



中科院计算培训中心

Part 1大数据挖掘及其背景

应用于大数据处理

- 杨文川



- 1) 大数据环境下数据分析
- 2) 数据挖掘定义
- 3) 数据挖掘相关技术
- 4) 大数据挖掘知识点
- 5) 模型及其评估指标



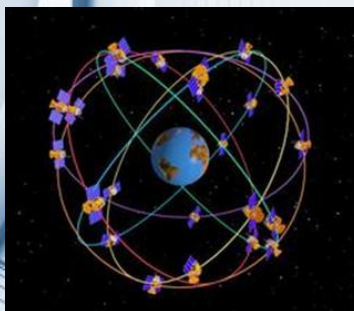
中科院计算培训中心

量化一切、利用所有的数据

大数据挖掘的基础

在数字化时代，获取数据正变得比以往任何时候都简单而不受限制

文字、方位、社交关系等都变成了数据





大数据挖掘

发现数据间的隐含信息

大数据挖掘的核心动力来源于人类了解和分析世界的渴望。

之前信息技术变革的重点在"T"（技术）上，而不是在"I"（信息）上。

现代信息系统让大数据成为了可能，人们更多的关注信息"I"本身。



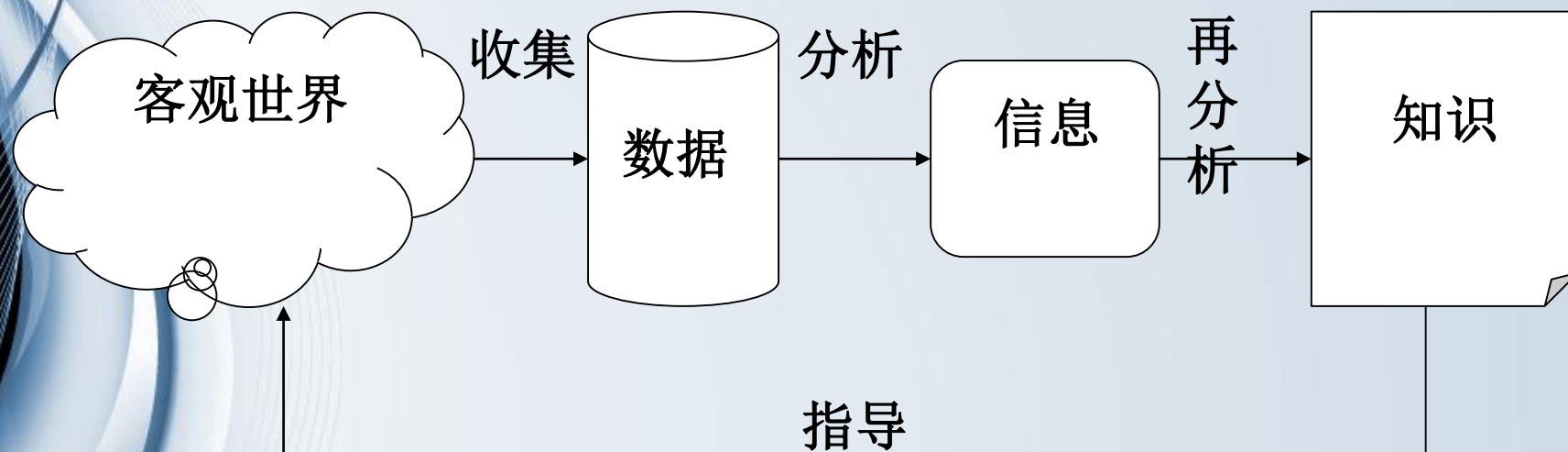
传统的数据挖掘

- 数据挖掘(Data Mining), 又称知识发现(KDD)
 - 是一个从大量数据中提取、挖掘出未知的、有价值的模式或规律等知识的复杂过程。
- 数据挖掘是一类深层次的数据分析方法。
 - 数据挖掘可以描述为: 按既定决策目标, 对大量的数据进行探索和分析, 揭示隐藏的、未知的或验证已知的规律性, 并进一步将其模型化的先进有效的方法。



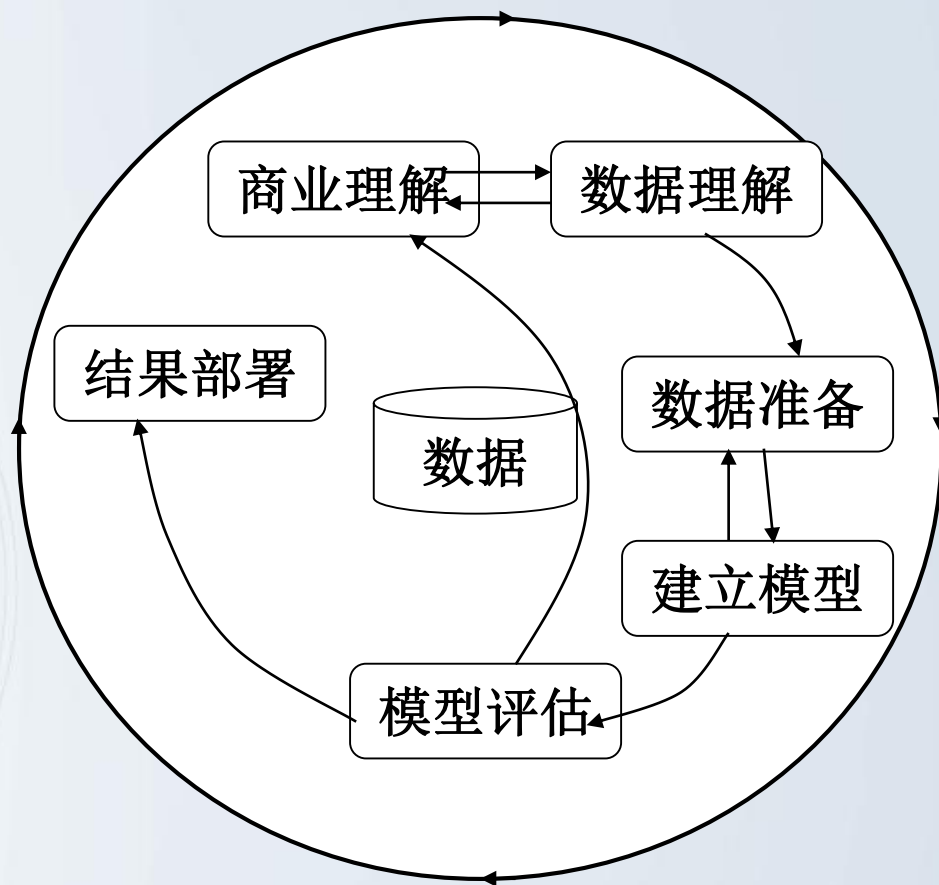
中科院计算培训中心

数据、信息与知识



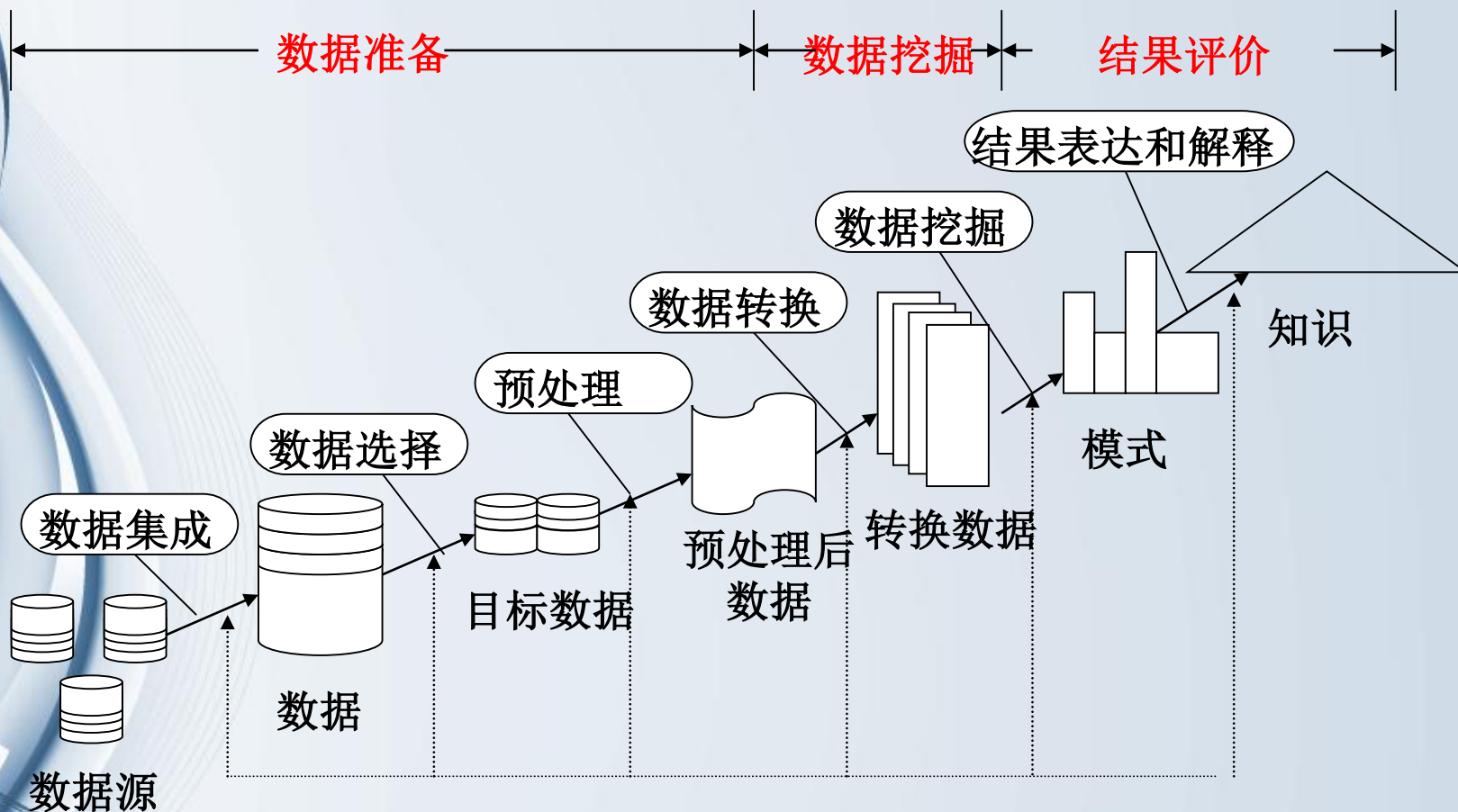


经典挖掘模型CRISP-DM





数据挖掘三阶段





常用的数据挖掘方法

关联规则
聚类分析
分类技术
时序模式
偏差检测
预测估计
.....

购物篮分析





传统的数据挖掘软件

- 专用挖掘工具、通用挖掘工具
 - QUEST
 - MineSet
 - DBMiner
 - Intelligent Miner
 - SAS Enterprise Miner
 - SPSS Modeler
 -





大数据挖掘面临的挑战

- 数据来源种类多且量大：
 - 现有的RDBMS无法处理如此巨大的数据
- 可扩展处理：
 - 挖掘计算可扩展，要反应及时
- 可靠性保证：
 - 分布式文件系统的备份恢复机制
- 并行计算模型：
 - 需要采用MapReduce的计算模型。



大数据挖掘的三个重要转变

- 首先，要分析与某事物相关的所有数据，而不是依靠分析少量的数据样本。
- 其次，接受数据的纷繁复杂，而不再追求精确性。
- 最后，不再探求难以捉摸的因果关系，转而关注事物的相关关系。



数据挖掘是数据模型的发现过程

- 数据挖掘(data mining)是数据"模型"的发现过程，而"模型"却可以有多种含义。
- 下面介绍在建模方面最重要的几个方向





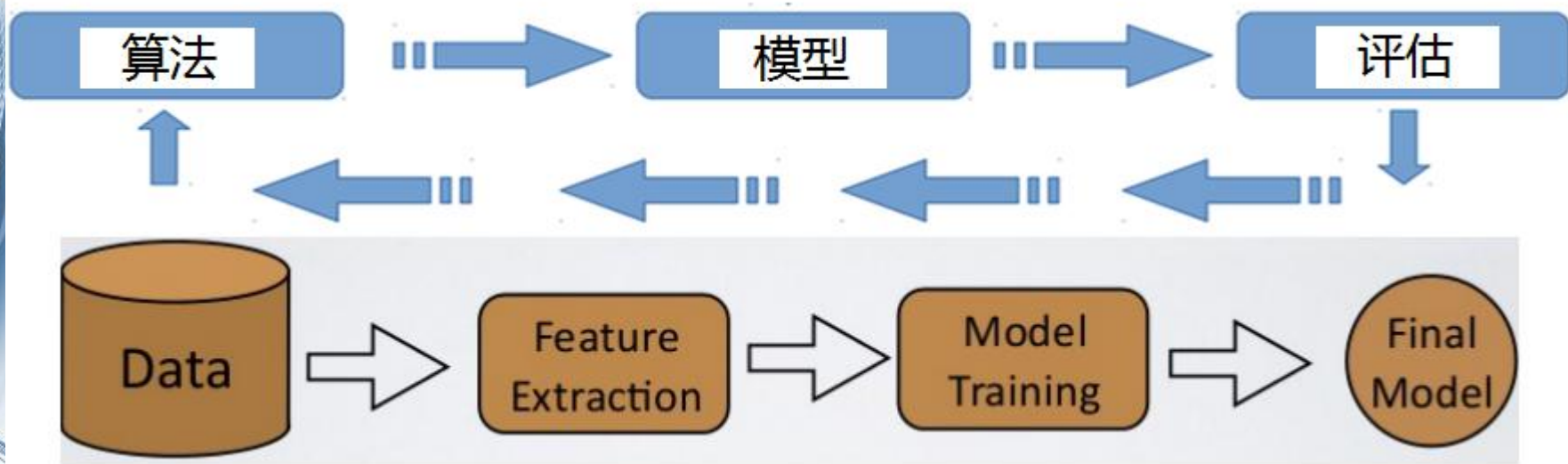
擅长的典型场景

- 数据挖掘擅长的，是当人们对数据中的寻找目标，几乎一无所知。
 - 比如，并不清楚到底是影片的什么因素，导致某些观众喜欢或者厌恶该影片。
 - 因此，在Netflix竞赛要求设计一个算法，来预测观众对影片的评分时，基于已有评分样本的数据挖掘算法获得了巨大成功。



大数据挖掘流程

- 大数据挖掘流程中最主要的三步是
– 算法设计， 模型获取， 评估效果





数据建模两种做法

- 数据建模方法，可描述为下列两种做法之一：
 - 1)对数据进行简洁的近似汇总描述；
 - 2)从数据中抽取出最突出的特征，代替数据，并忽略剩余内容



数据汇总

- 一种数据汇总形式是PageRank，谷歌成功的关键算法
 - Web的整个复杂结构，可由每个页面所对应的一个数字(PageRank值)归纳而成。
- 另一种数据汇总形式是聚类
 - 在聚类中，数据被看成是多维空间下的点，空间中相互邻近的点将被赋予相同的类别。
 - 这些类别的概括信息综合在一起，形成了全体数据集合的数据汇总结果。



特征抽取

- 基于特征的模型，会从数据中寻找某个现象的最极端样例，并用其表示数据。
- 大数据下的一些重要的特征抽取类型，包括：
 - 1) 频繁项集(frequent itemset)
 - 2) 相似项(similar item)



1) 频繁项集

- 该模型适用于多个项集组成的数据，其原始应用发生在真实的购物篮场景下：
 - 在超市结账的时候，某些物品会被顾客同时购买，例如热狗和芥末，这些物品组成了项集
 - 寻找那些在很多购物篮中，同时出现的项集（频繁项集），这就是要找的，用以刻画数据的特征。



2) 相似项

- 有时数据看上去像一系列集合，这时的目标是，寻找那些共同元素比例较高的集合对。
 - 由于顾客大都对许多不同的商品感兴趣，寻找兴趣相似的那部分顾客，并根据这些关联对数据进行表示的做法会更有用。
 - 为向顾客推荐感兴趣的商品，Amazon先寻找与他相似的顾客群，并把其中大部分人购买过的商品也推荐给他，该过程称为协同过滤



大数据挖掘知识点

- 对数据挖掘研究有益的一些知识
 - (1)用于度量词语重要性的TF.IDF指标
 - (2)哈希函数及其使用
 - (3)二级存储器(磁盘)及其对算法运行时间的影响；
 - (4)自然对数的底 e 及包含它的一系列恒等式
 - (5)幂定律(power law)



TF.IDF

- 假定文档集中有 N 篇文档， f_{ij} 为词项 i 在文档 j 中出现的频率(即次数)，词项 i 在文档 j 中的词项频率 TF_{ij} 定义为

$$TF_{ij} = \frac{f_{ij}}{\max_k f_{kj}}$$

- 假定词项 i 在文档集的 n_i 篇文档中出现，那么词项 i 的IDF定义

$$IDF_i = \log_2 \frac{N}{n_i}$$

- 具有最高TF.IDF得分的那些词项，通常都是刻画文档主题的最佳词项



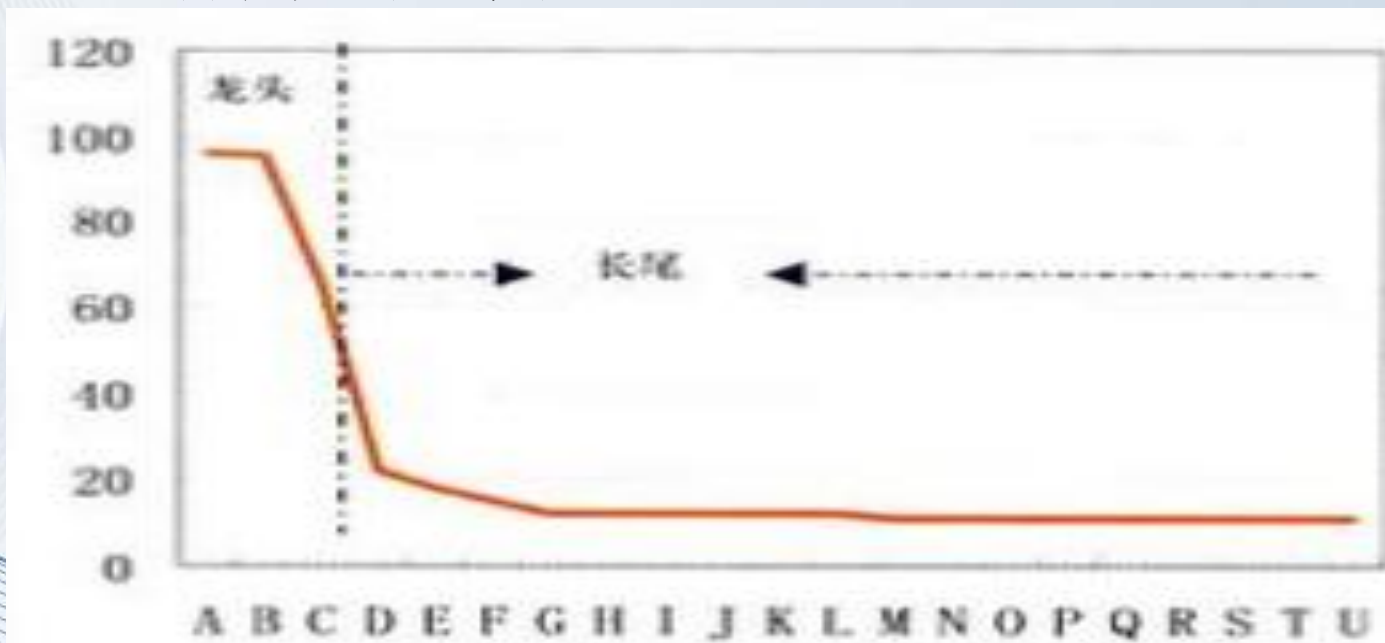
正态分布

- 假定现有的数据是一系列数字。
 - 统计学家可能会判定这些数字，来自一个高斯分布(即正态分布)，并利用公式来计算该分布最有可能的参数值。
 - 该高斯分布的均值和标准差，能够完整地刻画整个分布，因而成为上述数据的一个模型



幂律分布

- 大数据变量间常呈现幂律(power law)关系
 - 两个变量在对数空间下，呈现出线性关系
- 图示为文章用词中的幂律关系
 - 也称为长尾效应





多处数据都满足幂律

- 1) Web图当中节点的度
- 2) 商品的销量
- 3) Web网站的大小
- 4) Zipf定律



数据挖掘中的模型评估

- 传统的数据挖掘的过程，就是通过利用已知的样本数据，发现和创建模型的过程
- 模型的好坏、准确与否，与样本数据的正确与否，关系密切
- 数据挖掘领域，目前有一整套评估和分析模型
- 大数据挖掘中，直接利用海量的生产数据进行建模和分析，模型评估更加重要
- 下面将对这些方法进行介绍



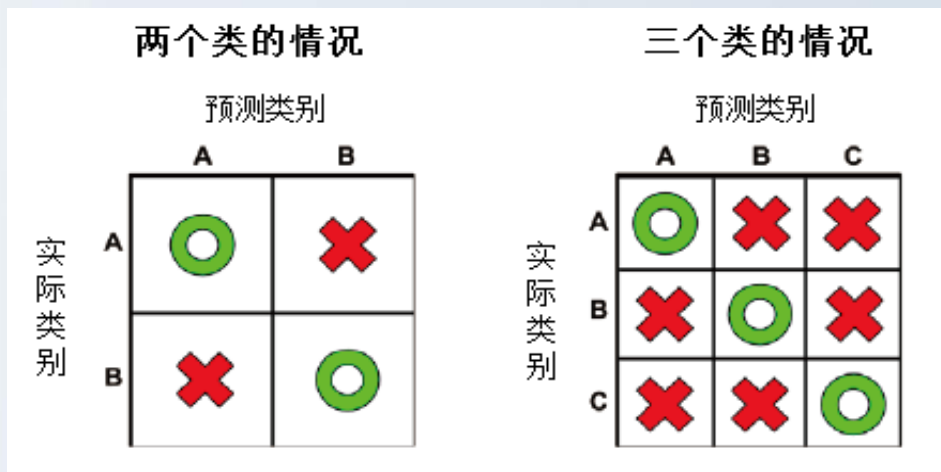
模型评估中的方法和指标

- 主要的模型评估方法和指标包括：
 - 1) 混淆矩阵
 - 2) 灵敏度与特异性
 - 3) 查准率和召回率
 - 4) ROC和AUC







1) 混淆矩阵

- 混淆矩阵(confusion matrix)是一张二维表，按预测值是否匹配数据的真实值，对预测值进行分类
 - 该表的第一个维度表示所有可能的预测类别，第二个维度表示真实的类别。
 - 下图展示了二值分类模型的混淆矩阵。对于三值分类模型，将是类似 3×3 的混淆矩阵。





表格矩阵

预测类别		no	yes
实际类别	no	 True Negative	 False Positive
	yes	 False Negative	 True Positive

- 可根据预测值是否落入下述4类中的某一个来创建这个表格矩阵：
 - 真阳性True Positive (TP): 正确的分类为感兴趣的类别。
 - 真阴性True Negative (TN): 正确的分类为不感兴趣的类别。
 - 假阳性False Positive (FP): 错误的分类为感兴趣的类别。
 - 假阴性False Negative (FN): 错误的分类为不感兴趣的类别。



2) 使用混淆矩阵度量性能

预测类别		no	yes
实际类别	no	TN True Negative	FP False Positive
	yes	FN False Negative	TP True Positive

- 使用 2×2 的混淆矩阵，可用公式来表示
- 准确度(accuracy, 有时也称为成功率):
- 准确度 = $(TP + TN) / (TP + TN + FP + FN)$

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- 因此，准确度表示真阳性和真阴性的数目，除以所有预测值的个数。



错误率

- 错误率(error rate)，或者说不正确分类的比例，定义如下：

错误率 = $(FP + FN) / (TP + TN + FP + FN) = 1 - \text{准确度}$

		预测类别	
		no	yes
实际类别	no	TN True Negative	FP False Positive
	yes	FN False Negative	TP True Positive

$$\text{error rate} = \frac{FP + FN}{TP + TN + FP + FN} = 1 - \text{accuracy}$$

- 注意，错误率可用1减去准确度来得到。直观的理解也是有道理的，如果一个模型有95%是预测正确的，那么意味着5%是预测错误的。



3) 查准率和召回率

- 与灵敏度和特异性紧密相关的，是另两个性能度量指标，预测查准率和召回率。
- 这两个统计量最开始用于信息检索领域，目的是提供对于模型结果的有趣和有关程度的描述，或者说预测是否会因为无意义的噪声而减弱。



查准率

预测类别		no	yes
实际类别	no	TN True Negative	FP False Positive
	yes	FN False Negative	TP True Positive

- 查准率(也称为阳性预测值)定义为真阳性在所有预测为阳性案例中的比例
- 查准率 = $TP / (TP + FP)$

$$\text{precision} = \frac{TP}{TP + FP}$$

- 考虑当模型不精确时，会发生结果不可信。
- 在信息检索领域，搜索引擎中，就好比Google老是返回不相关的结果。最终用户将会转向其竞争对手。



召回率

- 召回率是关于结果完备性的度量，它定义为真阳性与阳性总数的比例。
- 召回率 = $TP / (TP + FN)$

$$\text{recall} = \frac{TP}{TP + FN}$$

预测类别		no	yes
实际类别	no	TN True Negative	FP False Positive
	yes	FN False Negative	TP True Positive

- 召回率与灵敏度是一样的。
- 召回率高的模型可捕捉大量的阳性样本，这意味着其具有很宽的范围。
 - 例如，高召回率的搜索引擎，可能返回大量与搜索词相关的文档。



4) F度量

- 将查准率和召回率合并，成一个单一值的模型性能度量方式是F度量(有时也称为F1记分或者F-score)。

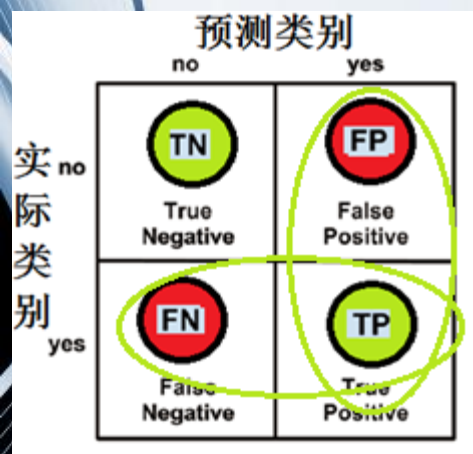
- F度量使用调和平均值，来整合查准率与召回率

- 因为预测查准率，和召回率都是0~1之间的比例，所以使用调和平均值，而不是更常用的算术平均值。

- F度量的公式：

- F度量 = $2 \times TP / (2 \times TP + FP + FN)$

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$





5) 灵敏度与特异性

- 在做决策时，分类经常要在过于保守和过于激进之间做平衡
 - 例如，一个邮件过滤器。这种权衡可由灵敏度和特异性这对度量方式实现。
- 模型的灵敏度(也称为真阳性率)度量了阳性样本被正确分类的比例。
 - 灵敏度是真阳性的数目除以数据中阳性的总数(包括正确分类的和错误分类的)
 - 灵敏度 = $TP / (TP + FN)$

预测类别		no	yes
实际类别	no	TN True Negative	FP False Positive
	yes	FN False Negative	TP True Positive

$$\text{sensitivity} = \frac{TP}{TP + FN}$$



特异性

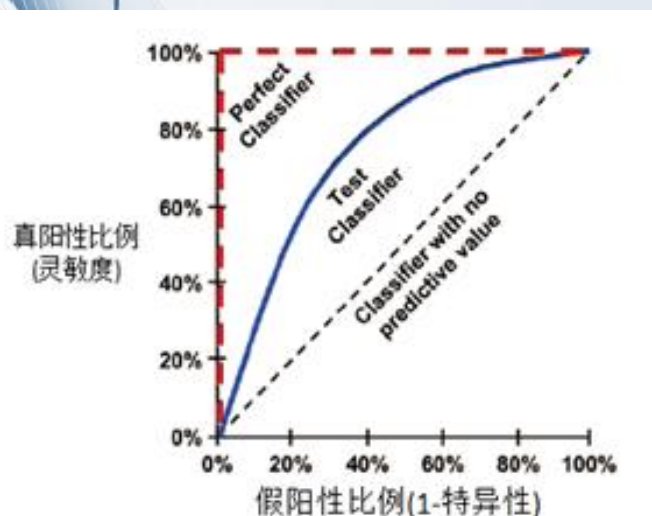
预测类别		no	yes
实际类别	no	TN True Negative	FP False Positive
	yes	FN False Negative	TP True Positive

- 模型的特异性(也称为真阴性率)度量了阴性样本被正确分类的比例。
- 特异性是真阴性的总数除以阴性的总数(包括真阴性和假阳性):
- 特异性 = $TN / (TN + FP)$
-

$$\text{specificity} = \frac{TN}{TN + FP}$$



6) ROC曲线

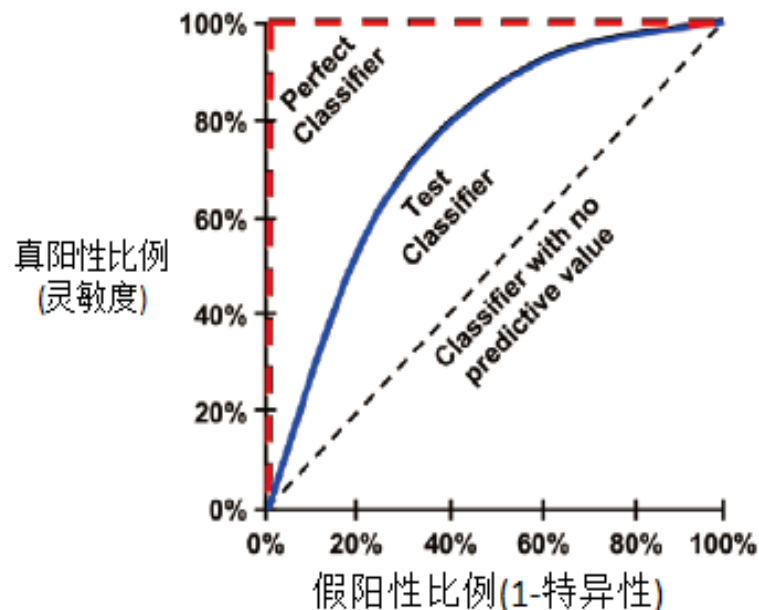
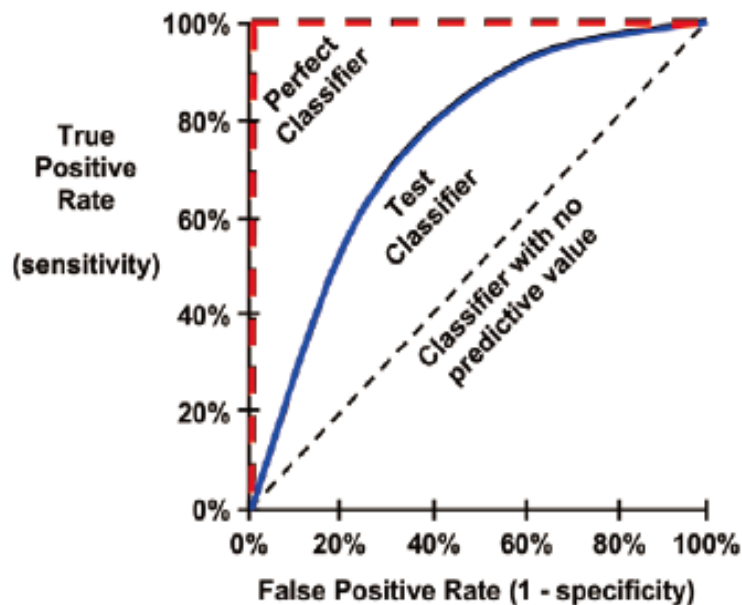


- ROC(Receiver Operating Characteristic, 受试者工作特征)曲线, 常用来检查在找出真阳性和避免假阳性之间的权衡。
 - 常用来可视化挖掘模型的功效。
- 典型ROC图形的特点在图中得到了显示。
 - 统计图形中的曲线, 纵轴表示真阳性的比例、横轴表示假阳性的比例



ROC

- ROC曲线上的点表示不同假阳性阈值上的真阳性的比例。因为这两个值分别等于灵敏度和1-特异性，所以该图形也称为灵敏度/特异性图
 - Y为真阳性比例(灵敏度)，X为假阳性比例(1-特异性)



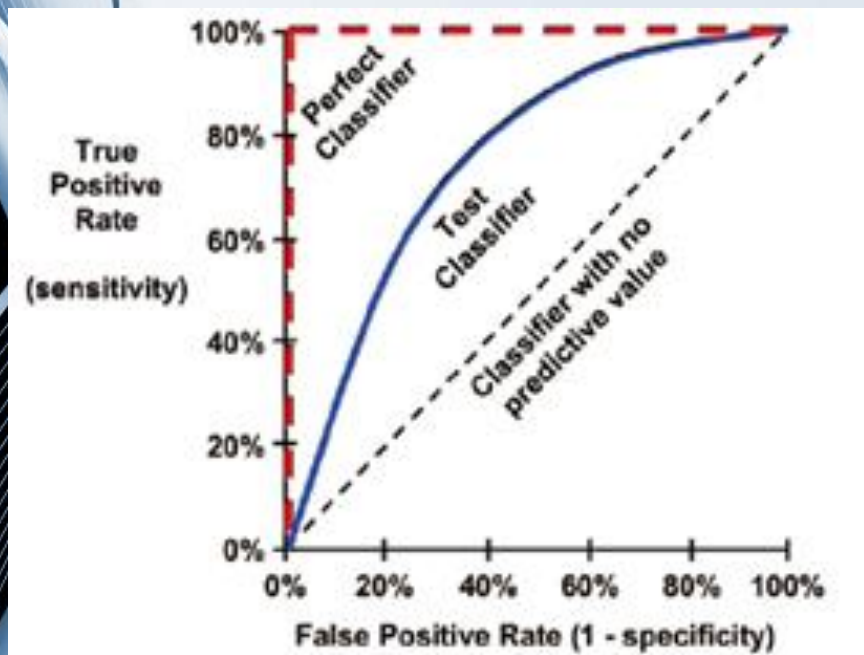


无预测价值的分类器

- 为了说明这个概念，在图中比较3个假设的分类器。

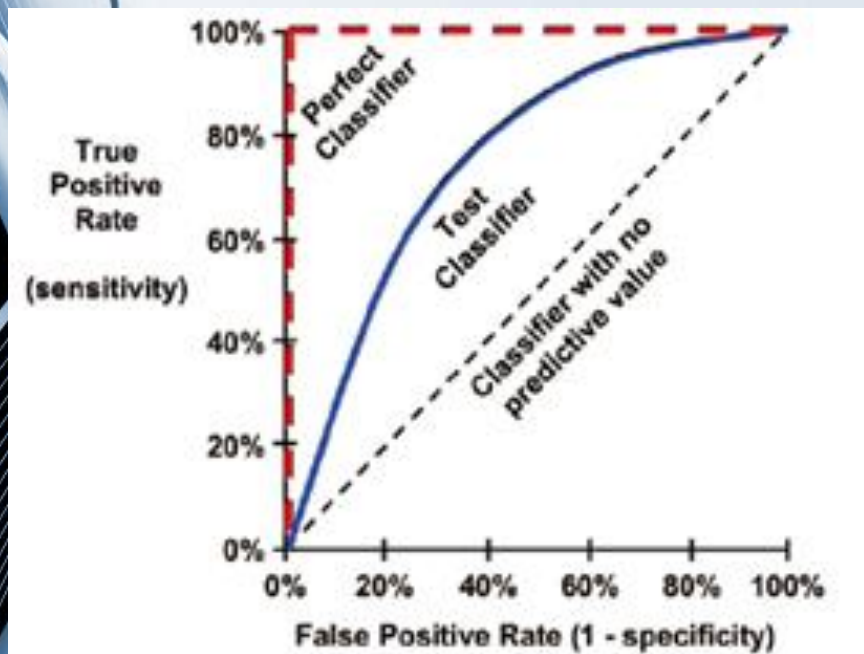
第一个是图中从左下角到右上角的直线，代表没有预测价值的分类器。

- 这种分类器发现真阳性和假阳性的比率完全相同，这意味着该分类器无法识别两者之间的差别





完美分类器

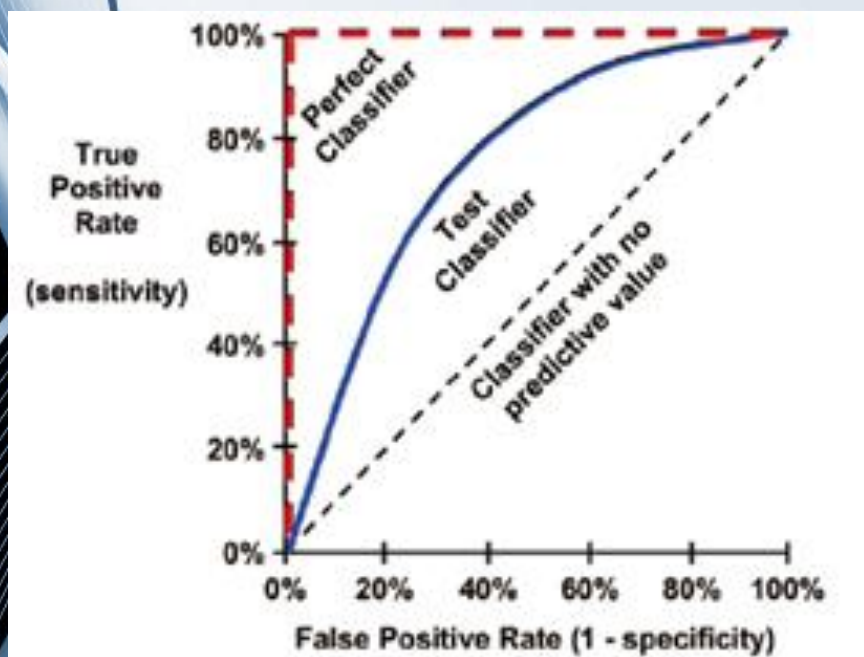


- 完美分类器拥有一条穿过了100%真阳性和0%假阳性点的曲线。
- 它在不正确地分出任何阴性的结果之前，已经正确地识别了所有的真阳性样本。



AUC

真实的分类器



- 真实的分类器，位于完美分类器，和无用分类器，之间的区域
- 离完美分类器越接近，说明能够越好地，识别阳性值。
- 可使用ROC曲线下面积 (Area Under the ROC, AUC)这个统计量来度量



AUC

- 与字面意思一样，AUC将ROC图看成是2维正方形，然后测量ROC曲线下的面积。
 - AUC的值从0.5(无预测值的分类器)到1.0(完美分类器)。
- 通常使用评分体系来解释AUC的得分：
 - $0.9 \sim 1.0 = A$ (优秀)。
 - $0.8 \sim 0.9 = B$ (良好)。
 - $0.7 \sim 0.8 = C$ (一般)。
 - $0.6 \sim 0.7 = D$ (很差)。
 - $0.5 \sim 0.6 = F$ (无法区分)。



中科院计算培训中心

谢 谢