上次课回顾

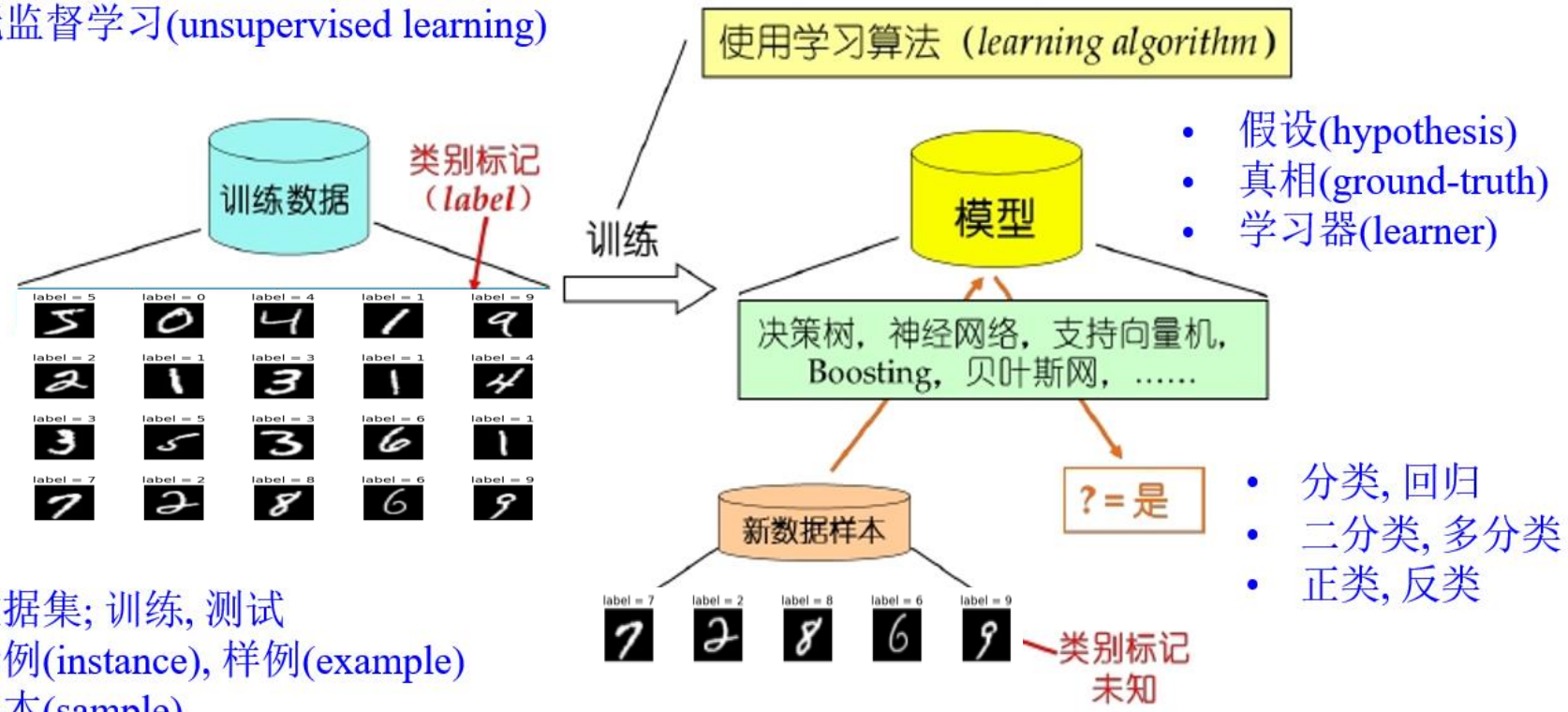
脱离具体问题，空泛地谈论“什么学习算法更好”，毫无意义！

一、基本术语

二、模型评估与选择

基本术语

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)



- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

- 分类, 回归
- 二分类, 多分类
- 正类, 反类

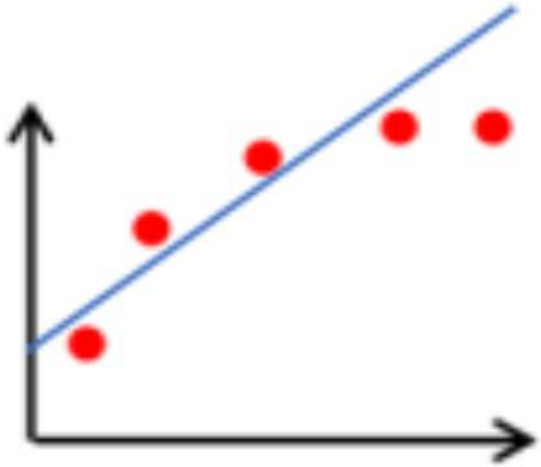
- 数据集; 训练, 测试
- 示例(instance), 样例(example)
- 样本(sample)
- 属性(attribute), 特征(feature); 属性值
- 属性空间, 样本空间, 输入空间
- 特征向量(feature vector)
- 标记空间, 输出空间

- 未见样本(unseen instance)
- 未知“分布”
- 独立同分布(i.i.d.)
- 泛化**(generalization)**

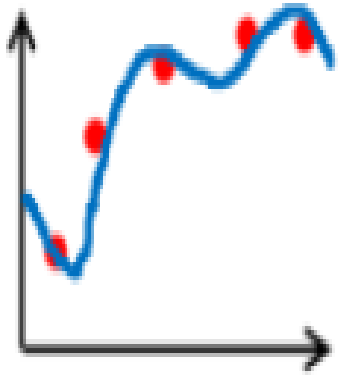
拟　合：　fitting

过拟合：　overfitting，在训练集上误差低，测试集上误差高；
拟　合：　fitting
欠拟合：　underfitting，模型在训练集上误差很高；

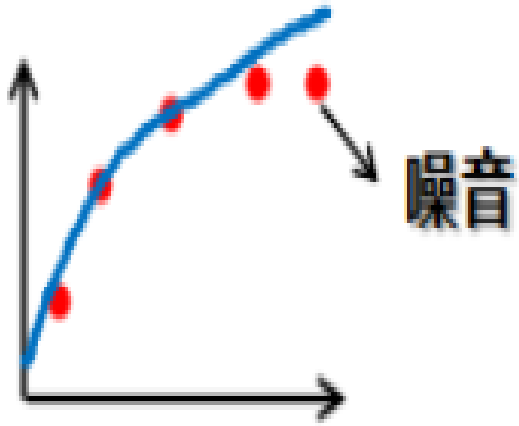
$$y = ax + b \qquad y = ax^4 + bx^3 + cx^2 + dx + e \qquad y = ax^2 + bx + c$$



欠拟合



过拟合



刚刚好

- 如何评价分类器的好坏？
- 在分类型模型评判的指标中，常见的方法有如下三种：

1、混淆矩阵：也称误差矩阵，Confusion Matrix

2、ROC曲线：Receiver Operating Characteristic Curve，受试者工作特征曲线

3、AUC面积：Area Under Curve，ROC曲线下与坐标轴围成的面积。

混淆矩阵

- 真正性（True Positive，TP）：样本的真实类别是正例，并且模型预测的结果也是正例
- 真反性（True Negative，TN）：样本的真实类别是负例，并且模型将其预测成为负例
- 假正性（False Positive，FP）：样本的真实类别是负例，但是模型将其预测成为正例
- 假反性（False Negative，FN）：样本的真实类别是正例，但是模型将其预测成为负例

混淆矩阵		预测值	
		positive	negative
真实值	Positive	TP	FN
	negative	FP	TN

混淆矩阵

例如： 有66只动物，其中13只猫，53只不是猫，分类器判断时这13只猫只有10只预测对了，其他动物也只预测对了45只。

混淆矩阵		预测值	
		猫	不是猫
真实值	猫	10	3
	不是猫	8	45

	公式	意义
准确率 ACC	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	分类模型所有判断正确的结果占总观测值的比重
精确率 PPV	$\text{Precision} = \frac{TP}{TP + FP}$	在模型预测是Positive的所有结果中，模型预测对的比重
灵敏度 TPR	$\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN}$	在真实值是Positive的所有结果中，模型预测对的比重
特异度 TNR	$\text{Specificity} = \frac{TN}{TN + FP}$	在真实值是Negative的所有结果中，模型预测对的比重

混淆矩阵

混淆矩阵		预测值	
		猫	不是猫
真实值	猫	10	3
	不是猫	8	45

Accuracy：在总共66个动物中，我们一共预测对了10 + 45=55个样本，所以准确率（Accuracy）=55/66 = 83.33%。

混淆矩阵

混淆矩阵		预测值	
		猫	不是猫
真实值	猫	10	3
	不是猫	8	45

Accuracy（猫）=55/66 = 83.33%

Precision（猫）= 10/18 = 55.6%

Recall（猫）= 10/13 = 76.9%

Specificity（猫）= 45/53 = 84.9%

表 2.1 分类结果混淆矩阵

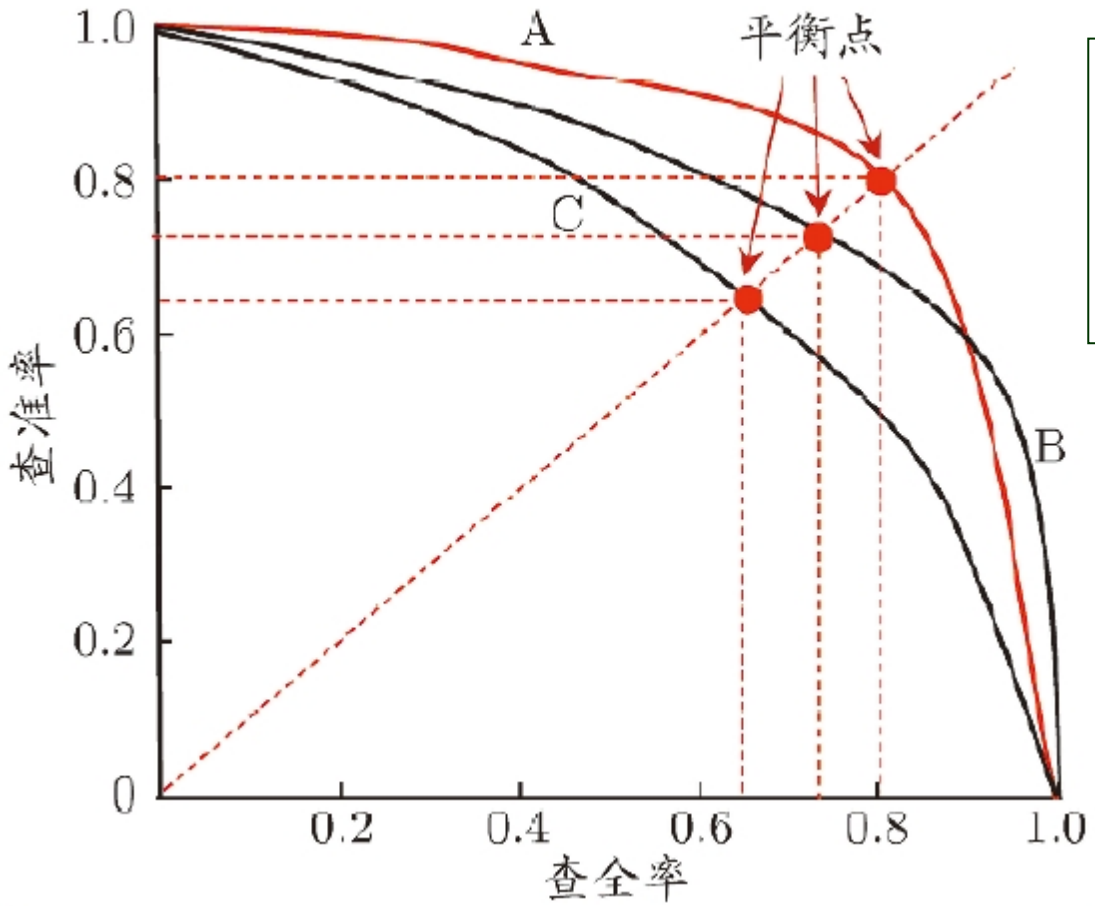
真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

- 查准率： $$P = \frac{TP}{TP + FP}$$

- 查全率： $$R = \frac{TP}{TP + FN}$$

PR图, BEP

根据学习器的预测结果按正例可能性大小对样例排序，并逐个把样本作为正例进行预测



PR图：

- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C
- 学习器 A ?? 学习器 B

BEP：

- 学习器 A 优于 学习器 B
- 学习器 A 优于 学习器 C
- 学习器 B 优于 学习器 C

混淆矩阵

F_1分数（ Score），又称平衡F_1分数（balanced F Score），它被定义为查准率和查全率的调和平均数。

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

除了F_1分数之外， F_2 分数和$F_{0.5}$ 分数在统计学中也得到大量的应用。

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

$\beta > 1$ 时查全率有更大影响; $\beta < 1$ 时查准率有更大影响

混淆矩阵

混淆矩阵		预测值	
		猫	不是猫
真实值	猫	10	3
	不是猫	8	45

对猫而言F1？

F1-Score=（2 * 0.769 * 0.556）/（ 0.769 + 0.556）= 64.54%

- 如何评价分类器的好坏？
- 在分类型模型评判的指标中，常见的方法有如下三种：

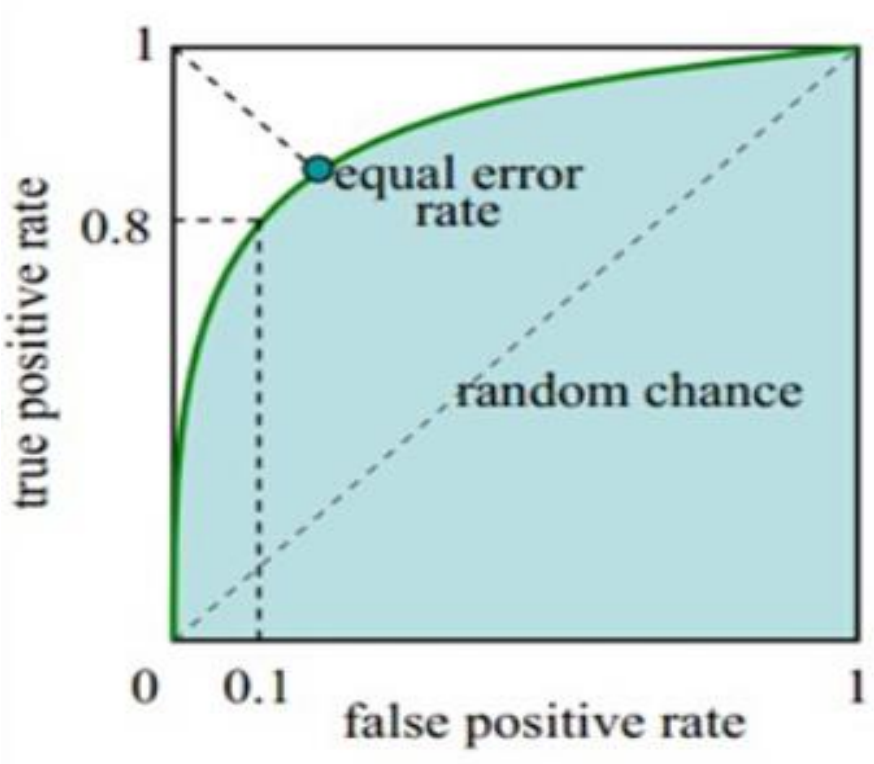
1、混淆矩阵：也称误差矩阵，Confusion Matrix

2、ROC曲线：Receiver Operating Characteristic Curve，受试者工作特征曲线

3、AUC面积：Area Under Curve，ROC曲线下与坐标轴围成的面积。

ROC曲线

ROC曲线：是一个概率曲线，在不同的阈值下绘制TPR与FPR的关系图，从本质上把“信号”与“噪声”分开。

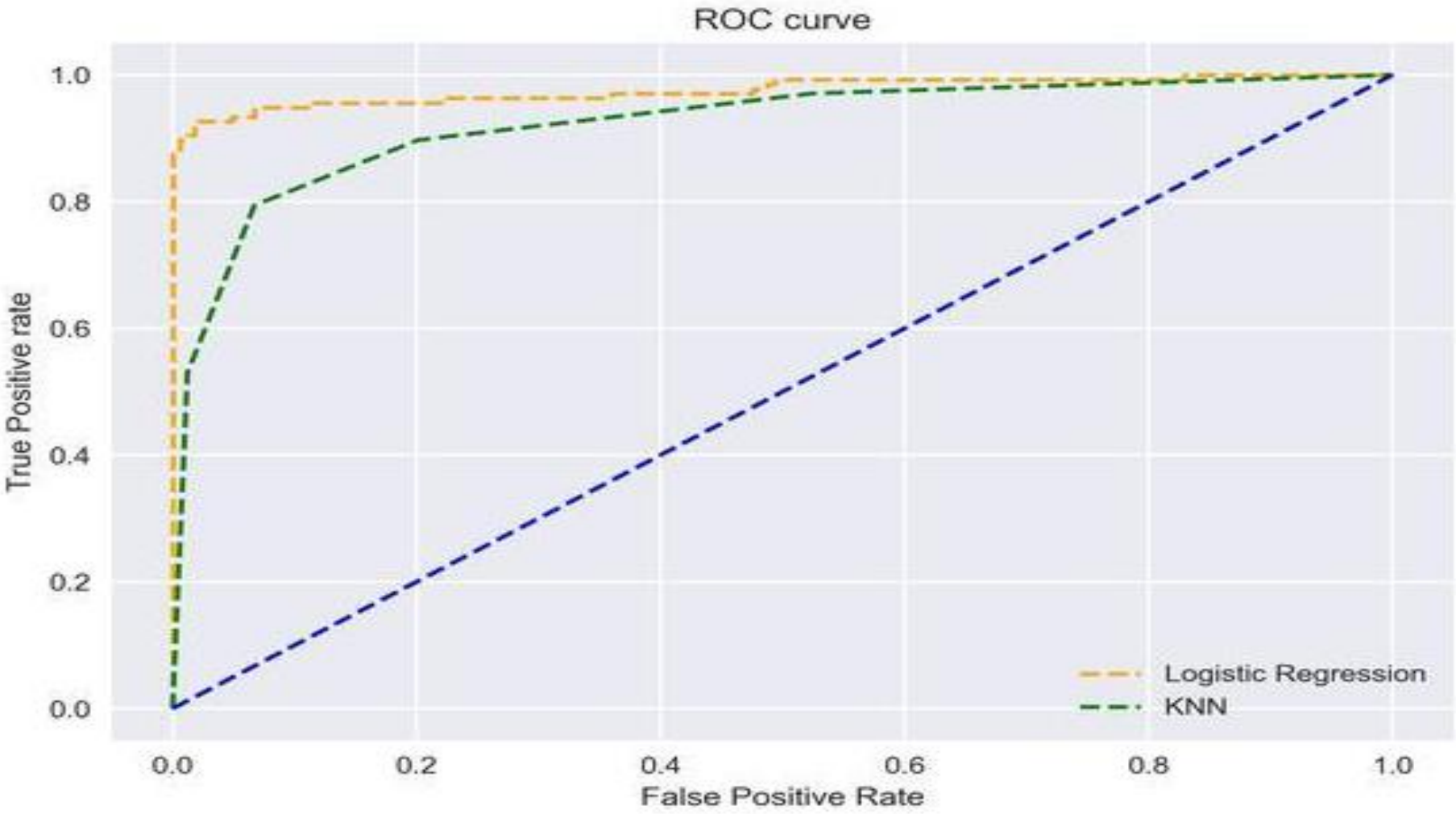


纵轴TPR：真正率，$tpr = \frac{TP}{TP + FN}$

TPR越大，预测正类中实际正类越多。

横轴FPR:假正率，$fpr = \frac{FP}{FP + TN}$

FPR越大，预测正类中实际负类越多。


ROC curve
True Positive rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression
KNN

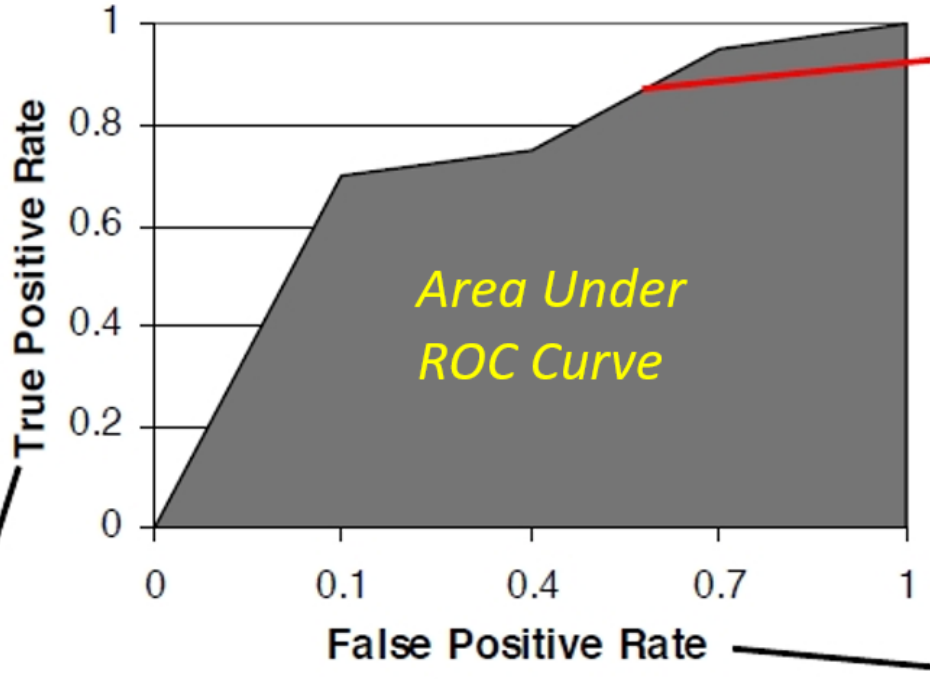
- 如何评价分类器的好坏？
- 在分类型模型评判的指标中，常见的方法有如下三种：

1、混淆矩阵：也称误差矩阵，Confusion Matrix

2、ROC曲线：Receiver Operating Characteristic Curve，受试者工作特征曲线

3、AUC面积：Area Under Curve，ROC曲线下与坐标轴围成的面积。

AUC: Area Under the ROC Curve



ROC (Receiver Operating Characteristic) Curve [Green & Swets, Book 66; Spackman, IWML'89]

The bigger, the better

$$tpr = \frac{TP}{TP + FN} = \frac{TP}{m_+}$$

$$fpr = \frac{FP}{FP + TN} = \frac{FP}{m_-}$$

混淆矩阵

混淆矩阵		预测值		
		猫	狗	猪
真实值	猫	10	1	2
	狗	3	15	4
	猪	5	6	20

课堂练习：计算图中狗的F1分数。

课后作业：用你熟悉的编程语言，编写一个2*2的混淆矩阵计算器，并计算图中猪的F1分数。

- 数据挖掘技术的产生与发展
- 数据挖掘研究的发展趋势
- 数据挖掘概念
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

数据挖掘方法与数据存储类型

- 近年来，数据存储类型越来越丰富
  - String、int、long、float、boolean等基本类型的存储
  - 自定义的对象数据类型<Key, Value>事务数据库（Transactional Database ）
  - 关系型数据库 （Related Database ）
  - 数据仓库（Data Warehouse ）
  - 在关系模型基础上发展的新型数据库
  - 面向应用的新型数据源
  - Web数据

数据挖掘技术可以应用到任何数据存储方式的知识挖掘中，但是因为源数据的存储类型的不同，挖掘的挑战性和技术会不同。

	块存储	文件存储	对象存储
概念	用高速（光纤）网络联接专业主机服务器的一种储存方式	使用文件系统，具有目录树结构	将数据和元数据当做一个对象，
速度			100ms - 1s，冷数据
可分步性			分步并发能力高
文件大小			适合各种大小
接口			Restful API
典型技术	SAN	HDFS、GFS	Swift、Amazon S3
适合场景	银行	数据中心	网络媒体文件存储

事务数据库中的数据挖掘

- 一个事务数据库是对事务型数据的收集。1993年，当Agrawal等开始讨论数据挖掘问题时，是以购物篮分析（Market Basket Analysis）作为商业应用背景的。
- 从事务数据库中发现知识是数据挖掘中研究较早但至今仍然很活跃的问题。通过特定的技术对事务数据库进行挖掘，可以获得动态行为所蕴藏的关联规则、分类、聚类以及预测等知识模式。
- 第三章将详细讲解

关系型数据库中的数据挖掘

- 关系型数据库是由一系列数据表组成的，相当成熟：
  - 成熟的语义模型（像实体-关系模型）；成熟的DBMS（像Oracle）
  - 成熟的查询语言（像SQL语言；可视化的辅助工具和优化软件。

- 一些更深入和亟待解决的问题：
  - **多维知识挖掘：** 传统的事务数据库挖掘所研究的知识一般是单维（Single-Demension）的，但是，在关系型数据库中，多维的知识更普遍和有应用价值。
    - 单维： “购买计算机的人也购买打印机”。
    - 多维： “什么样购买计算机的人也购买打印机的可能性更大？”。

关系型数据库中的数据挖掘

- 一些更深入和亟待解决的问题：
  - **多表挖掘：**关系型数据库是一系列表的集合。因此，多表挖掘是必然的。
  - **数量数据挖掘：** 关系型数据库经常包含非离散数量属性（如工资）。
  - **多层知识挖掘：**数据及其关联总是可在多个不同的概念层上来理解它。
  - **知识评价问题：**对传统的数据挖掘框架的知识评价问题，也是关系型数据库中数据挖掘走向实际应用必须要解决的问题。
  - **约束数据挖掘问题：**数据挖掘系统在用户的约束指导下进行，可以提高挖掘效率和准确度。

新型数据库中的数据挖掘

- 对象—关系型数据库（Object-Ralational Database）挖掘；
- 面向对象数据库的挖掘；
- 空间数据库的挖掘；
- 时态数据库的挖掘；
- 工程数据库（Engineering Database）的挖掘；
- 多媒体数据库（Multimedia Database）的挖掘；
- 等等

Web数据源中的数据挖

- 随着Internet的广泛使用，Web这一巨大的海洋中蕴藏着极其丰富的有用信息。
- 面向Web的数据挖掘比面向数据库和数据仓库的数据挖掘要复杂得多：
  - **异构数据源环境：**Web网站上的信息是异构：每个站点的信息和组织都不一样；存在大量的无结构的文本信息、复杂的多媒体信息；站点使用和安全性、私密性要求各异等等。
  - **数据的是复杂性：**有些是无结构的（如Web页），通常都是用长的句子或短语来表达文档类信息；有些可能是半结构的（如Email，HTML页）。当然有些具有很好的结构（如电子表格）。揭开这些复合对象蕴涵的一般性描述特征成为数据挖掘的不可推卸的责任。
  - **动态变化的应用环境：**
    - Web的信息是频繁变化的，像新闻、股票等信息是实时更新的。
    - 这种高变化也体现在页面的动态链接和随机存取上。
    - Web上的用户是难以预测的。
    - Web上的数据环境是高噪音的。

Web挖掘的研究主要流派

- **Web结构挖掘：挖掘Web上的链接结构。**
  - 通过Web页面间的链接信息可以识别出权威页面（Authoritative Page）、安全隐患（非法链接）等。
- **Web使用挖掘对Web上的Log日志记录的挖掘**
  - Web上的Log日志记录了包括URL请求、IP地址以及时间等的访问信息。
  - 分析和发现Log日志中蕴藏的规律可以帮助我们识别潜在的客户、跟踪Web服务的质量以及侦探非法访问的隐患等。
- **Web内容挖掘：Web的内容是丰富的，而且构成成分是复杂的（无结构的、半结构的等），对内容的分析是重要而艰巨的工作。**
  - Web的内容主要是包含文本、声音、图片等的文档信息。
  - 文本挖掘（Text Mining）和Web搜索引擎（Search Engine）等相关领域的研究。目
  - 多媒体信息挖掘技术。

- 数据挖掘技术的产生与发展
- 数据挖掘研究的发展趋势
- 数据挖掘概念
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

粗糙集理论

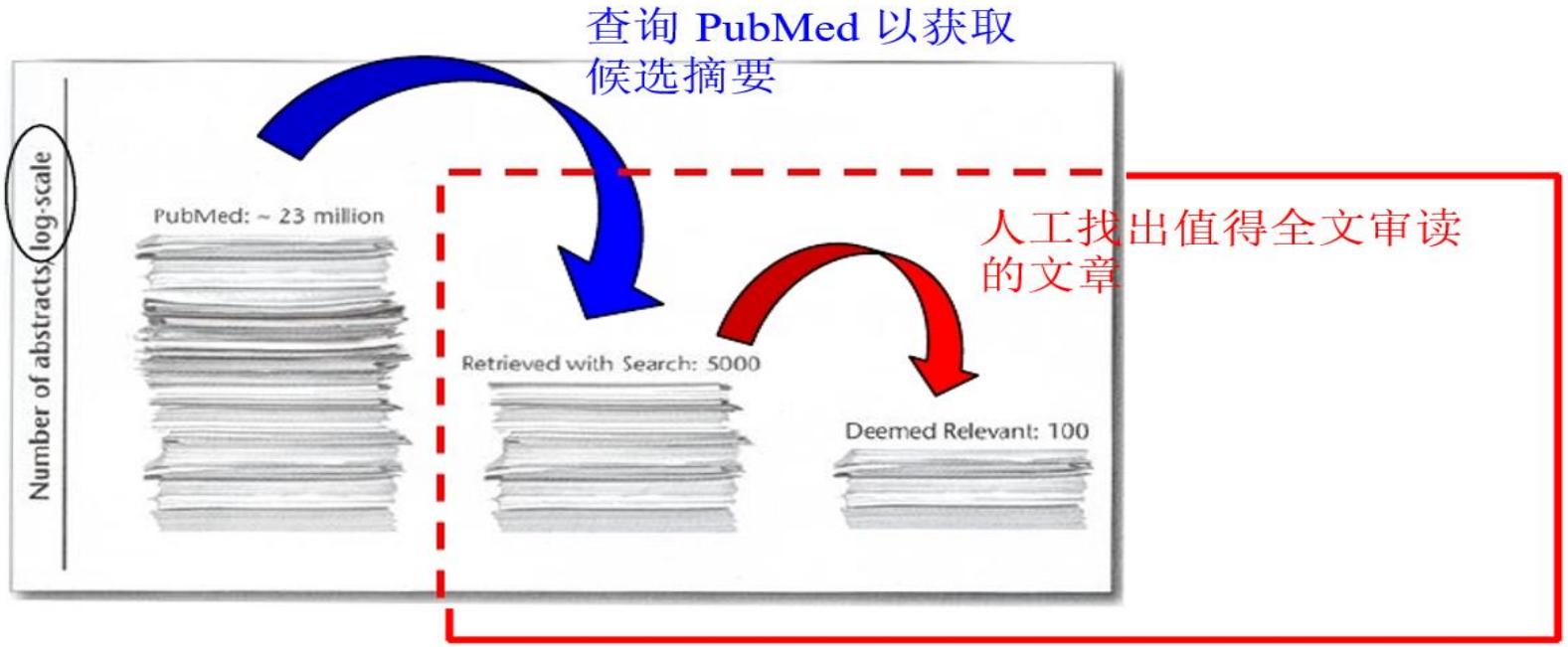
- 粗糙集理论是一种研究不精确、不确定性知识的数学工具，由波兰科学家Z. Pawlak在1982年首先提出的。
- 粗糙集一经提出就立刻引起数据挖掘研究人员的注意，并被广泛讨论。
- 粗糙集的知识形成思想可以概括为：一种类别对应于一个概念，知识由概念组成。
- 粗糙集对不精确概念的描述方法是通过下近似（Lower Approximation）和上近似（Upper Approximation）概念来表示：
  - 一个概念（或集合）的下近似概念（或集合）中的元素肯定属于该概念（或集合）
  - 一个概念（或集合）的上近似概念（或集合）只是可能属于该概念。

- 数据挖掘技术的产生与发展
- 数据挖掘研究的发展趋势
- 数据挖掘概念
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析

数据挖掘的应用分析

“文献筛选”的故事

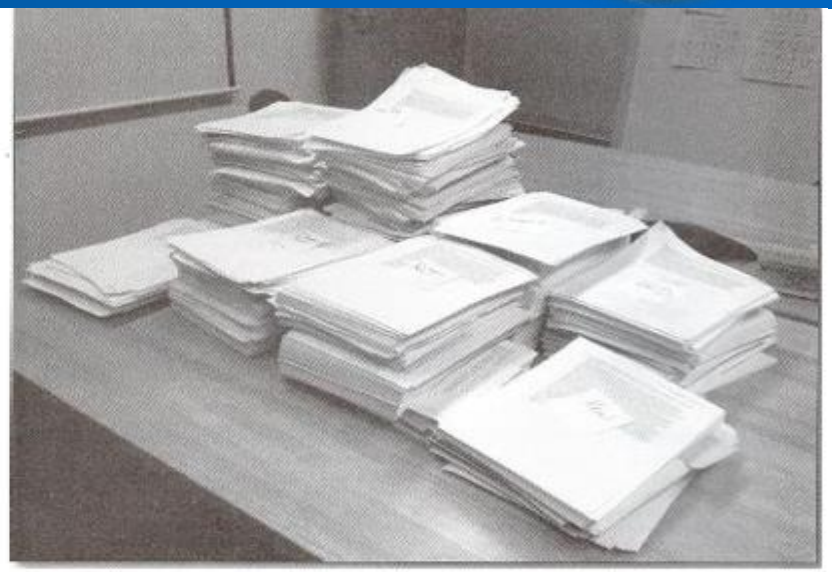
在“循证医学”（evidence-based medicine）中，针对特定的临床问题，先要对相关研究报告进行详尽评估



数据挖掘的应用分析

“文献筛选”的故事

在一项关于婴儿和儿童残疾的研究中，美国Tufts医学中心筛选了约 33,000 篇摘要



a portion of the 33,000 abstracts

尽管Tufts医学中心的专家效率很高，对每篇摘要只需 30 秒钟，但该工作仍花费了 250 小时

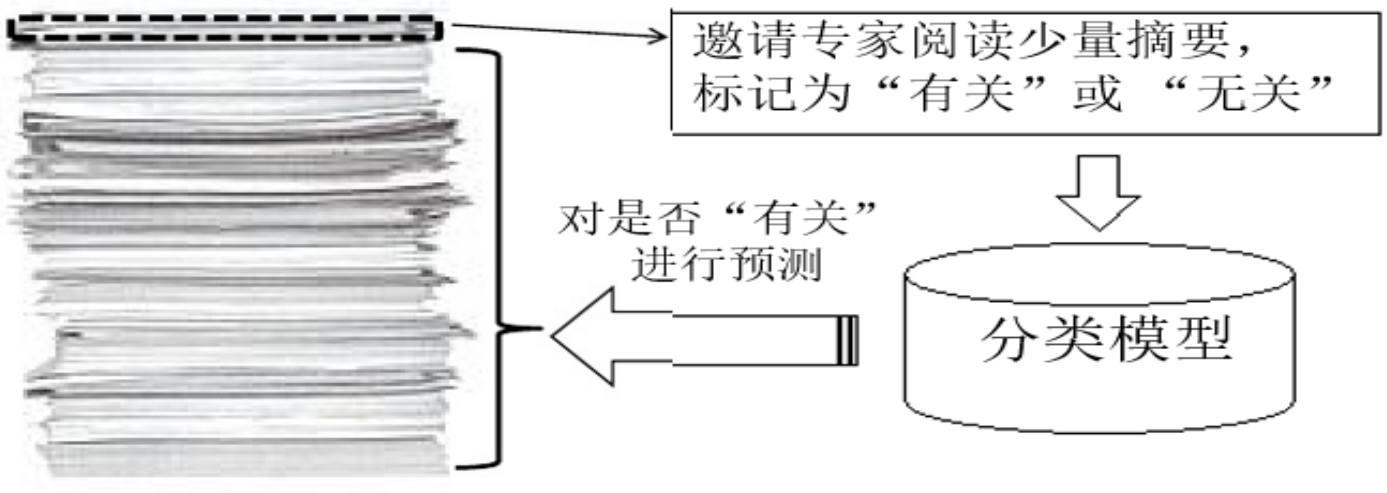
每项新的研究都要重复这个麻烦的过程！

需筛选的文章数在不断显著增长！

数据挖掘的应用分析

“文献筛选”的故事

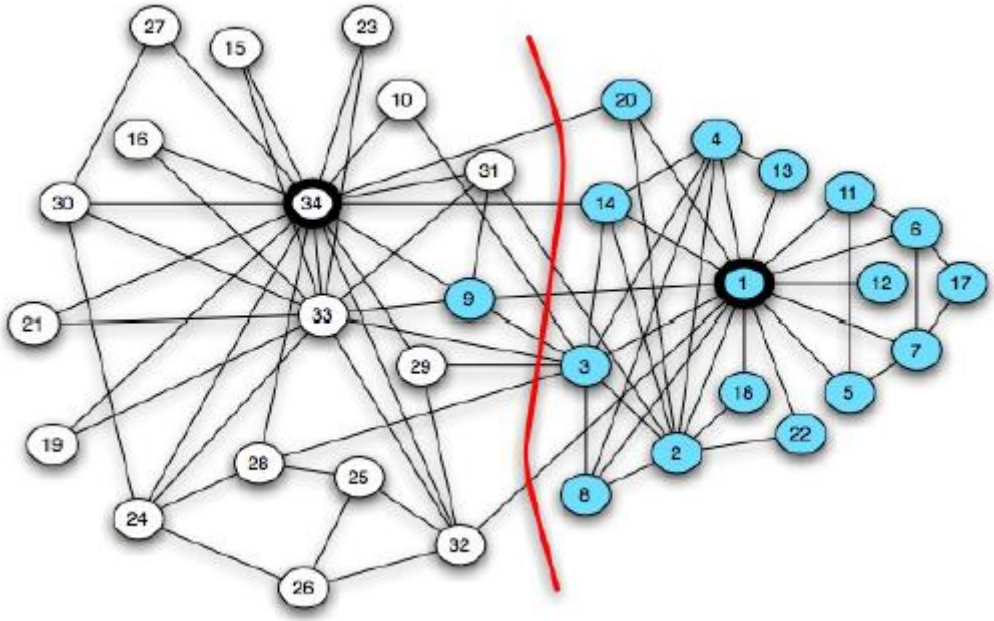
为了降低昂贵的成本, Tufts医学中心引入了机器学习技术



人类专家只需阅读 **50** 篇摘要， 系统的自动筛选精度就达到 **93%**
人类专家阅读 **1,000** 篇摘要， 则系统的自动筛选敏感度达到 **95%**
(人类专家以前需阅读 **33,000** 篇摘要才能获得此效果)

数据挖掘与社会网络

- 社会网络是指社会个体成员之间因为互动而形成的相对稳定的关系网络体系。社会网络分析技术可以帮助人们发现潜在的社会关系或者社会结构性知识。

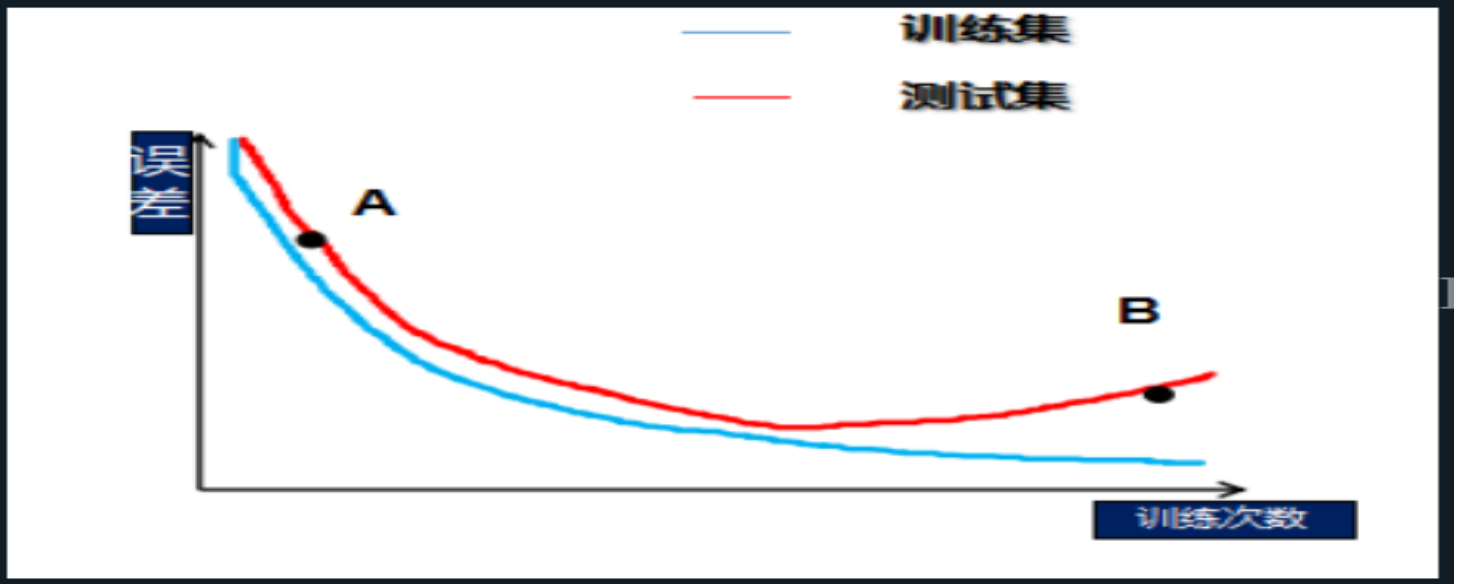


- 许多例子可以说明社会网络离不开计算机技术支持，如：
  - 观察互联网中网页的链接规律(极多数页面有很少的链接数，而极少数页面却有很大的链接数)发现了“幂次定律”在社会网络中普遍存在。
  - 现实世界许多现象符合社会网络的“小世界模型”。
  - 现代社交媒介是大数据环境，社会网络是了解它的很好途径。利用数据挖掘技术等进行大数据自动化分析有很高应用价值。 如利用Facebook(脸谱)等社交媒介来帮助美国总统选举。

作业：

混淆矩阵		真实值		
		猫	狗	猪
预测值	猫	10	1	2
	狗	3	15	4
	猪	5	6	20

课后作业：用你熟悉的编程语言，编写一个2*2的混淆矩阵计算器，并计算图中猪的F1分数。



1、

请问，图中A与B分别处于什么状态？

- A、欠拟合，欠拟合
- B、欠拟合，过拟合
- C、过拟合，欠拟合
- D、过拟合，过拟合

1、假设，我们现在利用5折交叉验证的方法来确定模型的超参数，一共有4组超参数，我们可以知道，5折交叉验证，每一组超参数将会得到5个子模型的性能评分，假设评分如下，我们应该选择哪组超参数？

- A、子模型1:0.8 子模型2:0.7 子模型3:0.8 子模型4:0.6 子模型5:0.5
- B、子模型1:0.9 子模型2:0.7 子模型3:0.8 子模型4:0.6 子模型5:0.5
- C、子模型1:0.5 子模型2:0.6 子模型3:0.7 子模型4:0.6 子模型5:0.5
- ✓ D、子模型1:0.8 子模型2:0.8 子模型3:0.8 子模型4:0.8 子模型5:0.6

作业：

继续以癌症检测系统为例，癌症检测系统的输出不是有癌症就是健康，这里为了方便，就用 1 表示患有癌症，0 表示健康。假设现在拿 10000 条数据来进行测试，其中有 9978 条数据

数据的真实类别是 1 却预测

1，有 8 条数据的真实类别

如果我们把这些结果组成如

真实预测	0	1
0	9978	12
1	2	8

混淆矩阵中每个格子所代表的的意义也很明显，意义如下：

真实预测	0	1
0	预测 `0` 正确的数量	预测 `1` 错误的数量
1	预测 `0` 错误的数量	预测 `1` 正确的数量

如果将正确看成是 True，错误看成是 False， 0 看成是 Negtive， 1 看成是 Positive。然后将上表中的文字替换掉，混淆矩阵如下：

真实预测	0	1
0	TN	FP
1	FN	TP