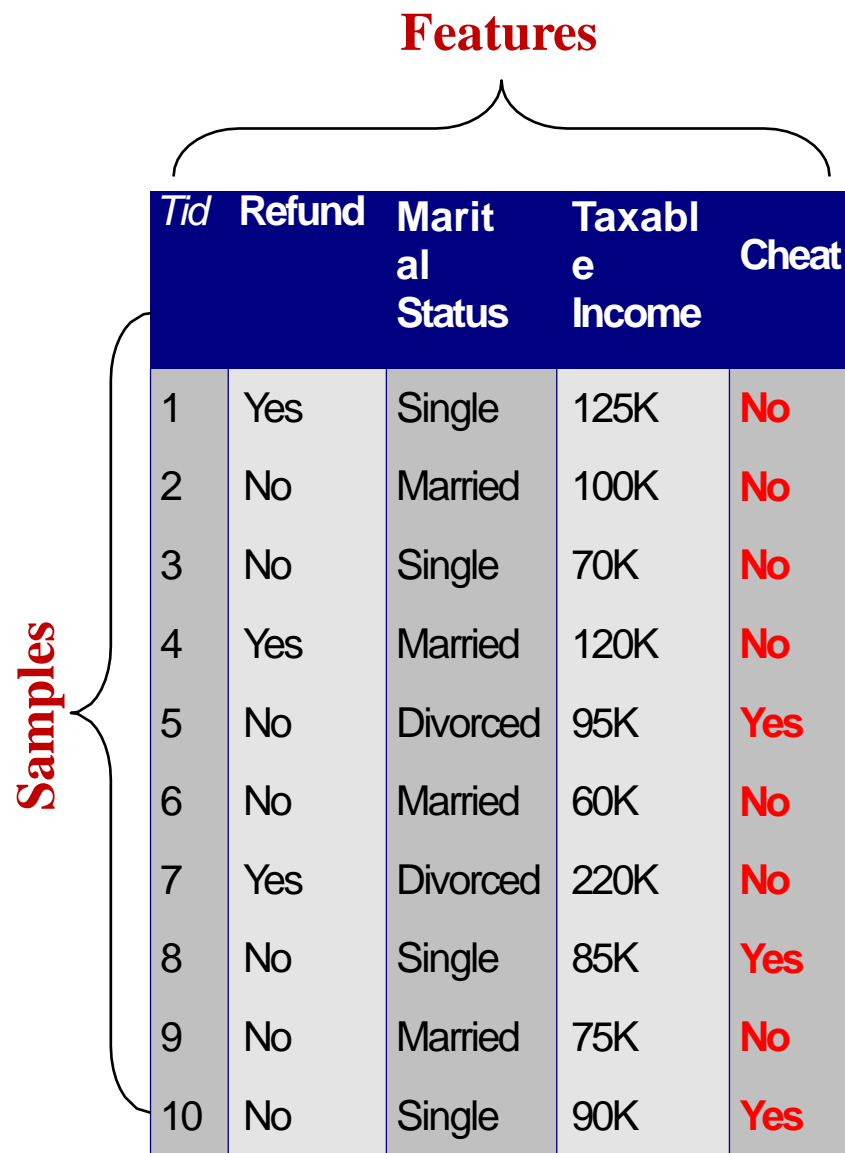
数据挖掘导论

第3章：特征工程

- (1) 陈封能等, 《数据挖掘导论》, 机械工业出版社, 2019
- (2) Alice Zheng, Amanda Casari, 《Feature Engineering for Machine Learning》, 2018.
(中文版《精通特征工程》, 人民邮电出版社, 2019)
- (3) 锡南·厄兹代米尔, 迪夫娅·苏萨拉, 《特征工程入门与实践》(中文版), 人民邮电出版社, 2019

数据 (data)

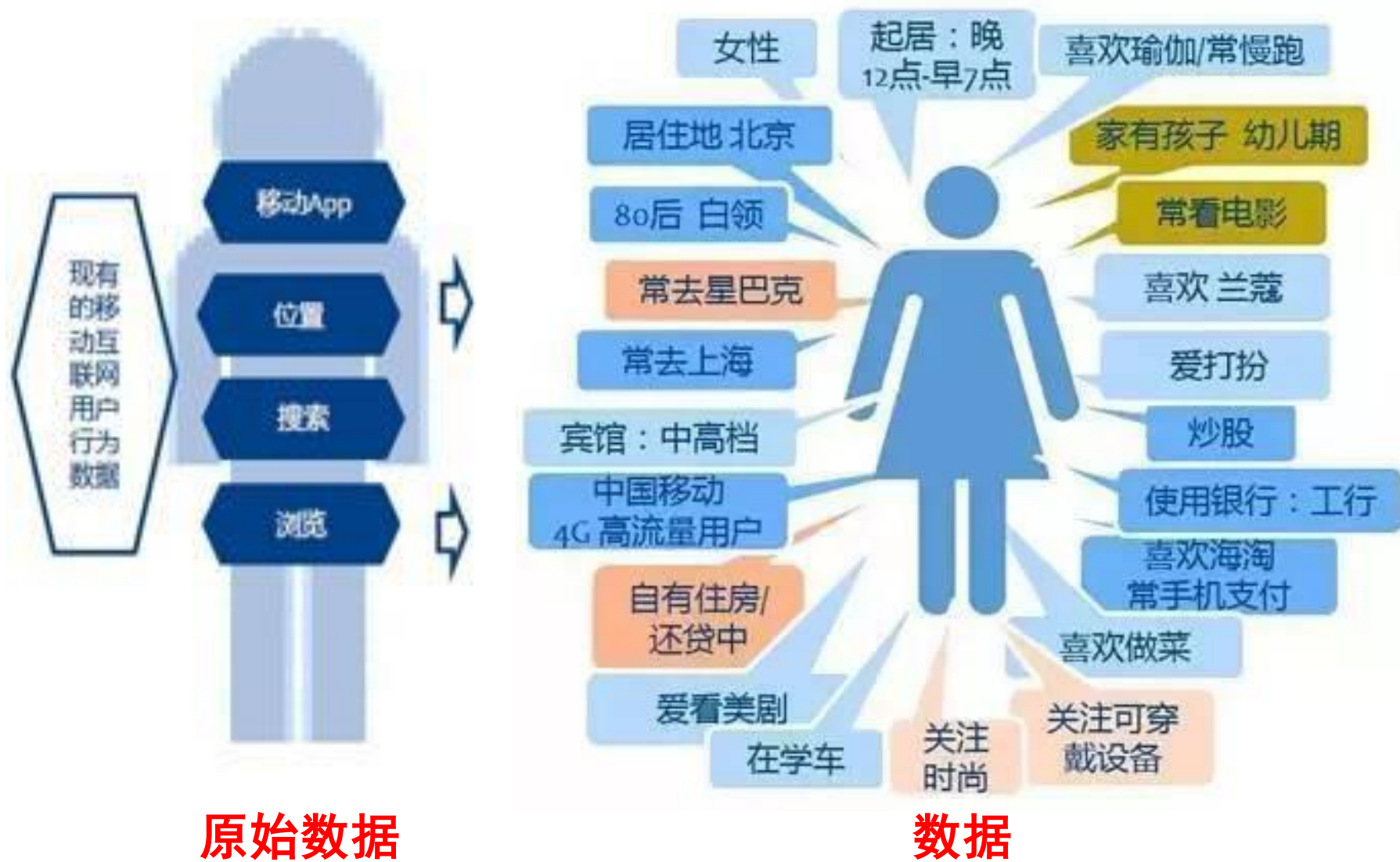
- **数据集**：一组含有**特征**的**样本**集合
- **特征**：描述样本某一方面性质的变量
 - 例如：某个人的身高、北京某一时刻的气温等等
 - Feature is also known as variable, field, characteristic, or attribute
- **样本**：由一组特征所描述的一个对象
 - Sample is also known as record, point, case, object, entity, or instance



The diagram illustrates the relationship between features and samples in a dataset. A bracket labeled "Features" spans the columns of the table, and a bracket labeled "Samples" spans the rows.

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

数据从哪里来?



特征 (Feature)

- **特征**是所有独立单元共享的属性，是进行分析或预测的对象。只要对模型有用，任何属性都可以是特征
- “A feature is a characteristic that might help when solving the problem”，可理解为：特征只有在某个具体需要解决的问题下讨论才具有意义。

特征的重要性

- DataRobot首席数据科学家、在2012-2013年Kaggle榜上排名第一的Xavier Conort曾说过，团队使用的算法在Kagglers中是十分常规的。他们将大多数精力花在**特征工程**上。他们也非常小心地放弃了可能会使模型产生过拟合的特征。
- “一些机器学习项目成功了而另外一些却失败了，是什么导致了这种差别呢？很简单，最关键的因素出现在他们所使用的**特征**上”，《终极算法：机器学习和人工智能如何重塑世界》作者Pedro Domingos的回答更为直接。

特征的重要性

(1) 特征越好，灵活性越强

只要特征选得好，即使是一般的模型（或算法）也能获得很好的性能，因为大多数模型（或算法）在好的数据特征下表现的性能都还不错。**好特征的灵活性在于它允许你选择不复杂的模型，同时运行速度也更快，也更容易理解和维护。**

(2) 特征越好，构建的模型越简单

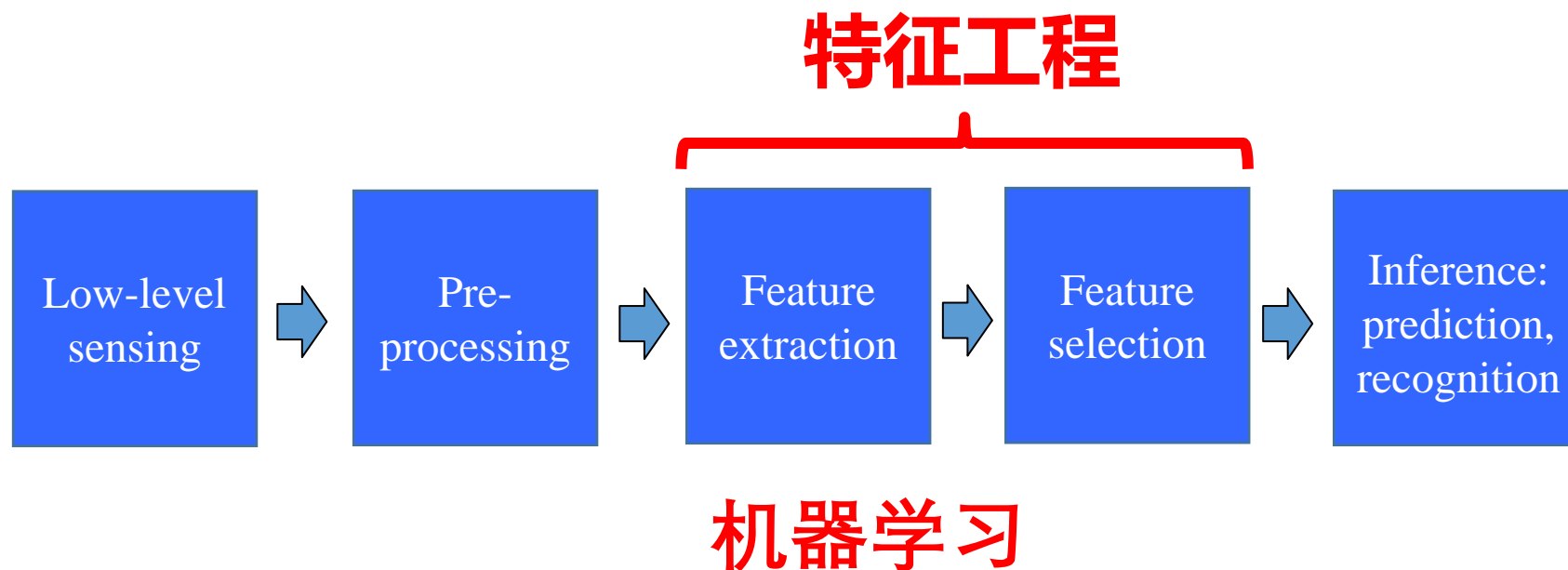
有了好的特征，即便你的参数不是最优的，你的模型性能也能仍然会表现的很好，所以你就不需要花太多的时间去寻找最优参数，这大大的降低了模型的复杂度，使模型趋于简单。

(3) 特征越好，模型的性能越出色

我们进行特征工程的最终目的就是提升模型的性能。

特征工程

- Feature engineering is the process of using domain knowledge of the data to create features that **make machine learning algorithms work**. (特征工程是利用数据领域的相关知识来创建能够使机器学习算法达到最佳性能的特征的过程) -Wikipedia
- Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive. (特征工程是机器学习应用的基础, 既困难又昂贵)



特征工程

- 特征工程就是在给定数据、模型和任务的情况下设计出最合适的特征的过程。
 - 《精通特征工程》
- 特征工程本质是一项工程活动，目的是最大限度地从原始数据中提取特征以供算法和模型使用
- 特征工程是一个非正式的话题，但在应用机器学习中却被认为是必不可少的，因为原始数据中通常有大量的特征，这是不可能全部处理的。特征工程消除了噪声，提高了模型的精度。它优化了训练过程，推理过程，消除了过拟合。

特征工程

- Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering. (创造特征是很困难的，既耗时，还需要专业知识。“应用机器学习”基本上是特征工程)

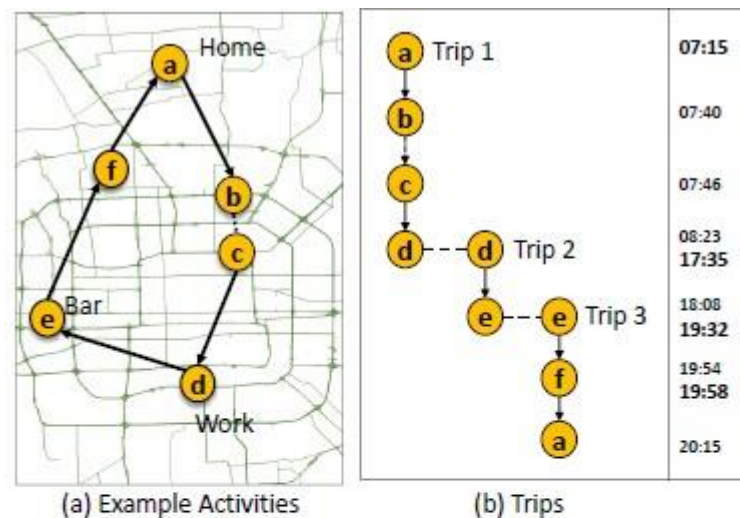
— Andrew Ng (吴恩达), Machine Learning and AI via Brain simulations (slide, 2013)

- 数据和特征决定了机器学习的**上限**，而模型和算法只是逼近这个上限而已。



如何提取特征？

原始数据



Smart Card ID	Route Number	Boarding Station	Boarding Time	Exiting Station	Exiting Time
4322	Route 52	a	07:15	b	07:40
4322	Route 26	c	07:46	d	08:23
4322	Route 11	d	17:35	e	18:08
4322	Route 11	e	19:32	f	19:54
4322	Route 16	f	19:58	a	20:15

(c) Transit Records

Figure 3: An example of trips and transit records.

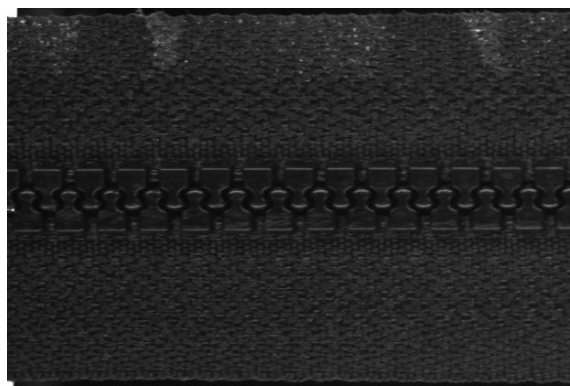


Table 1: Categories of Functional Regions

Category	Examples	Frequency
Home	Apartment buildings	28,731
Work	Government or office buildings	71,364
Education	Schools, training centers	3,527
Food	Restaurants and dining	56,906
Shopping	Shopping malls and outlets	24,310
Entertainment	Museums, theaters, clubs	18,223
Scenic Spot	Parks, sports fields	2,362
Transportation	Airports, transit centers	15,287
Healthcare	Hospitals, pharmacy	8,685
Car services	Car sales, repairs	1,781

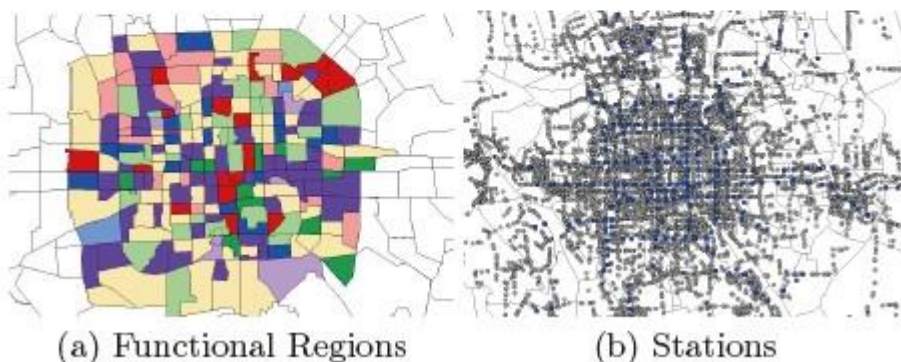


Figure 4: Geographical information.

#抓小偷为百姓服务# 7月10日早7时40分，便衣反扒小伟队长带大彪组在349路程庄路口东车站抓获一名男小偷，现案为乘客挽回公交IC卡一张，真应了那句话叫：贼不走空！嫌疑人吴某，27岁，健壮有力、聪明伶俐，选择不劳而获确实是不应该，目前被公交警方刑事拘留。PS：想起《小花》了😊



今天 11:00 来自 微博 weibo.com

(a) Police report: "At 7:40am on July 10th, a thief was caught at Route 349 East Chengzhuanglukou Station."

Figure 5: Example incident reports on Weibo.

特征工程

- 数据清洗 **Data cleaning**
- 数据预处理 **Data Pre-processing**
- 特征构建 **Feature Construction**
- 特征提取 **Feature Extraction**
- 特征选择 **Feature Selection**

数据清洗

(Data cleaning)

数据质量

- 数据质量问题对于数据处理的效果有着非常严重的影响
 - “The most important point is that poor data quality is an **unfolding disaster**. Poor data quality costs the typical company at least ten percent (10%) of revenue; twenty percent (20%) is probably a better estimate.”
 - (最重要的一点是，不良的数据质量是一个不断发展的灾难。不良的数据质量使典型的公司损失了至少百分之十（10%）的收入；百分之二十(20%)可能是一个更好的估计)

----Thomas C. Redman, DM Review, August 2004

• 实际案例

- 使用poor data进行银行贷款用户的征信评估
 - 一些credit-worthy用户的申请可能会被拒绝
 - 一些贷款可能会被放给不可靠的用户

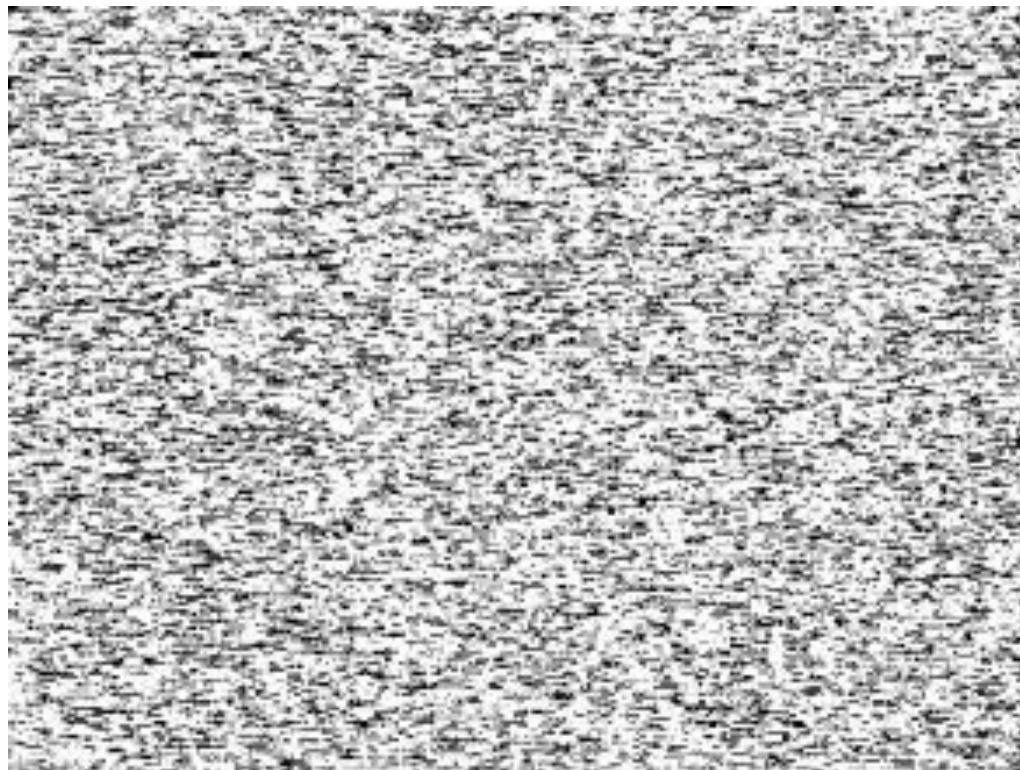
数据质量问题

- 数据当中存在哪些类型的质量问题？
- 如何检测出来这些质量问题？
- 对于这些质量问题，我们可以做什么？

- 数据质量问题的一些例子
 - 噪声与离群点
 - 数据缺失 Missing value
 - 数据冗余 Duplicate data

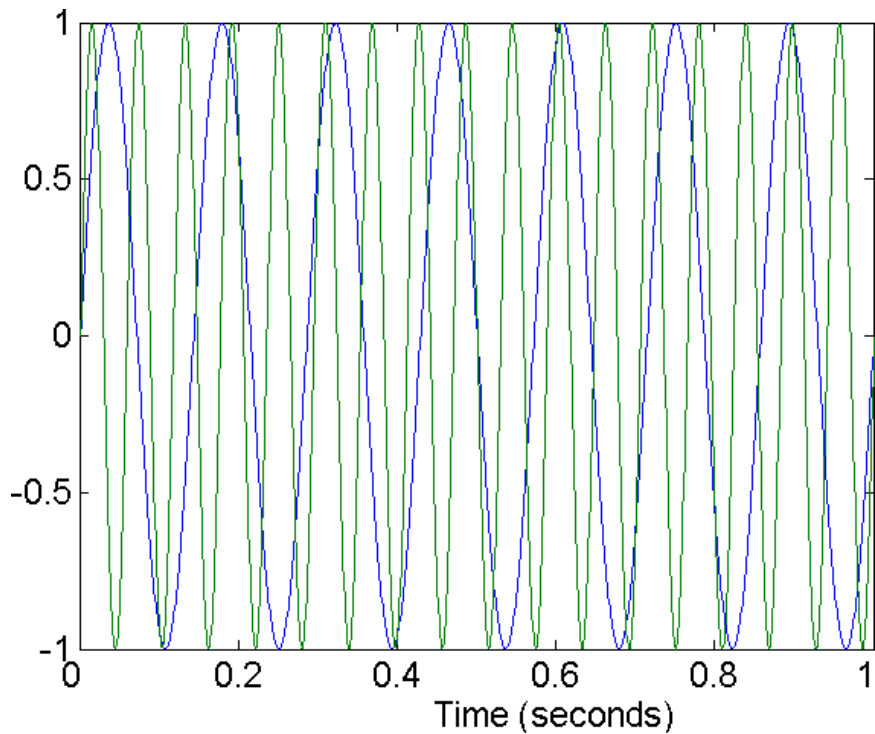
噪声 Noise

- Noise refers to modification of original values
 - 噪声是测量误差的随机部分
 - 例如：电视机中的“雪花”信号

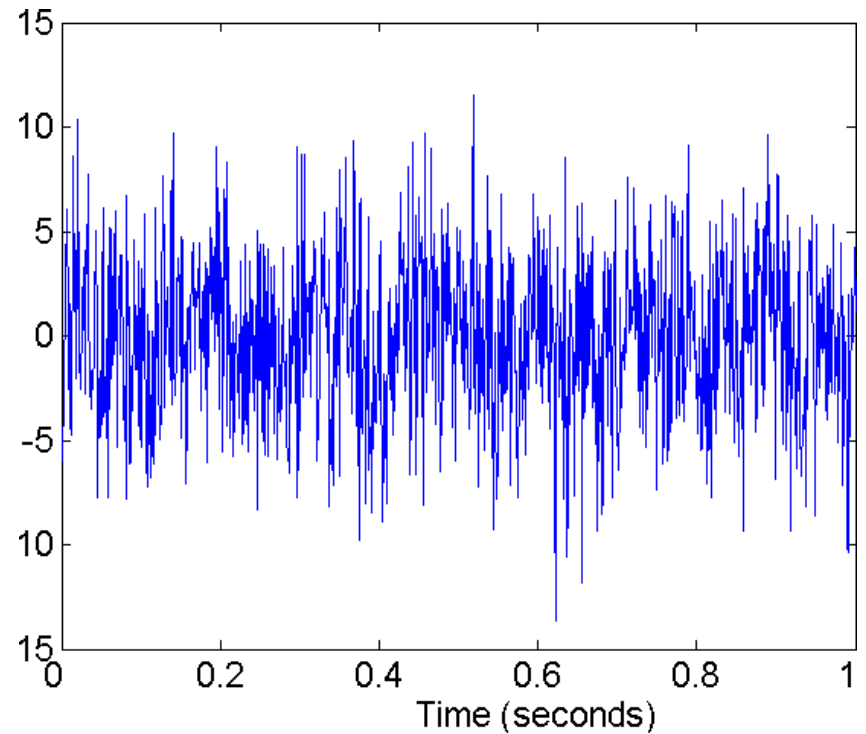


噪声 Noise

- Noise refers to modification of original values
 - 噪声是测量误差的随机部分



Two Sine Waves

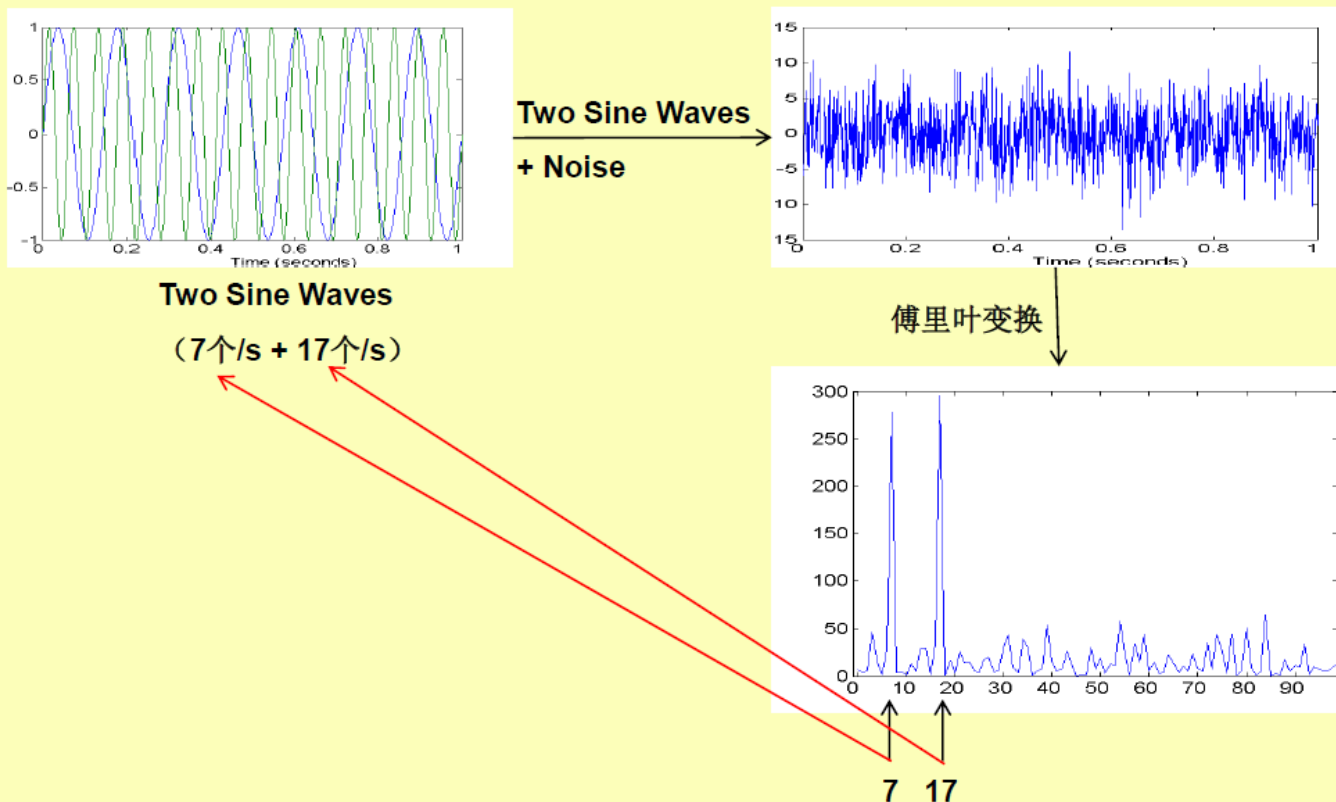


Two Sine Waves + Noise

噪声 Noise

• 噪声的处理：引入降噪算法

□ 正弦波噪声的例子

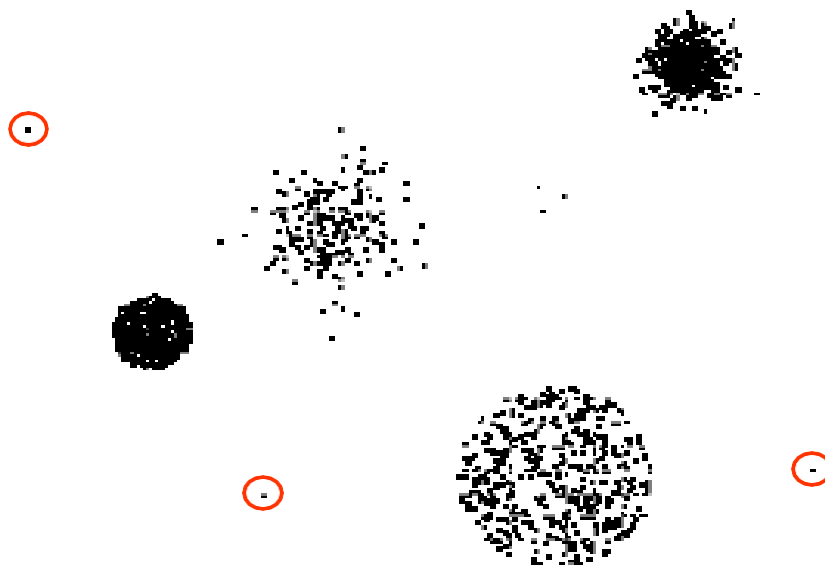


傅里叶变换的应用：识别时间序列数据中的基本频率

映射数据到新的空间。尽管有噪声，图中有2个尖峰，对应于2个原来的、无噪声的时间序列的周期

离群点 Outliers

- 离群点是具有**不同于数据集中其他大部分数据**样本特征的数据样本
 - Case1: 离群点可能是噪声 (**不合法**)
 - Case2: 离群点可能是**合法数据**，即异常检测的目标
 - 例如：信用欺诈、入侵检测



缺失数据 Missing Value

- 数据在收集过程中缺失

- 完全随机缺失 (missing completely at random, MCAR)

- 指的是数据的缺失是完全随机的，不依赖于任何不完全变量或完全变量，不影响样本的无偏性。如家庭地址缺失。

- 随机缺失(missing at random, MAR)

- 指的是数据的缺失不是完全随机的，即该类数据的缺失依赖于其他完全变量。例如财务数据缺失情况与企业的大小有关。

- 非随机缺失(missing not at random, MNAR)

- 指的是数据的缺失与不完全变量自身的取值有关。如高收入人群的不原意提供家庭收入。

缺失数据 Missing Value

- 如何处理缺失数据

- 删除相应的数据对象
- 在计算和分析过程中忽略缺失值
- 对缺失数据进行估计
 - 用平均值、中值、分位数、众数、随机值等替代。
 - 用其他变量做预测模型来算出缺失变量。有一个根本缺陷，如果其他变量和缺失变量无关，则预测的结果无意义。如果预测结果相当准确，则又说明这个变量是没必要加入建模的。
 - 把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。连续型变量也可以这样处理。这样做的好处是完整保留了原始数据的全部信息，缺点是计算量大大提升。

重复的数据 Duplicate Data

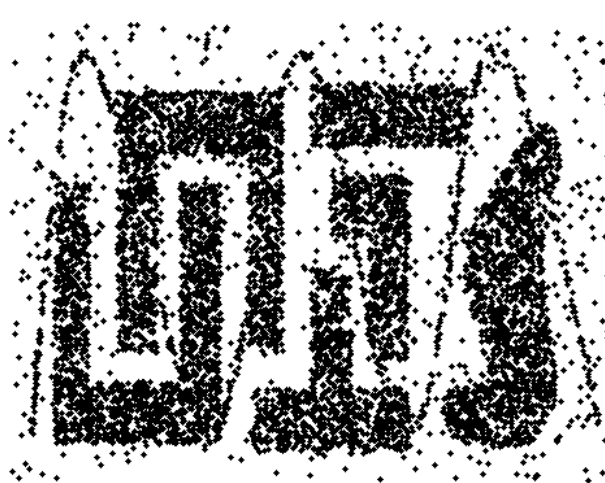
- 数据集可能包含重复或者几乎重复的数据
 - 两个或多个对象实际代表单个对象
 - 例如：同一人拥有的多个电子邮件地址
 - 避免将两个相似但并非重复的数据对象组合在一起
 - 例如：同名同姓的人
- 需要注意区分是**数值重复** 不**数据对象重复**
- 产生的主要原因
 - 对异种数据源进行数据合并时
- 数据清洗
 - 去重处理

数据矛盾

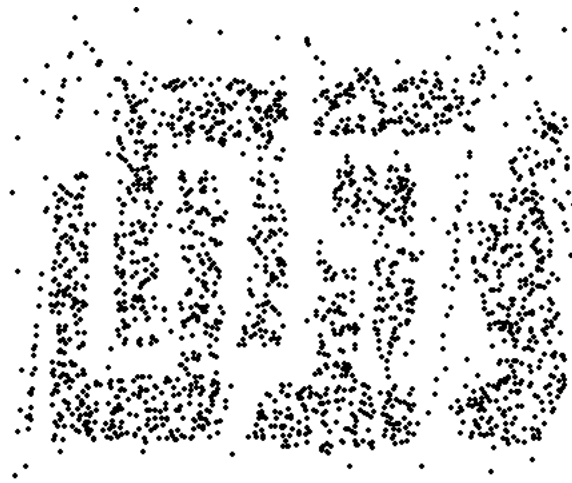
- 数据样本中存在自相矛盾的特征
 - 是否打过疫苗=Y, 是否打过乙肝疫苗=Y
 - 是否打过疫苗=Y, 是否打过乙肝疫苗=N
 - **是否打过疫苗=N, 是否打过乙肝疫苗=Y**
- 处理方式
 - 和数据来源进行确认, 评估不同特征的可靠性
 - 修正矛盾的数据。

采样 Sampling

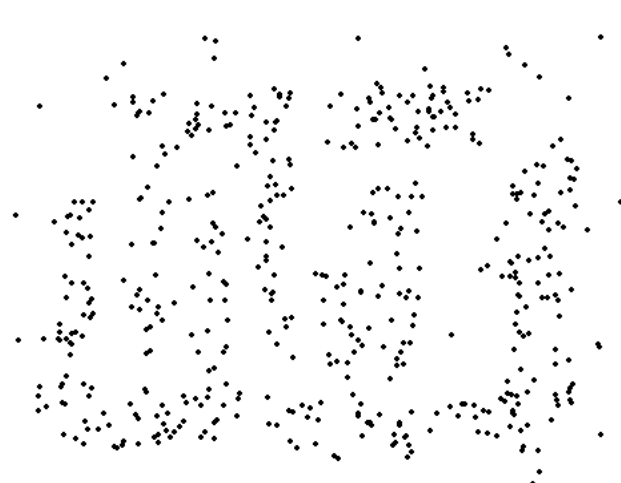
- 数据的采样主要用在当获得全量数据非常困难或是（时间）成本过高时
- 进行数据采样成功的关键在于：采样样本本身能够代表全样本的统计特性



8000 points



2000 Points



500 Points

采样 Sampling

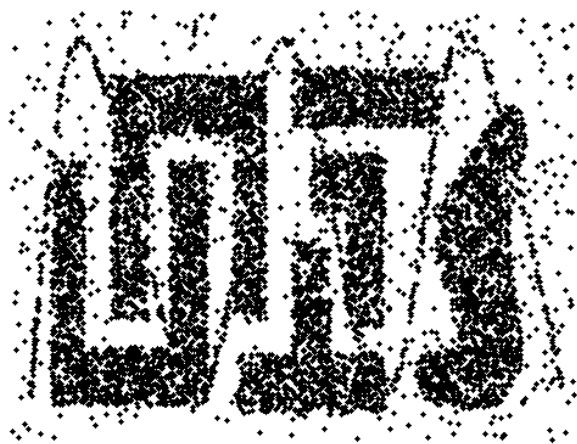
- 采样是一种选择数据对象子集进行分析的常用方法
 - 常用于数据的事先调查和最终的数据分析
 - 获得感兴趣的整个数据集的成本太高、太费时间
 - 处理整个数据集的成本太高、太费时间
- 进行数据采样成功的关键在于：**采样样本本身能够代表全样本的统计特性**
- 有效采样的原则：
 - 代表性：如果样本是有代表性的，则使用样本与使用整个数据集的效果几乎一样
 - 保留原数据集的性质：如果样本近似地具有原数据集相同的性质，则称它是有代表性的

采样方法

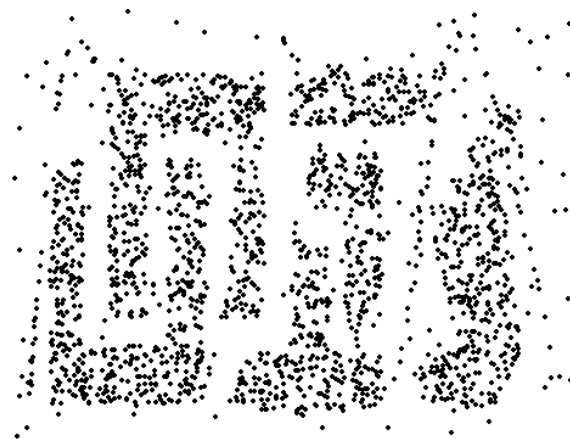
- 简单采样(Simple Random Sampling)
 - 选取任何特定项的概率相等
 - 简单无放回采样(Sampling without replacement)
 - 每个选中项立即从构成总体的所有对象集中删除
 - 简单有放回采样(Sampling with replacement)
 - 对象被选中时不从总体中删除
 - 在有放回采样中, 相同的对象可能被多次抽中
- 分层采样(Stratified sampling)
 - 将数据分成几个组; 再分别从各个组中随机采样
 - 每组抽相同个数 vs 按比例
- 自适应(adaptive)或渐进采样(progressive sampling)
 - 原因: 有时难以预先确定样本集大小
 - 方法: 从一个小样本开始, 然后增加样本容量直至得到足够容量的样本

样本容量: 例子

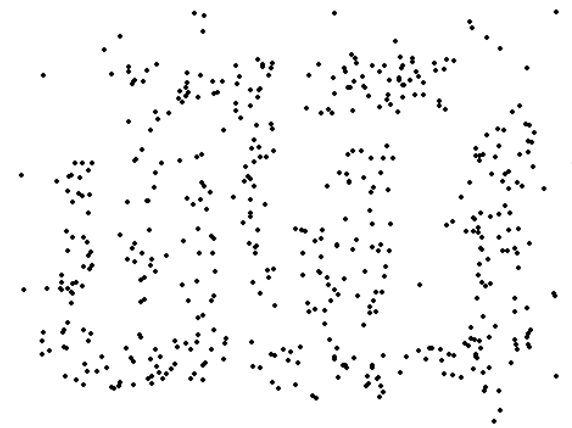
- 从8000个点分别抽2000和500个点
 - 2000个点的样本保留了数据集的大部分结构
 - 500个点的样本丢失了许多结构



8000 points



2000 Points



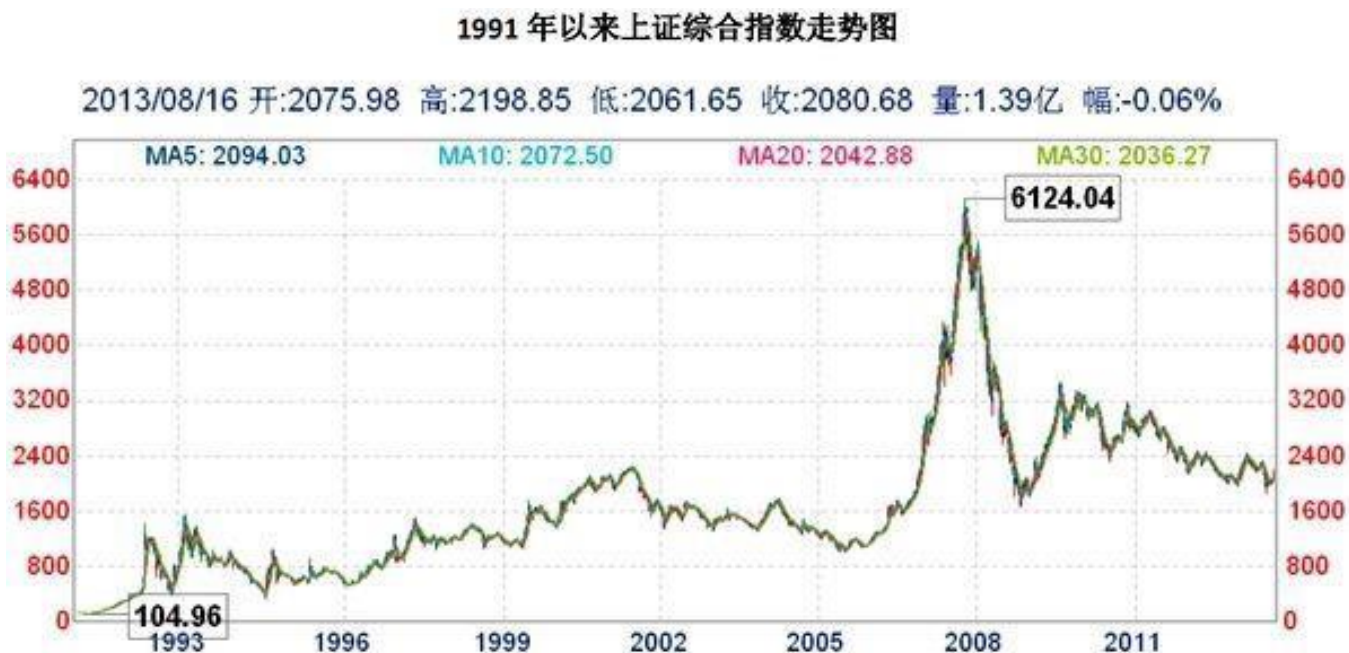
500 Points

数据预处理

(Data Pre-processing)

无量纲化

- 特征不属于同一量纲，即特征的规格不一样，不能够放在一起比较。
 - 特征的量纲标准不同：如英寸和厘米如何比较？
 - 特征的量纲类型不同：公斤与厘米如何比较？
 - 特征的分布尺度不同：人类的身高和大象的身高如何比较？



• 归一化

- 样本除以总样本的均值

$$AM(x_1, \dots, x_n) = \frac{1}{n} (x_1 + \dots + x_n) \quad \text{算术平均}$$

$$\hat{x}_n = \frac{x_n}{\bar{x}} \quad GM(x_1, \dots, x_n) = \sqrt[n]{|x_1 \times \dots \times x_n|} \quad \text{几何平均}$$

$$HM(x_1, \dots, x_n) = \frac{n}{\frac{1}{x_1} + \dots + \frac{1}{x_n}} \quad \text{调和平均}$$

$$\min \leq HM \leq GM \leq AM \leq \max$$

无量纲化

• Min-max 标准化

- **区间缩放法**：将特征的取值区间缩放到某个特点的范围，例如[0, 1]等。

$$x' = \frac{x - Min}{Max - Min}$$

• Z-score 标准化

- 标准化的前提是特征值服从**正态分布**，标准化后，其转换成**标准正态分布**。

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

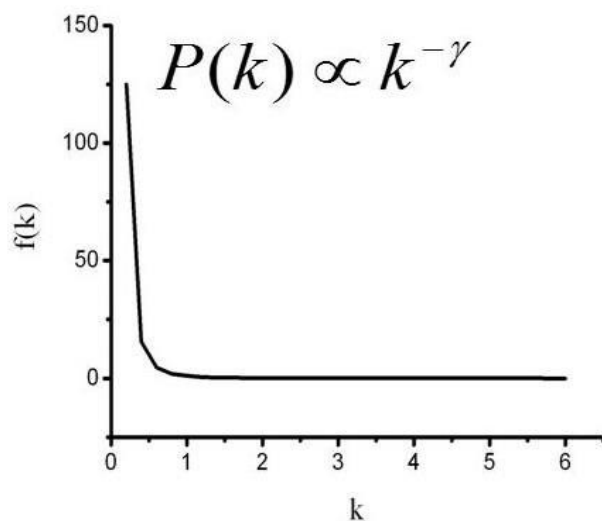
μ is the mean of the population.

σ is the standard deviation of the population.

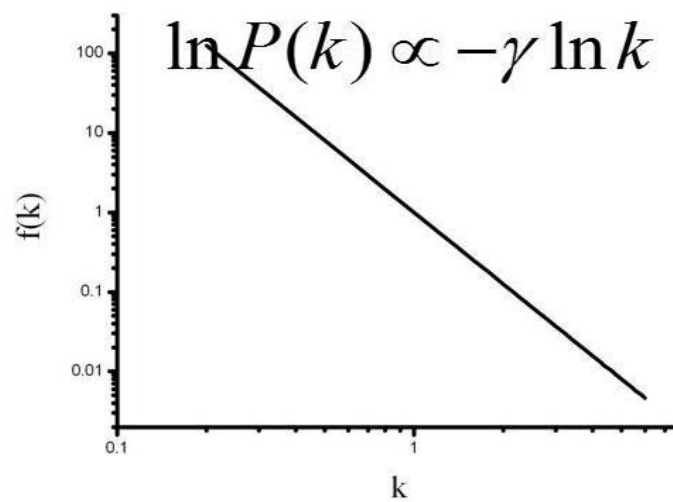
无量纲化

- 对数Logistic标准化

$$\hat{x}_n = \frac{\ln(1 + x_n)}{\ln(1 + x_{max})}.$$



原始坐标系



双对数标系

标准化方法

标准化方法	适用	特点	公式	范围
Min-max标准化：原始数据进行线性变换	适合于最大最小值明确不变。如果最大最小值易变，容易使得归一化结果不稳定，使得后续使用效果也不稳定。实际应用中经常用常量来替代。	简单易理解，不改变数据分布	$(\text{原数据} - \text{极小值}) / (\text{极大值} - \text{极小值})$	$[0, 1]$
z-score标准化：	SPSS默认的标准化方法。适合于最大值和最小值未知的情况，接近正态分布	改变数据分布，对离群点规范化效果好	$(\text{原数据} - \text{均值}) / \text{标准差}$	$(?, ?)$
对数Logistic标准化	长尾分布的数据，需要对样本作分段操作	完全改变数据分布	$\ln(\text{原数据}) / \ln(\text{最大值})$	$(0, 1]$
Decimalscaling 小数定标标准化	只适合数据初期探索。不消除属性间的权重差异	直观简单，不改变数据分布。	$\text{原数据} / (10^j)$	$(0, 1]$
排序归一	适用待归一的具体值并不是很关心，更关心相对排序	数据变成直线分布	$\text{Rank} / \text{数据记录数} N$	$(0, 1]$
分段归一	较复杂的归一操作，适合分布有明显分段特征的数据	更能贴近业务，整合多个方法优点	先根据业务经验分段，在不同段内用各种方法	$[0, 1]$

特征离散化

- 信息冗余:

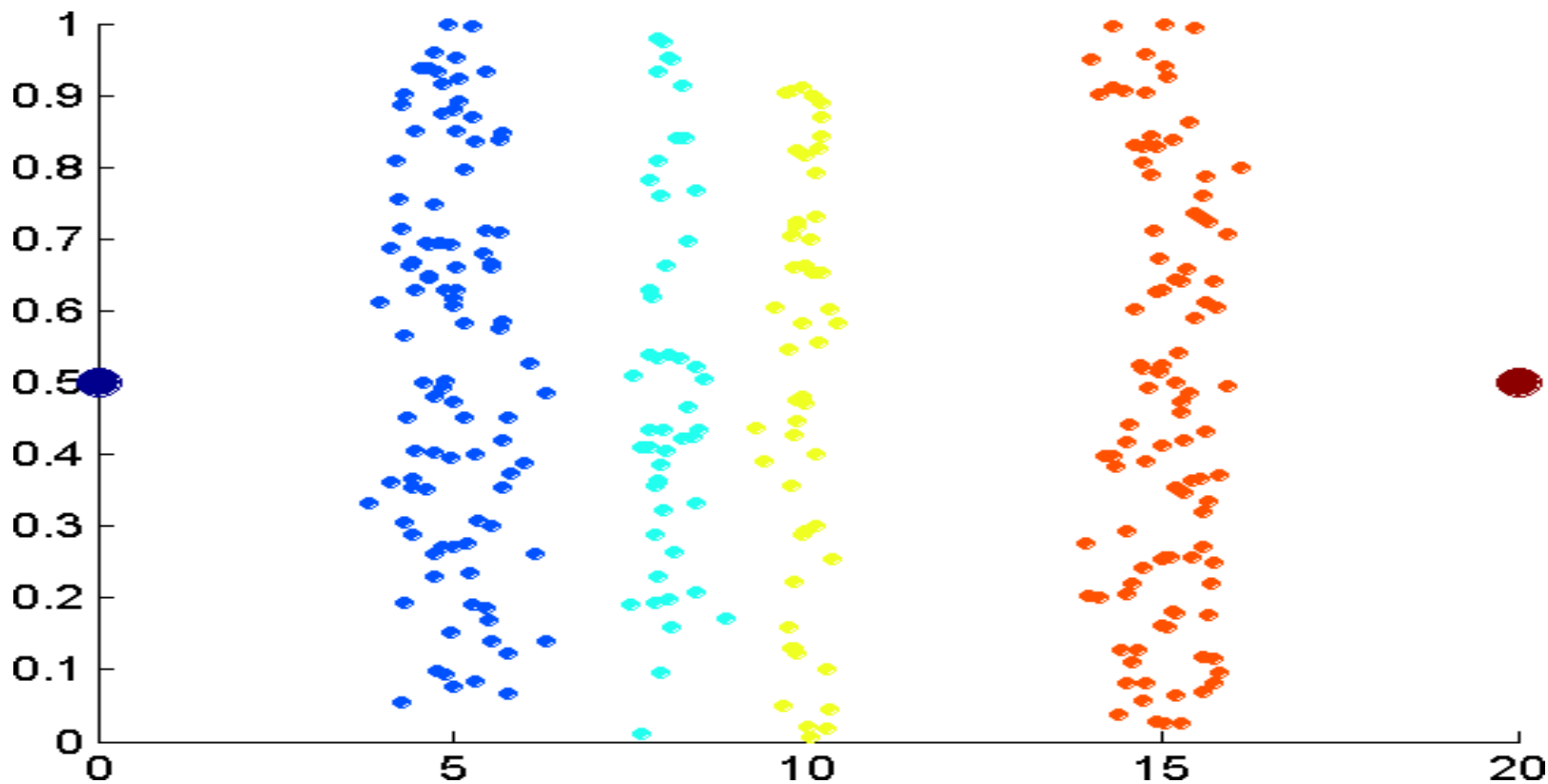
- 对于某些定量特征，其包含的有效信息为区间划分，例如学习成绩，假若只关心“及格”或“不及格”，那么需要将定量的考分，转换成“1”和“0”表示及格和未及格。二值化可以解决这一问题。

- 定量特征二值化

- 设定一个阈值，大于阈值的赋值为1，小于等于阈值的赋值为0，公式表达如下：

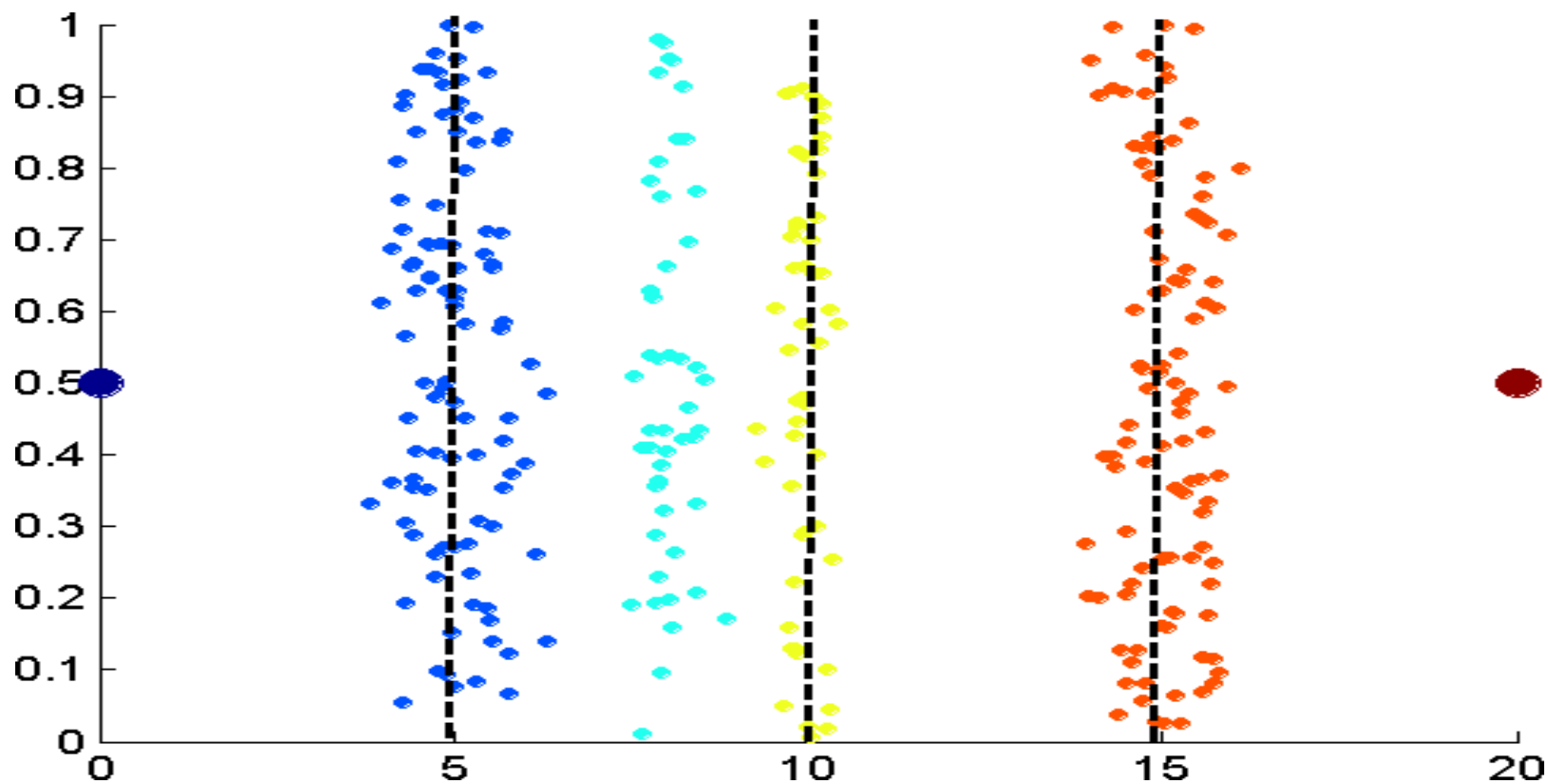
$$x' = \begin{cases} 1, & x > threshold \\ 0, & x \leq threshold \end{cases}$$

特征离散化-unsupervised



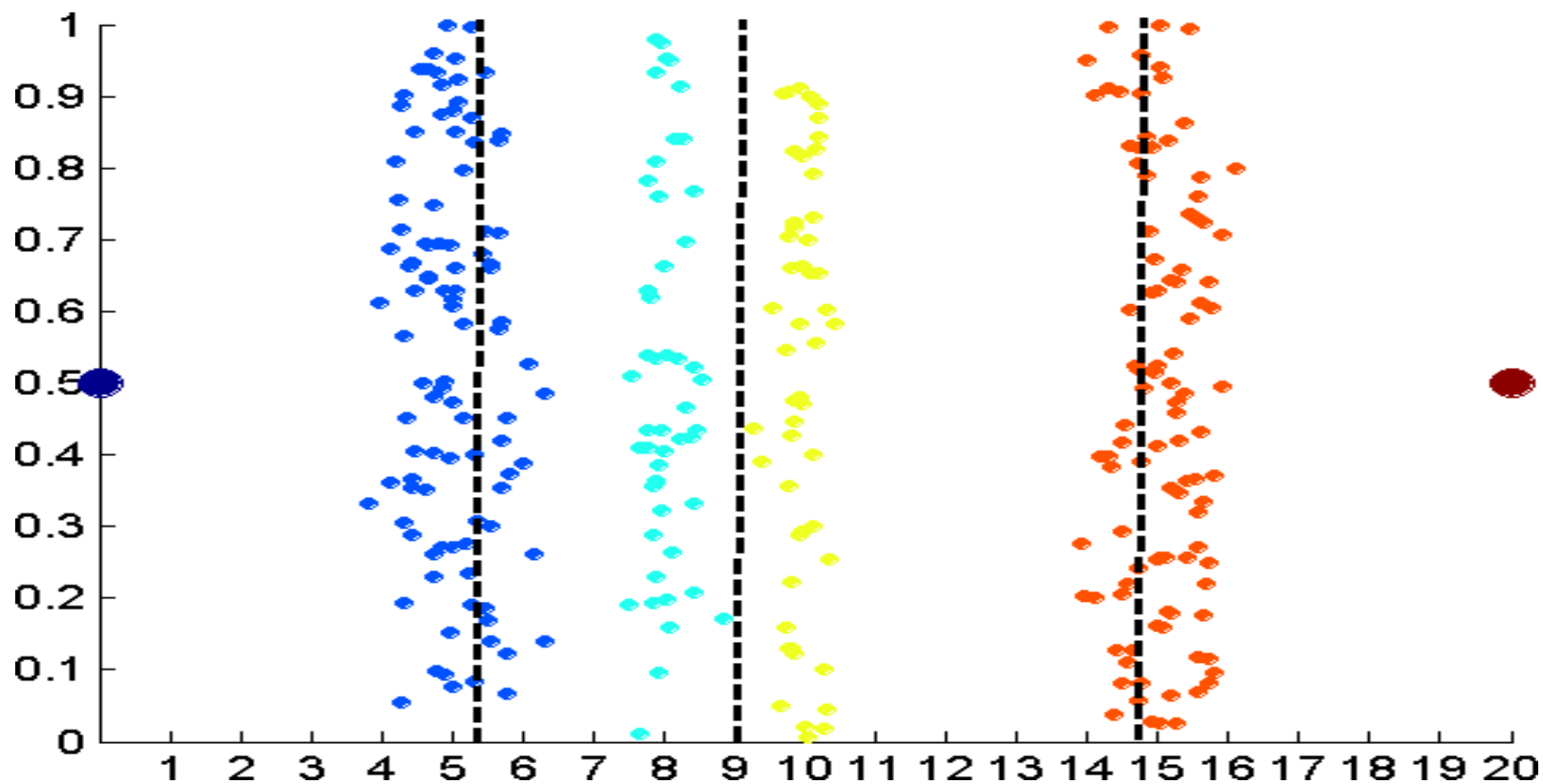
数据当中包含了四个group以及两个离群点，数据本身是一维的，为了避免数据的重叠，我们将其进行了二维化。

特征离散化-unsupervised



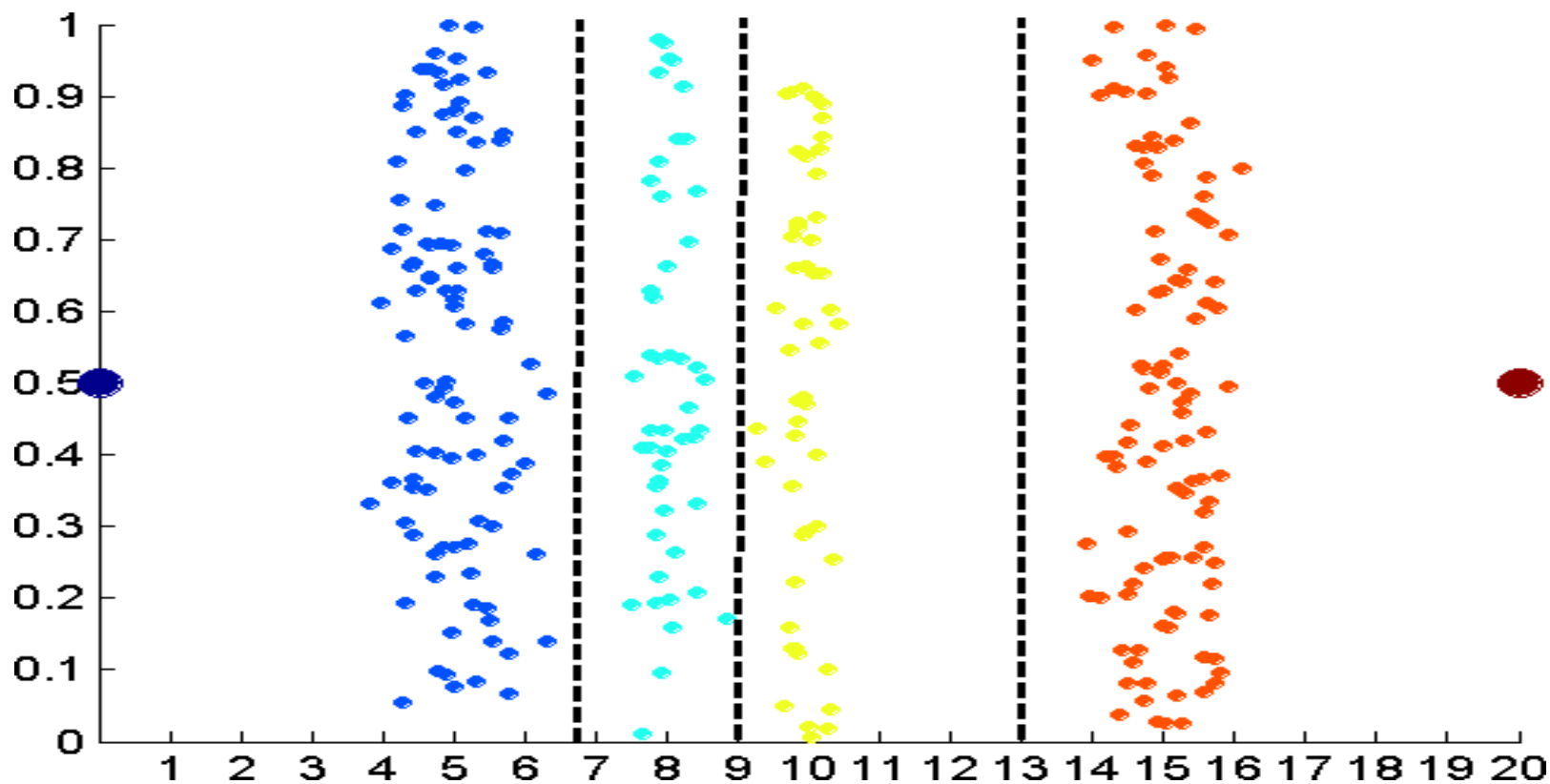
等间距划分为4类

特征离散化-unsupervised



等频繁划分为4类

特征离散化-unsupervised



使用 K-means算法划分为4类 期望最大化

特征创建 (Feature creation)

特征创建Feature Creation

- **特征创建**指的是从原始数据中人工的构建新的特征，使其能够捕获原来特征所不能够捕获的重要信息
- 这需要我们花大量的时间去研究真实的数据样本，思考问题的潜在形式和数据结构，同时能够更好地应用到预测模型中
- 特征构建需要很强的洞察力和分析能力，要求我们能够从原始数据中找出一些具有物理意义的特征

特征创建常用三种方法

- **特征提取**

- 特征提取算子。高度针对具体领域 (domain-specific)(如： 拉链缺陷检测、 指纹识别)
- 提取统计特征， 例如： 数据的均值、 方差等中心特征

- **将数据投影到一个新的空间**

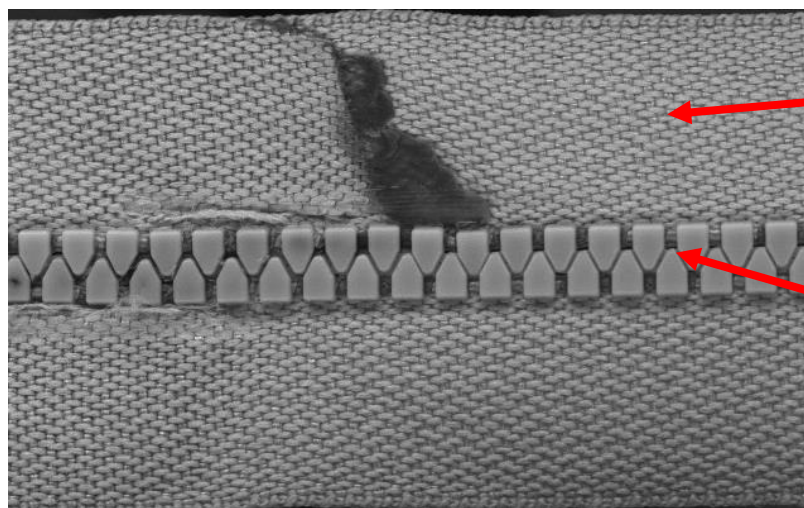
- 例如： 傅立叶变换、 小波变换

- **构造新的特征**

- 从多个原特征构造新的特征 （ combining features ）
- 例如： 质量除以体积获得密度

特征提取 (Feature Extraction)

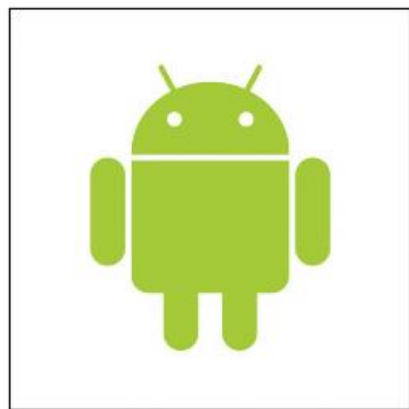
- 特征提取的对象是原始数据，它的目的是自动地构建新的特征，将原始特征转换为一组具有明显物理意义 (Gabor、几何特征[角点、不变量]、纹理[LBP, HOG]) 或者统计意义或核的特征
- 对于图像数据，可能还包括了线或边缘检测
- 高度针对具体领域 (domain-specific) (如：拉链缺陷检测、指纹识别)



灰度共生矩阵 + 局部二值模式(LBP) 提取纹理特征

方向梯度直方图(HOG)提取边缘特征

特征提取 (Feature Extraction)

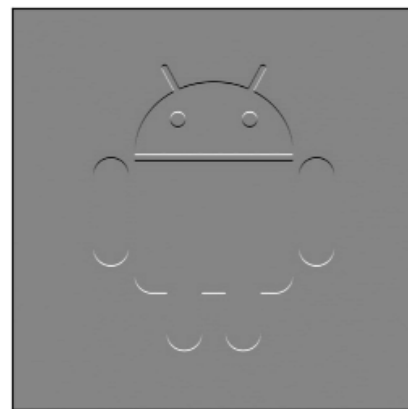


原始图像

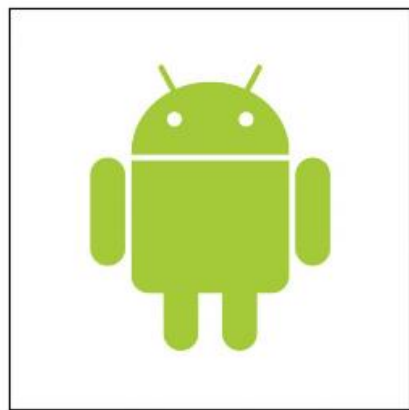


-1	-2	-1
0	0	0
1	2	1

水平 Sobel 滤波



提取水平方向边缘特征



原始图像



-1	0	1
-2	0	2
-1	0	1

垂直 Sobel 滤波



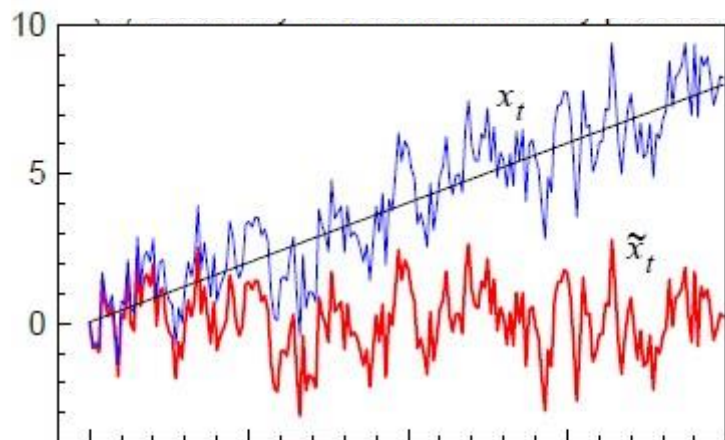
提取垂直方向边缘特征

Sobel滤波器是边缘检测器。水平Sobel滤波检测水平边缘，而垂直Sobel滤波检测垂直边缘。输出图像中的亮像素（值较高的像素）表明原始图像周围有很强的边缘。

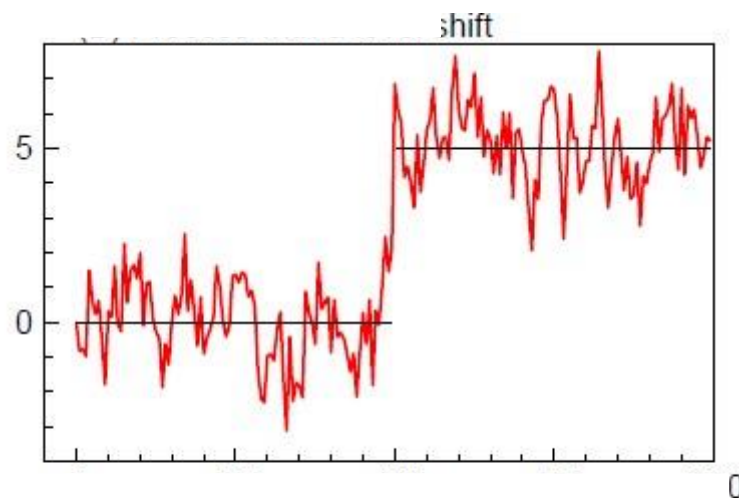
提取统计特征

- 使用统计特征来表征全局特征

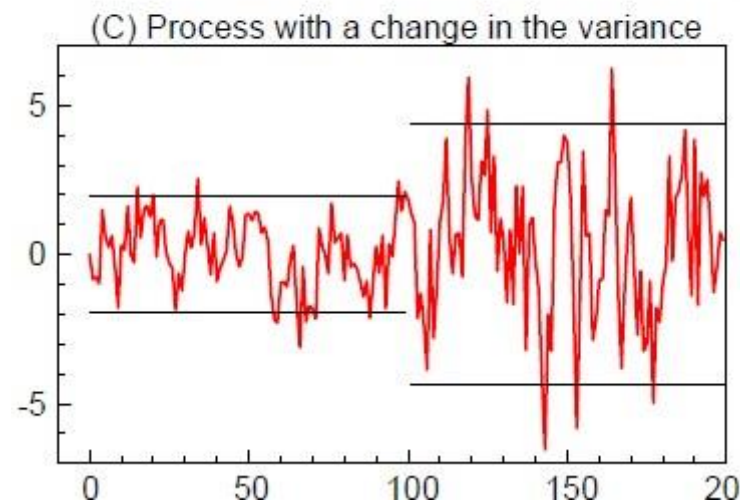
斜率



均值



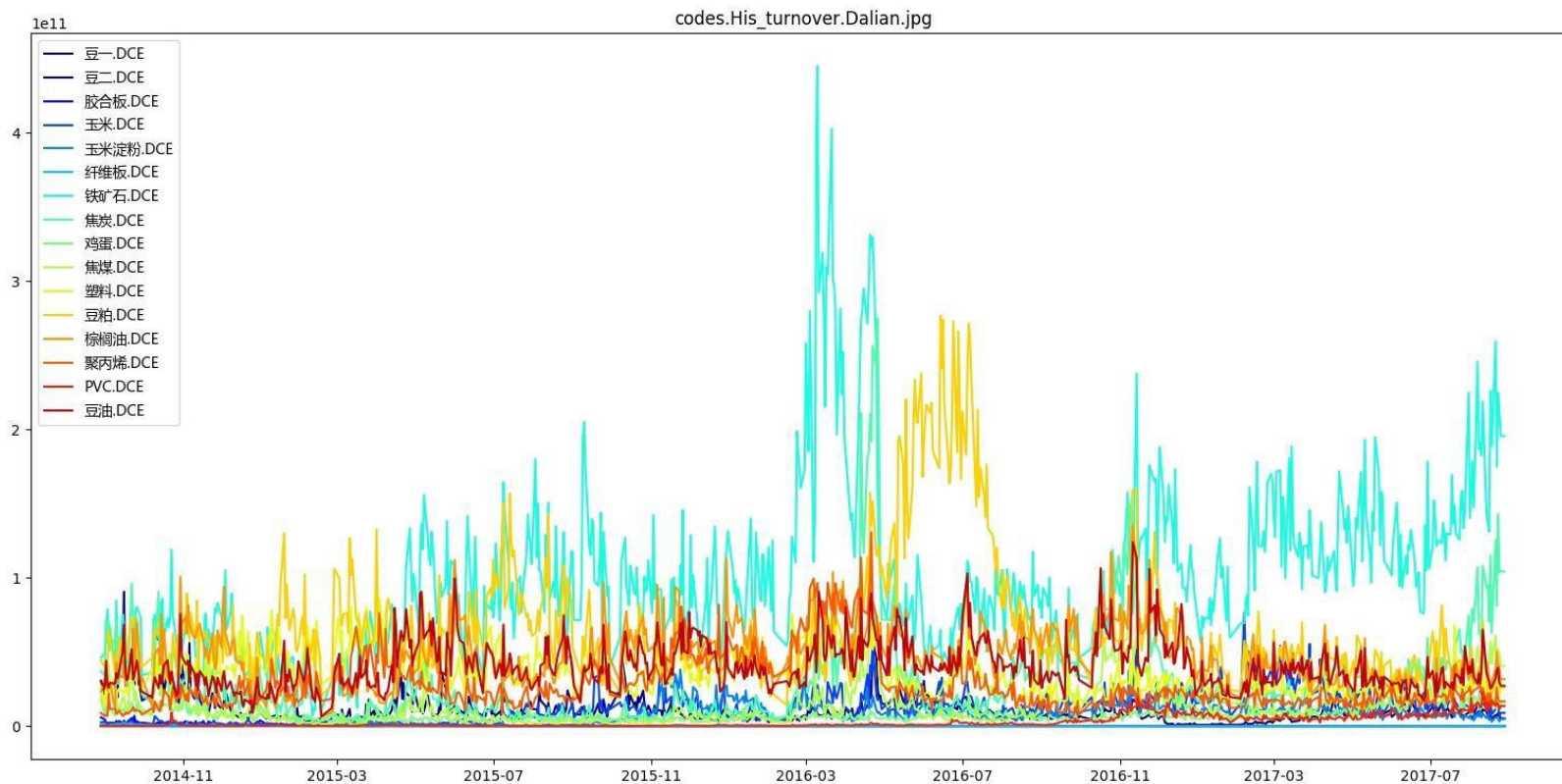
波动



提取统计特征

- **特征组合**是将两个或更多的类别属性组合成一个。当组合的特征要比单个特征更好时，这是一项非常有用的技术
- **数据聚集的目的**
 - 减少数据规模
 - 调整数据的尺度
 - 将地块数据聚集成为城区、城市、地区、省等
 - 突出数据的统计特性
 - 聚集后的对象具有更明显的统计特性，例如男女比例统计

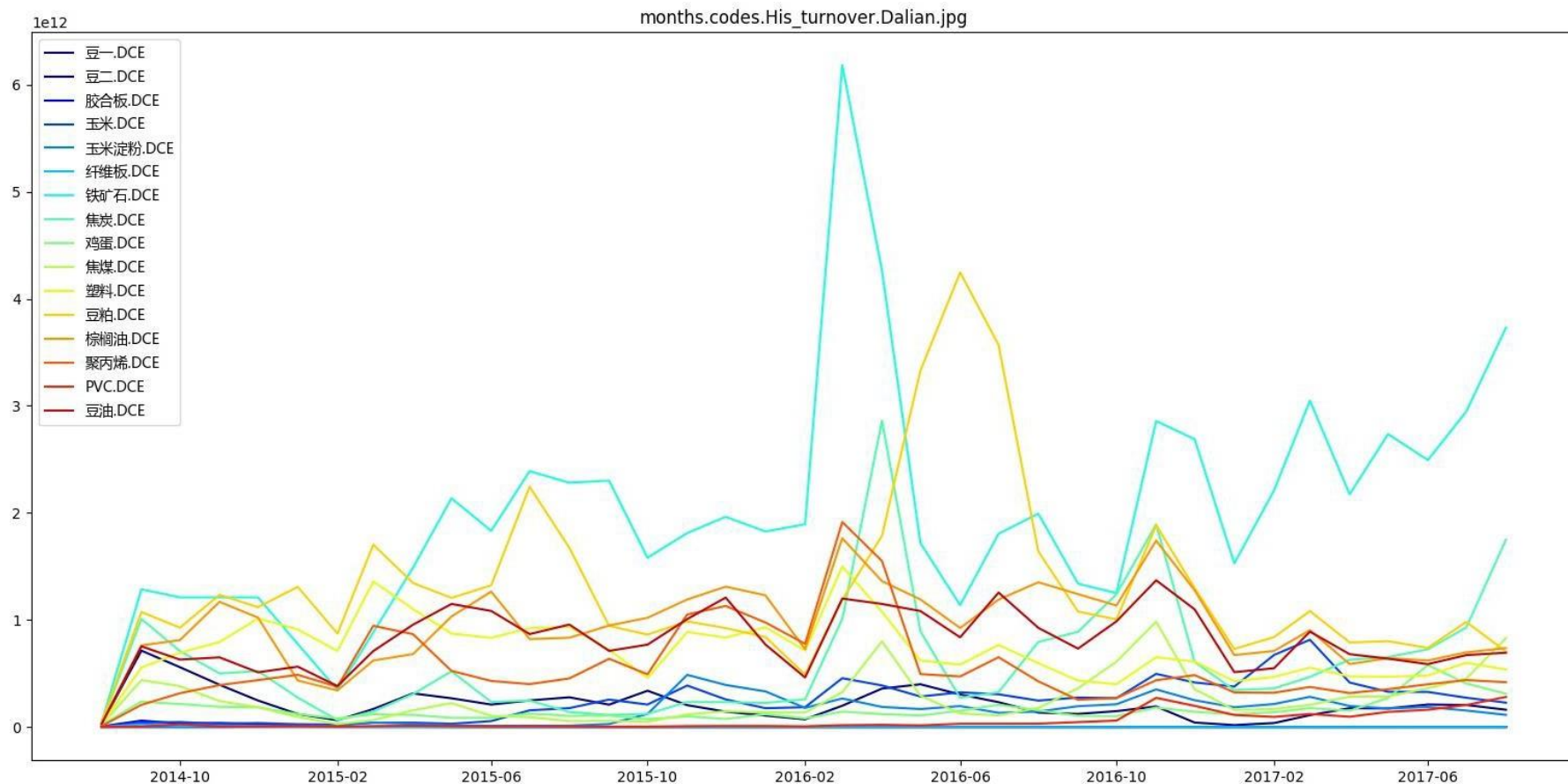
提取统计特征



大连交易所三年不同期货成交额变化（单位：天）

大连交易所铁矿石、豆粕占比较大

提取统计特征



大连交易所三年不同期货成交额变化（单位：月）

大连交易所铁矿石、豆粕占比较大

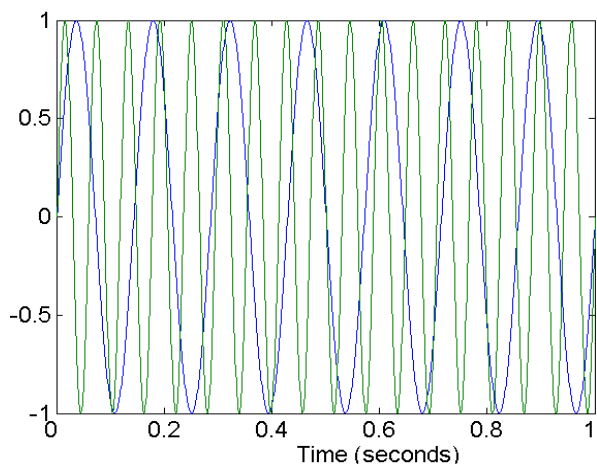
特征变换

- 特征变换是指使用 **(非线性)** 函数将特征投影到一个新的空间当中
- 常见的特征变换有基于多项式的、基于指数函数的、基于对数函数的等
 - Simple functions: x^k , $\log(x)$, e^x , $|x|$
- 4个特征 x_1, x_2, x_3, x_4 , 的2次多项式特征变换如下

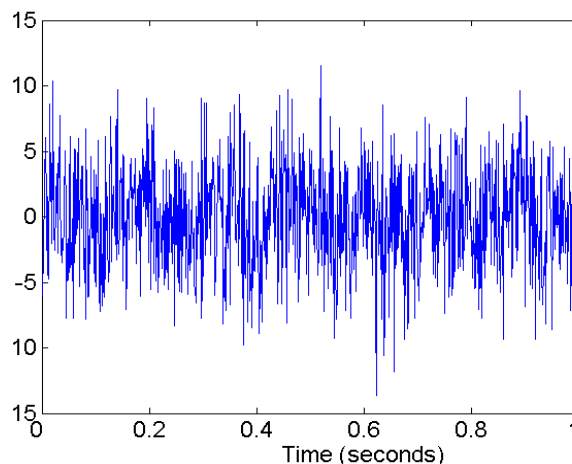
$$\begin{aligned} & (x'_1, x'_2, x'_3, x'_4, x'_5, x'_6, x'_7, x'_8, x'_9, x'_{10}, x'_{11}, x'_{12}, x'_{13}, x'_{14}, x'_{15}) \\ &= (1, x_1, x_2, x_3, x_4, x_1^2, x_1 * x_2, x_1 * x_3, x_1 * x_4, x_2^2, x_2 * x_3, x_2 * x_4, x_3^2, x_3 * x_4, x_4^2) \end{aligned}$$

特征创建：映射数据到新空间

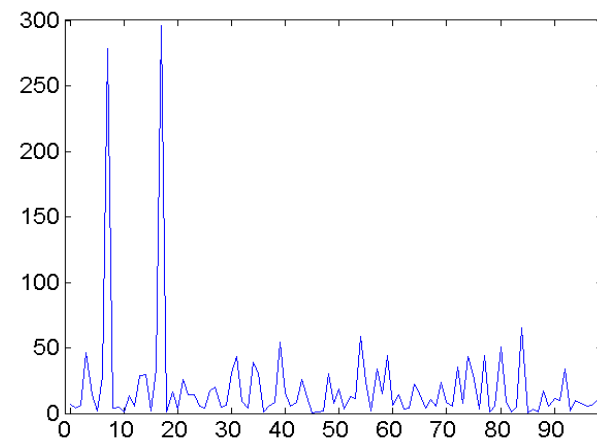
- 傅里叶变换(Fourier transform)
- 小波变换(Wavelet transform)
- 例: 傅里叶变换
 - 左: 两个sin波;
 - 中: 两个sin波之和+噪声; 检测不到模式
 - 右: 傅里叶变换到频谱; 两个尖峰对应于两个无噪声的时间序列



两个正弦波



两个正弦波+噪声



功率频谱

特征提取-降维

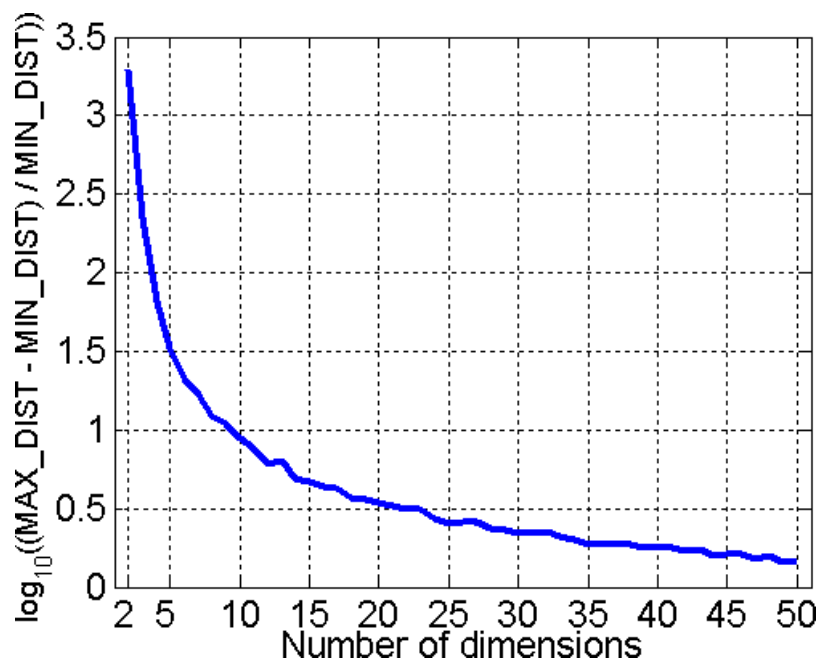
(Feature extraction- Dimensionality Reduction)

特征提取

- 特征提取是从最初的一组测量数据开始，构建旨在提供信息且非冗余的派生值（特征），以促进后续的学习和泛化过程。
- 特征提取是一个**降维**过程，将原始变量的初始集合降维至更易于管理的组别（特征）进行处理，同时仍然准确、完整地描述原始数据集
- 当一个算法的输入数据太大而不便处理，并且可能是冗余时（例如，不同单位下的相同测量），则可以将其转换为一组简化的特征（也称为特征向量）。所选特征将包含来自输入数据的相关信息，因此可以使用这种简化的表示来执行所需的任务，而不是使用完整的初始数据。

降维 Dimensionality Reduction

- 维度灾难
 - 当数据的维度增加时，数据将会变得特别**稀疏**
- **数据的高度稀疏会使得“距离”、“密度”、“聚类”、“异常检测”都变的无意义**



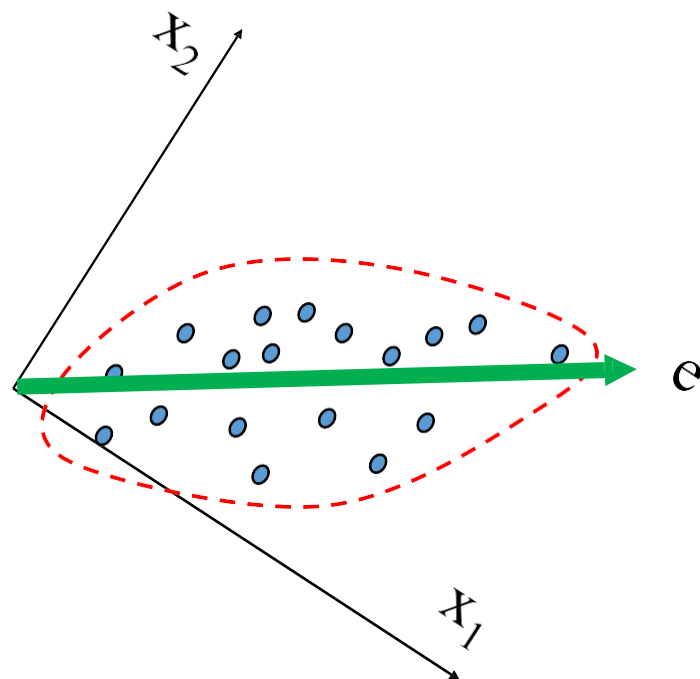
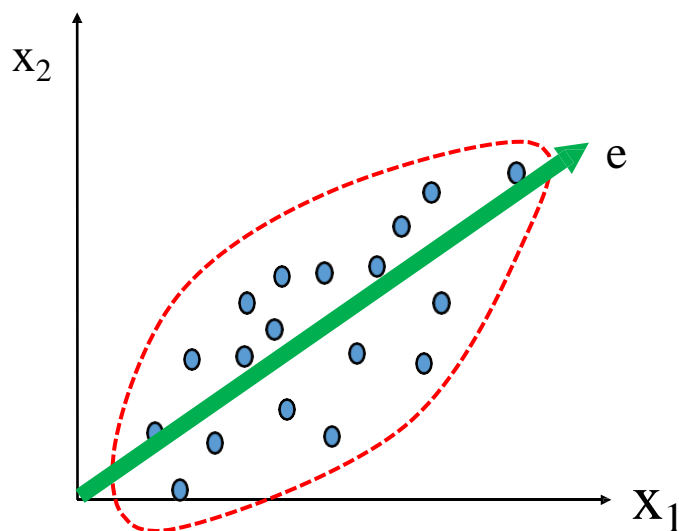
- Randomly generate 500 samples
- Compute difference between max and min distance between any pair of samples

降维 Dimensionality Reduction

- 降维的目的
 - 避免维度灾难
 - 减少数据挖掘所需要的时间和内存要求
 - 使得数据更加方便的被可视化
 - 排除数据中不重要的数据特征以及降低数据噪声
- 降维技术
 - 无监督：主成分分析 (PCA)
 - 有监督：线性判别分析 (LDA)
 - 其它有监督的和非线性技术

主成分分析

- 主成分分析(Principal components analysis, PCA)是迄今为止最流行的降维方法, 在数据压缩、消除冗余和数据噪音消除等方面有广泛的应用
- 减少数据集的维度, 同时保持数据集中**对方差贡献最大的特征**。



参考《机器学习》周志华著, P₂₂₉

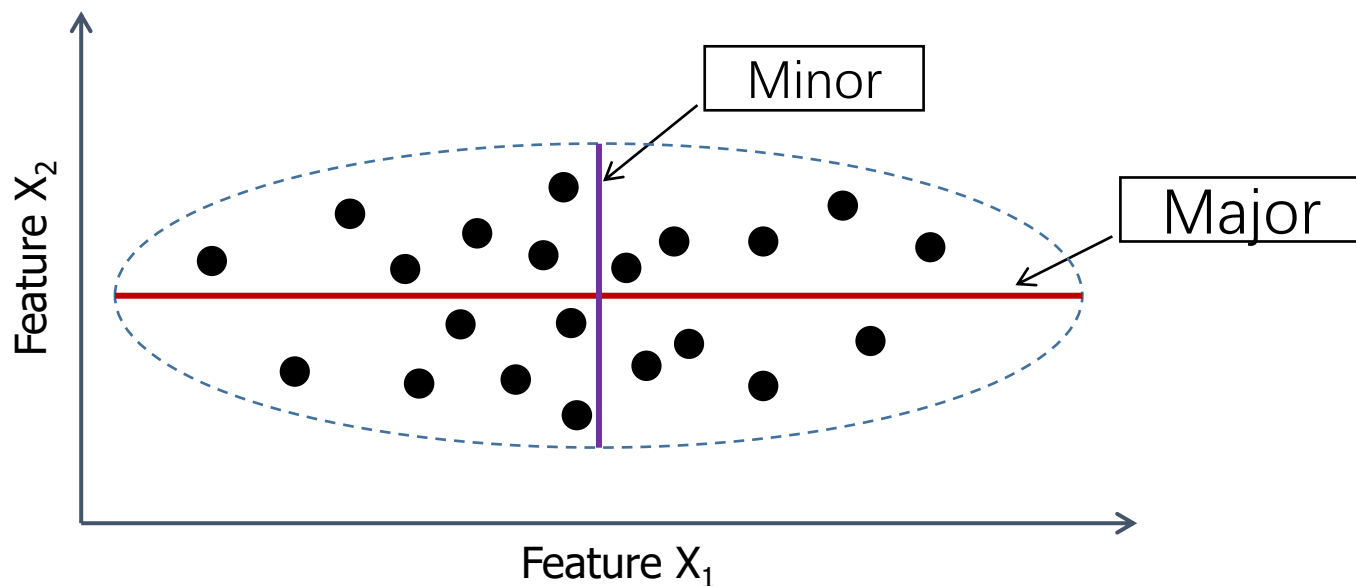
参考《统计学习方法》李航著, 2019, 第2版, P₂₉₇

参考《数据挖掘导论-附录》陈封能等著, 第2版, 2019

PCA和LDA的细节推荐参考课程群中补充材料和《统计学习方法》李航

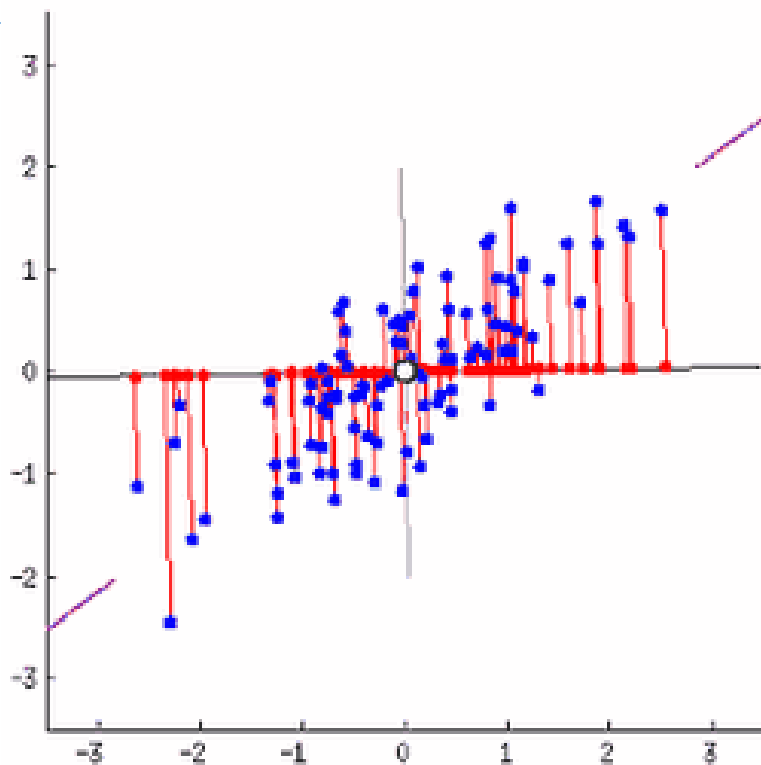
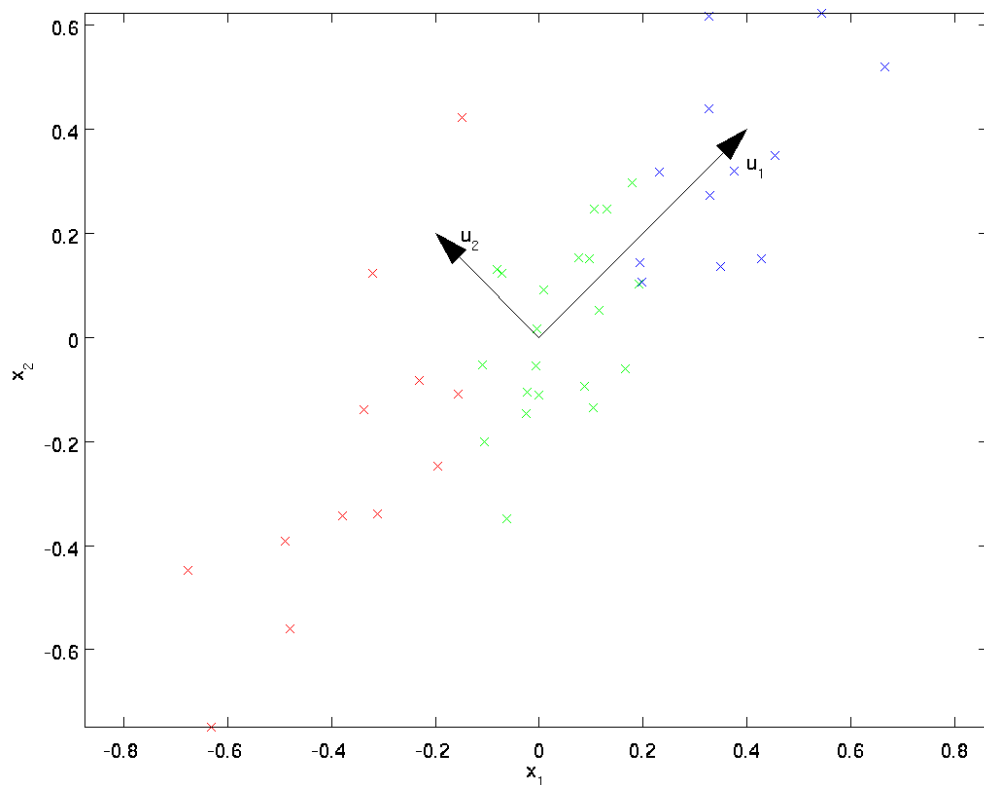
2D 例子

- 数据(Data): 高斯分布
- 属性的方差(Variance of the attribute): **信息** (Information)
- 椭圆(Ellipse): 长轴 vs 短轴 (Major Axis vs. Minor Axis)
- 选择和长轴(major)对应的属性



主成分分析 (PCA)

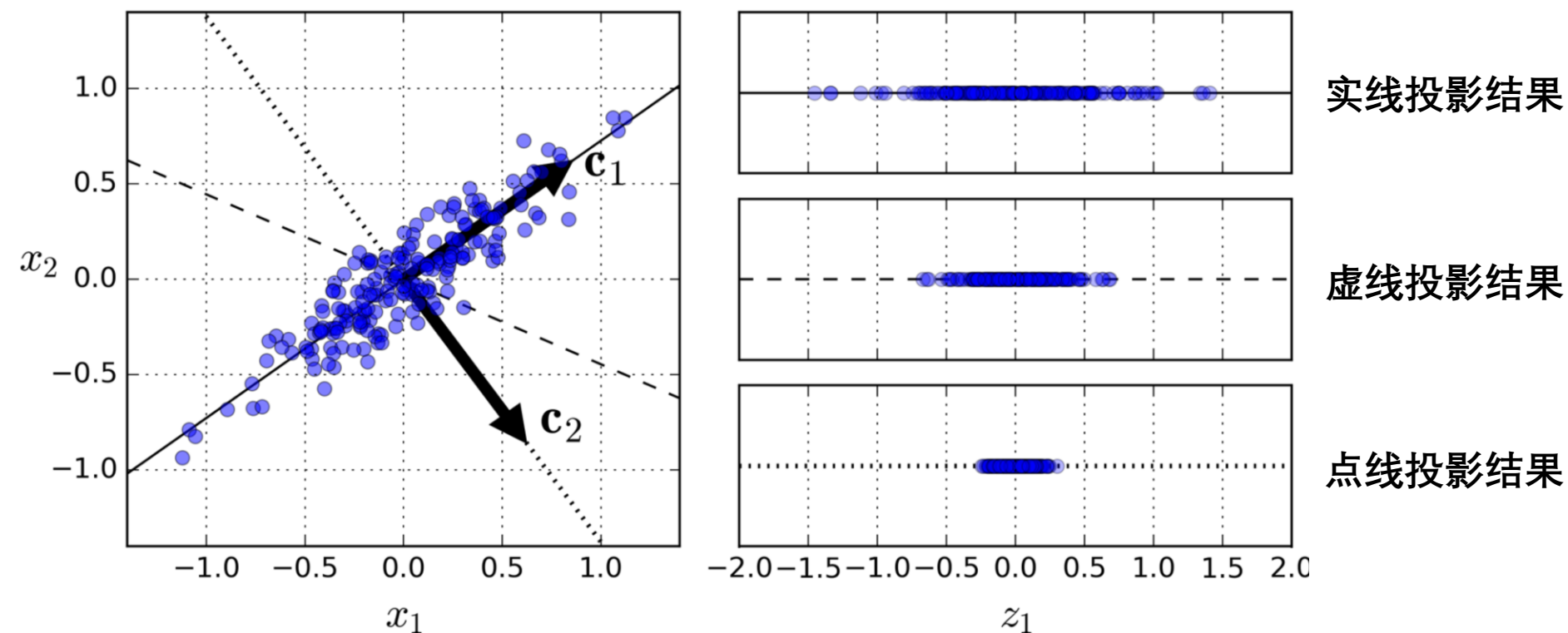
- PCA就是找出数据中最主要的方面，用数据中最主要的方面来代替原始数据。假如我们的数据集是 n 维的，共有 m 个数据 $\{x_1, x_2, \dots, x_m\}$ ，我们将这 m 个数据从 n 维降到 d 维，希望这 m 个 d 维的数据集尽可能的代表原始数据集。
- 我们知道从 n 维降到 d 维肯定会有损失，但是希望损失尽可能的小，那么如何让这 d 维的数据尽可能表示原来的数据呢？
- 首先来看最简单的情况，即将二维数据降到一维，也就是 $n=2, d=1$ 。数据如下图所示，我们希望找到某一个维度方向，它可以代表这两个维度的数据。图中列了两个向量，也就是 u_1 和 u_2 ，那么哪个向量可以更好的代表原始数据集呢？



直观上看 \mathbf{u}_1 比 \mathbf{u}_2 更好，为什么呢？可以有两种解释，第一种解释是样本点在这个直线上的投影尽可能的分开(方差大)，第二种解释是样本点到这个直线的距离足够近。假如我们把 d 从1维推广到任意维，则我们希望降维的标准为样本点在这个超平面上的投影尽可能分开，或者说样本点到这个超平面的距离足够近。

投影后数据差异最大的超平面-例子

- 左图为简单的2D数据集, 沿3个不同的轴(即一维超平面,实线、虚线、点线)进行投影, 右图为将数据集投影到每条轴上的结果
- 在实线上的投影保留了最大的差异性, 在虚线上的投影差异性居中, 而在点线上的投影只保留了非常小的差异性

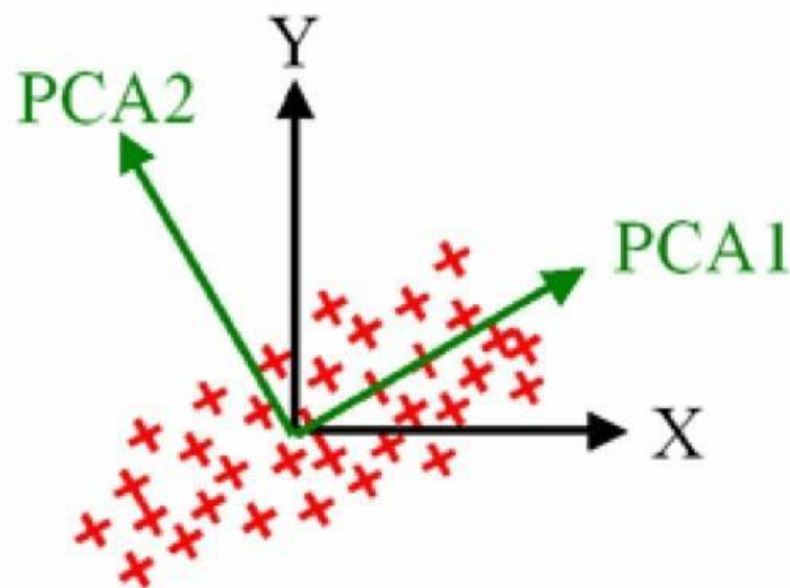
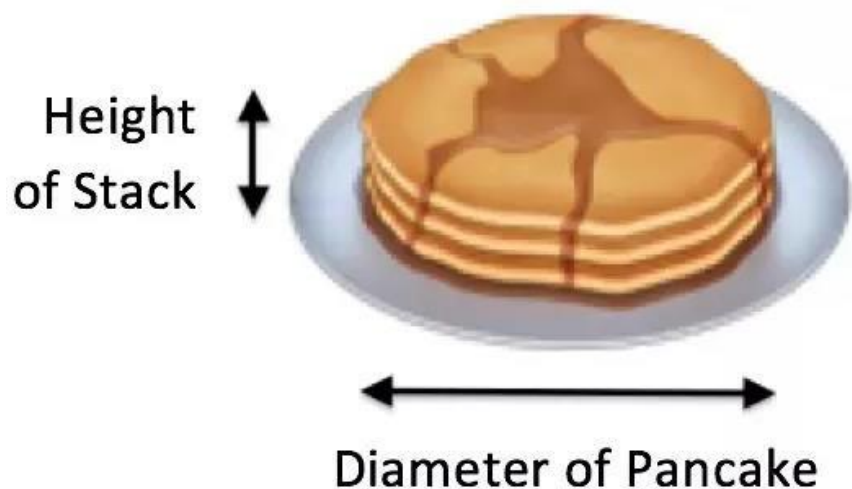


投影后数据差异最大的超平面

- 选择投影后保留最大差异性的轴看起来比较合理, 因为它可能比其它两种投影丢失的信息更少
- 要证明这一选择, 可以比较原始数据集与其他轴上的投影之间的均方误差距离, 使这个均方距离最小的轴是最合理的选择, 也就是实现代表的轴. 这正是PCA背后的简单思想
- 主成分分析可以在训练集中识别出哪条轴对差异性贡献度最高. 同时也找出了第二条轴, 它对剩余差异性贡献度最高, 与第一条垂直. 因为这是二维的, 故除了点线再没有其他
- 高维数据集中, PCA还会找到与前两条都正交的第三条轴, 以及第四条, 第五条, 等等--轴的数量与数据集的维度相同

主成分分析

- PCA的特点
 - 主成分对应的基为数据方差贡献最大方向
 - 成分的重要性逐步递减
 - 成分之间相互**正交**



特征选择

(Feature selection)

特征子集选择

- 特征选择，又称变量选择、属性选择或变量子集选择，是选择相关特征（变量、预测器）子集用于模型构造的过程
- 简单地说：检测相关特征，摒弃冗余特征，获得特征子集，从而以最小的性能损失更好地描述给出的问题
- 冗余特征
 - 特征之间存在部分或全部的特征冗余
 - 例如：“产品价格”与“产品的消费税”
- 无关特征
 - 包含了和数据挖掘任务无关的特征
 - 例如：学生的学号通常和预测学生GPA没有太明显的关系

特征工程、特征提取、特征选择三者之间的特点

- **特征工程**：从已有数据中创建新的特征；注入领域知识。
 - **特征提取**：将原始数据转换为特征，会创建新特征；降维过程；
 - **特征选择**：选择特征子集；不创建新特征
-
- 特征工程和特征提取都是将原始数据转换为适合建模的特征，有些情况下可以互换概念，但**特征提取更注重数据降维**。
 - 特征选择不创建新特征，注重删除无用特征。有时候还会看到特征转换,它属于数据转换，目的是提高算法的精度。
 - 特征提取主要用在图像、信号处理和信息检索领域，在这些领域，模型精确度比模型可解释性要重要；
 - 特征选择主要用于数据挖掘，如文本挖掘，基因分析和传感器数据处理

特征提取和特征工程（从中提取出信息）的例子

- 文本（ngrams、word2vec、tf-idf）
- 图像（CNN）
- 地理空间数据（经纬度）
- 日期和时间（日、月、周、年）
- 时间序列、网络
- 数据降维

特征选择（基于选择的特征构建模型）

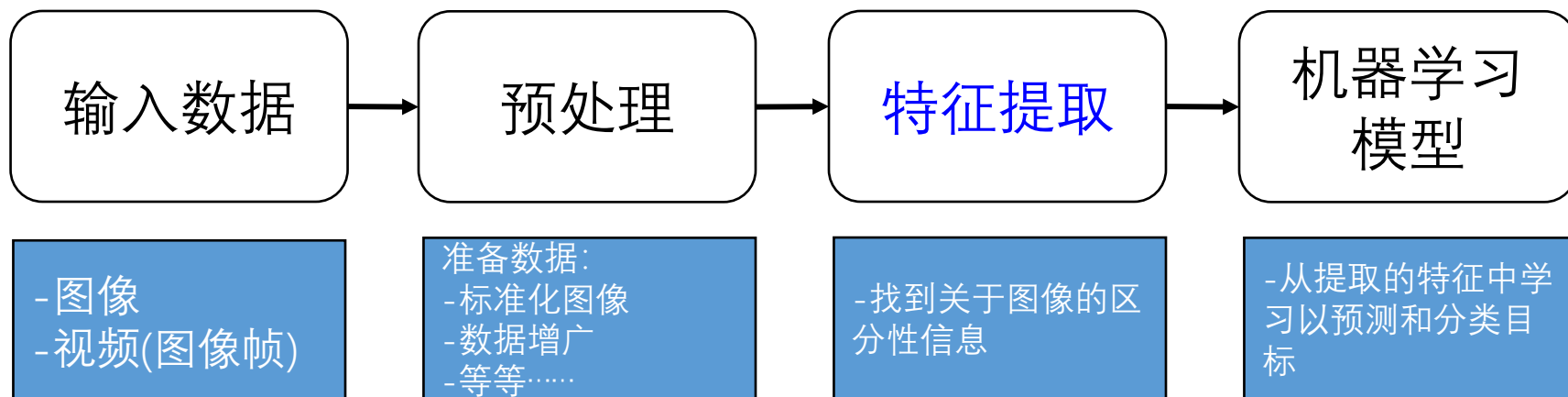
- 统计方法
- 交叉验证
- 特征加权算法（ReliefF）

特征转换（转换为有意义的）

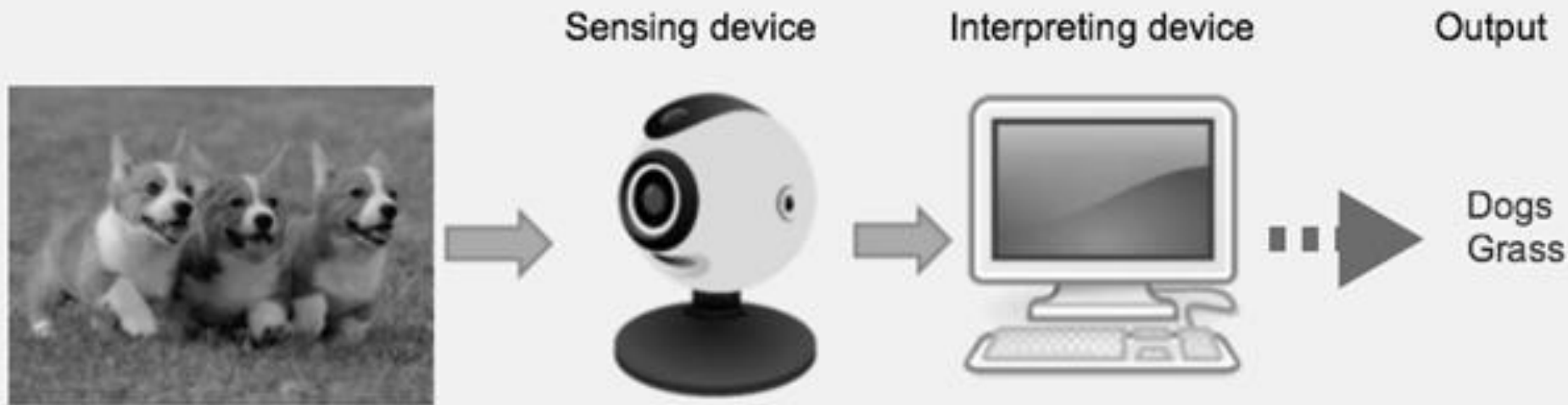
- 规范化和改变分布（例如缩放）
- 交互作用
- 填写缺失值（中间值填充等）

机器学习在计算机视觉 中的特征表示

计算机视觉检测流程



Computer Vision System



为什么使用图像特征

机器学习/深度学习的面临的语义鸿沟(Semantic Gap)



29	142	142	75	22	109	111	28	6	5
137	168	41	206	100	70	219	127	114	191
205	154	226	14	89	86	242	67	203	15
247	47	128	123	253	229	181	251	232	28
68	75	24	99	93	63	215	222	102	180
206	246	85	103	215	3	62	64	77	216
126	80	165	149	196	75	186	60	179	193
44	253	164	253	14	216	175	30	46	254
137	23	33	203	241	21	144	63	244	188
32	214	142	121	249	109	99	232	183	71
45	36	152	27	190	137	61	1	237	247
1	14	241	70	2	30	151	67	169	205
32	80	102	32	99	169	91	166	73	214
186	219	9	203	209	240	40	249	119	122
177	252	38	203	119	0	217	139	139	157
154	145	49	251	150	185	235	23	230	156
157	168	223	60	247	118	5	180	16	206
102	208	195	246	140	138	54	191	139	79
17	233	85	169	166	24	49	40	160	97
84	242	247	144	203	3	19	24	198	88
67	67	185	98	123	106	168	105	127	153
37	113	214	252	203	80	146	211	7	16
142	241	66	86	214	133	146	253	189	200
67	215	174	111	189	54	144	56	59	163

人的视角：包含狗的图像

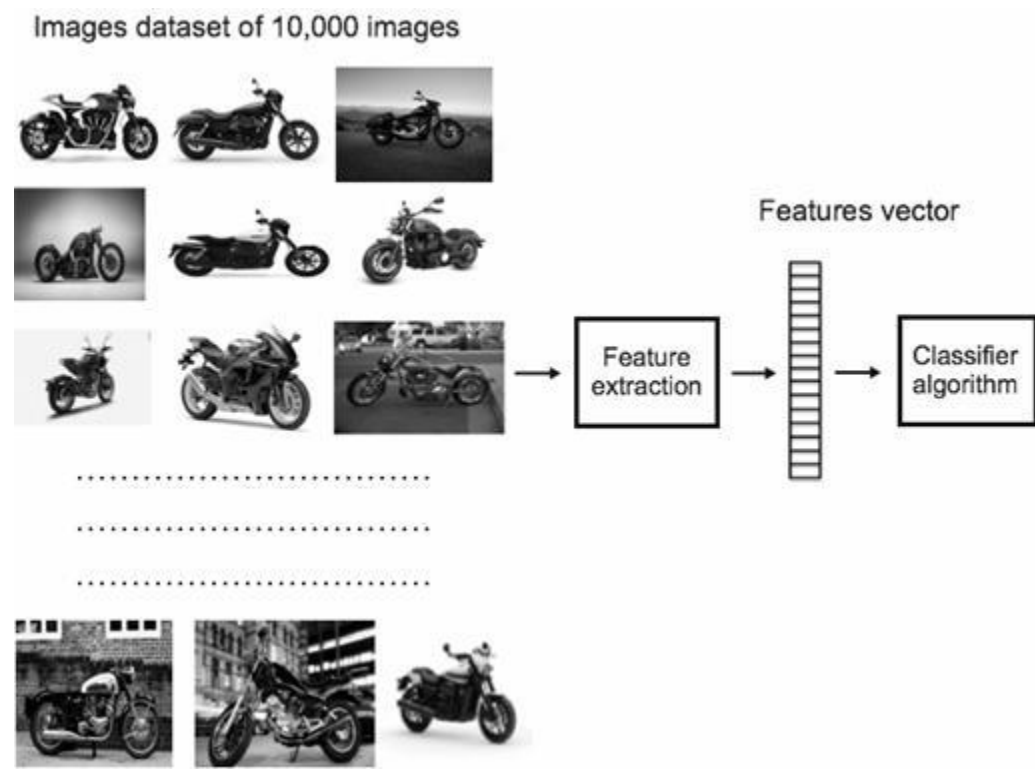
计算机视角：二维矩阵

语义鸿沟是人类如何感知图像内容与如何以计算机理解过程的方式表示图像(数字矩阵)之间的差异。

解决方法：应用特征提取来量化图像的内容

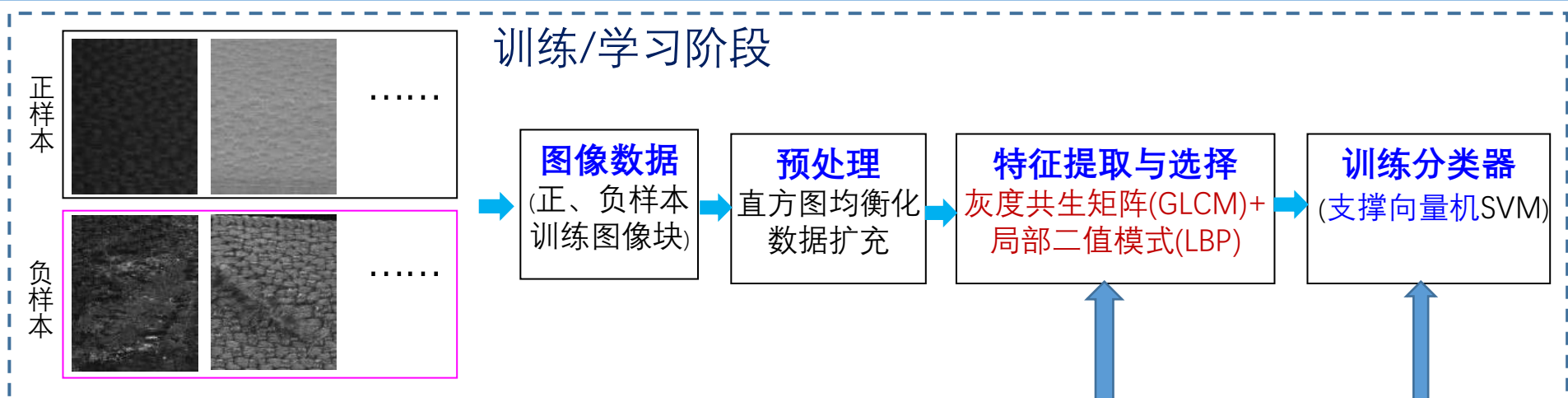
为什么使用图像特征

- 输入图像包含过多不必要的分类信息。
- 提取图像中包含的重要信息并丢弃不必要的信息来简化图像。
- 通过提取重要的颜色或图像的部分片段，我们可以将复杂的大型图像数据转换为较小的特征集。
- 这使得完成基于图像特征的图像分类的任务更加简单快捷。



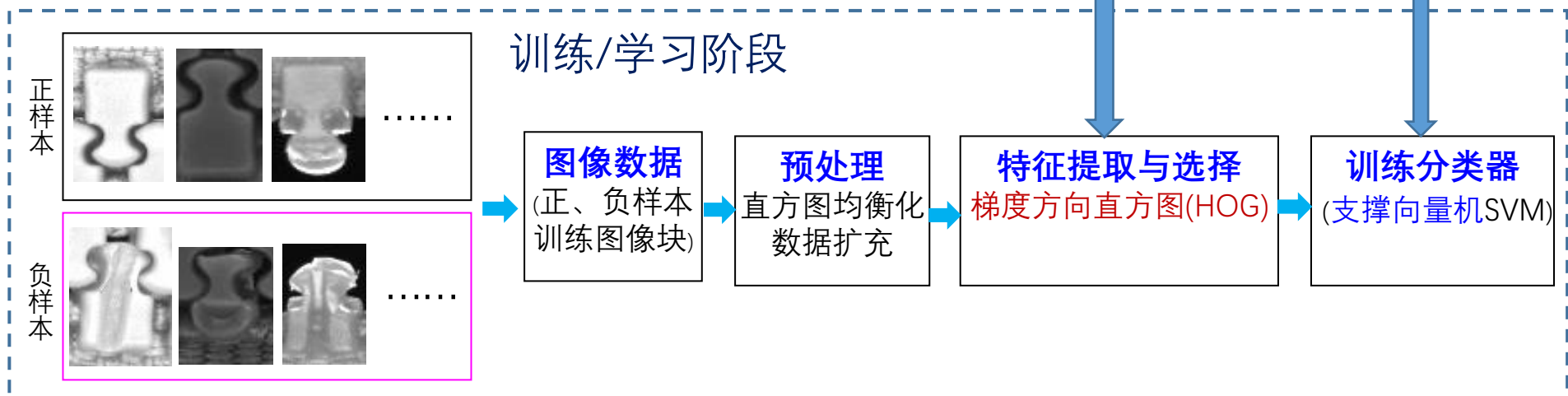
- 当将这数千张图像输入到特征提取算法时，我们将丢失所有对于摩托车识别不重要的不必要数据，并且仅保留有用特征的合并列表，然后可以将其输入分类器。
- 与让分类器查看10000张图像的数据集以了解摩托车的特性相比，此过程要简单得多。

为什么使用图像特征



拉链布带图像缺陷检测

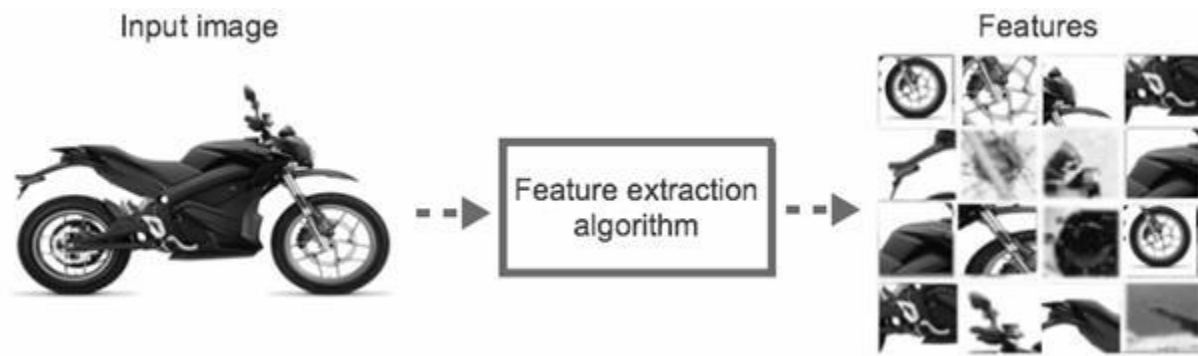
不同的特征 相同的分类器



拉链链齿图像缺陷检测

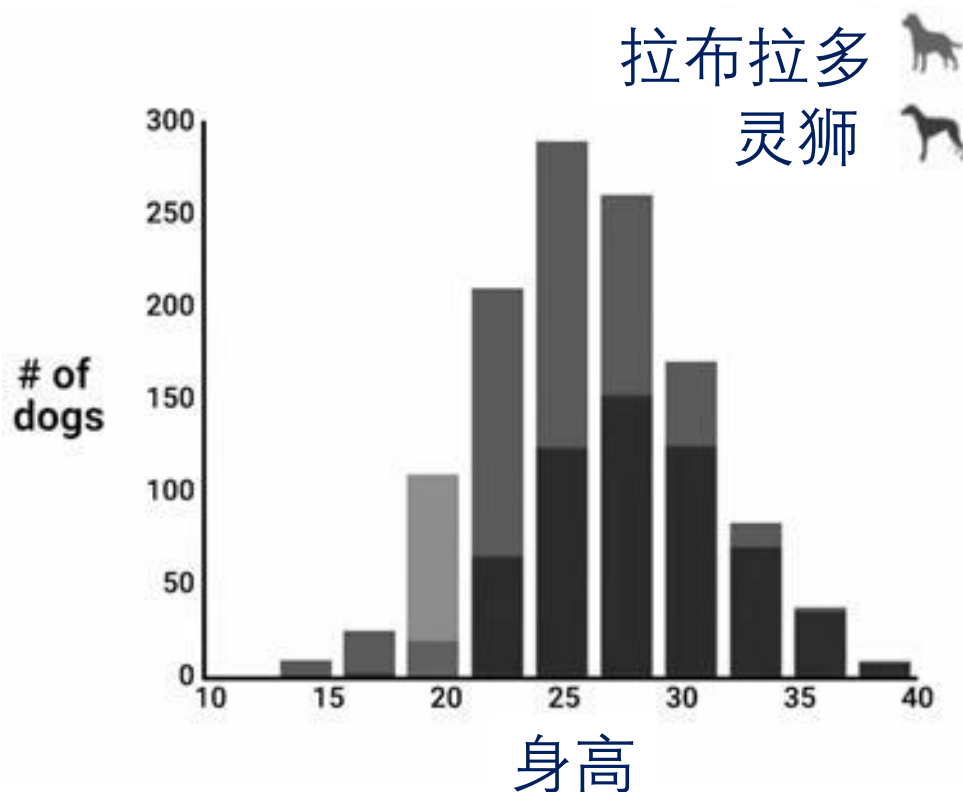
什么是好的（有用的）特征？

- 特征是图像中此特定目标所独有的可测量数据。它可能是图像中的不同颜色，也可能是特定的形状，例如线条、边缘或图像部分区域。
- 如果给一个像轮子这样的特征，问你猜这个物体是摩托车还是狗。你可能猜测是一辆摩托车。正确！在这种情况下，车轮是很强的特征，可以清楚地区分摩托车和狗。
- 如果给相同的特征（车轮），并让你猜测目标是自行车还是摩托车。在这种情况下，此特征不足以区分两个目标。我们需要寻找更多的特征，例如镜子，牌照，或者可能是一个可以共同描述物体的踏板。



什么是好的（有用的）特征？

- 建立一个分类器来区分两种狗：灵狮和拉布拉多犬
- 我们来评估两个特征：1) 狗的身高和 2) 眼睛的颜色。



灵狮
Greyhound

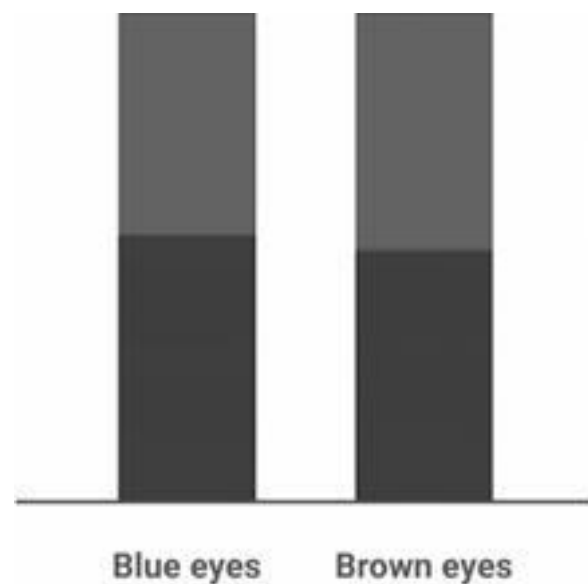


拉布拉多
Labrador



什么是好的（有用的）特征？

- 建立一个分类器来区分两种狗：灵狮和拉布拉多犬
- 我们来评估两个特征：1) 狗的身高和 2) 眼睛的颜色。



拉布拉多

灵狮



灵狮
Greyhound

拉布拉多
Labrador



对于大多数值，两种类型的分布均约为50/50
上述结果表明该特征与狗的类型无关。因此，它不会区分灵狮和拉布拉多。

什么是好的（有用的）特征？

- 良好的特征将帮助我们以各种方式识别物体。 好的特征特点：
 - (1) 可识别
 - (2) 轻松跟踪和比较
 - (3) 在不同尺度、光照条件和视角下保持一致
 - (4) 在有噪声的图像中或仅可见目标的一部分时仍然可以捕捉

Python 自主练习实践

- 阅读 Module 3 Data Exploration(tutorial3).pdf
- 动手实践 tutorial3.ipynb
- 阅读 Module 4 Data Preprocessing(tutorial4).pdf
- 动手实践 tutorial3.ipynb + tutorial4 Dataset.rar