模式识别与机器学习

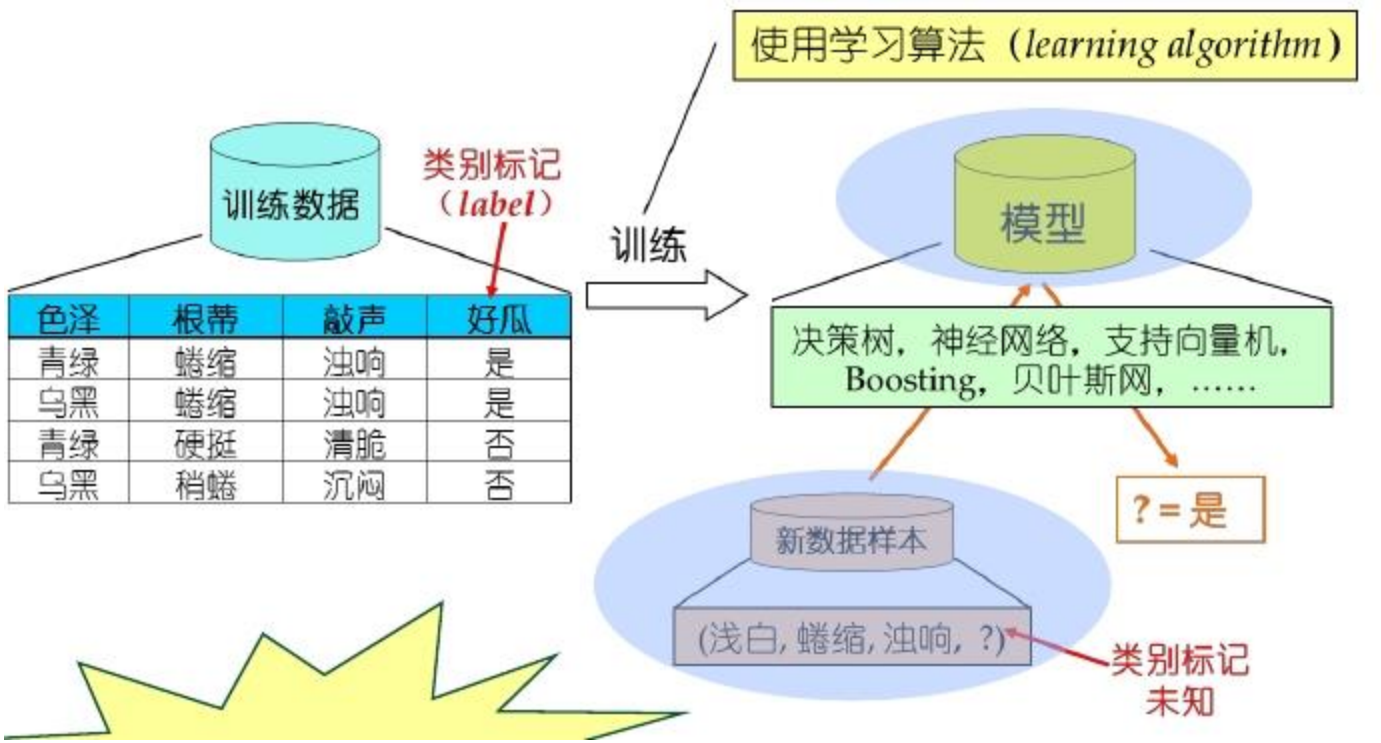
01 绪论
（第2部分）

大连理工大学 人工智能学院
School of Artificial Intelligence, Dalian University of Technology

模型评估与选择

Model Evalution and Selection

典型的机器学习过程



色泽	根蒂	敲声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

泛化能力强！

能很好地适用于未知样本

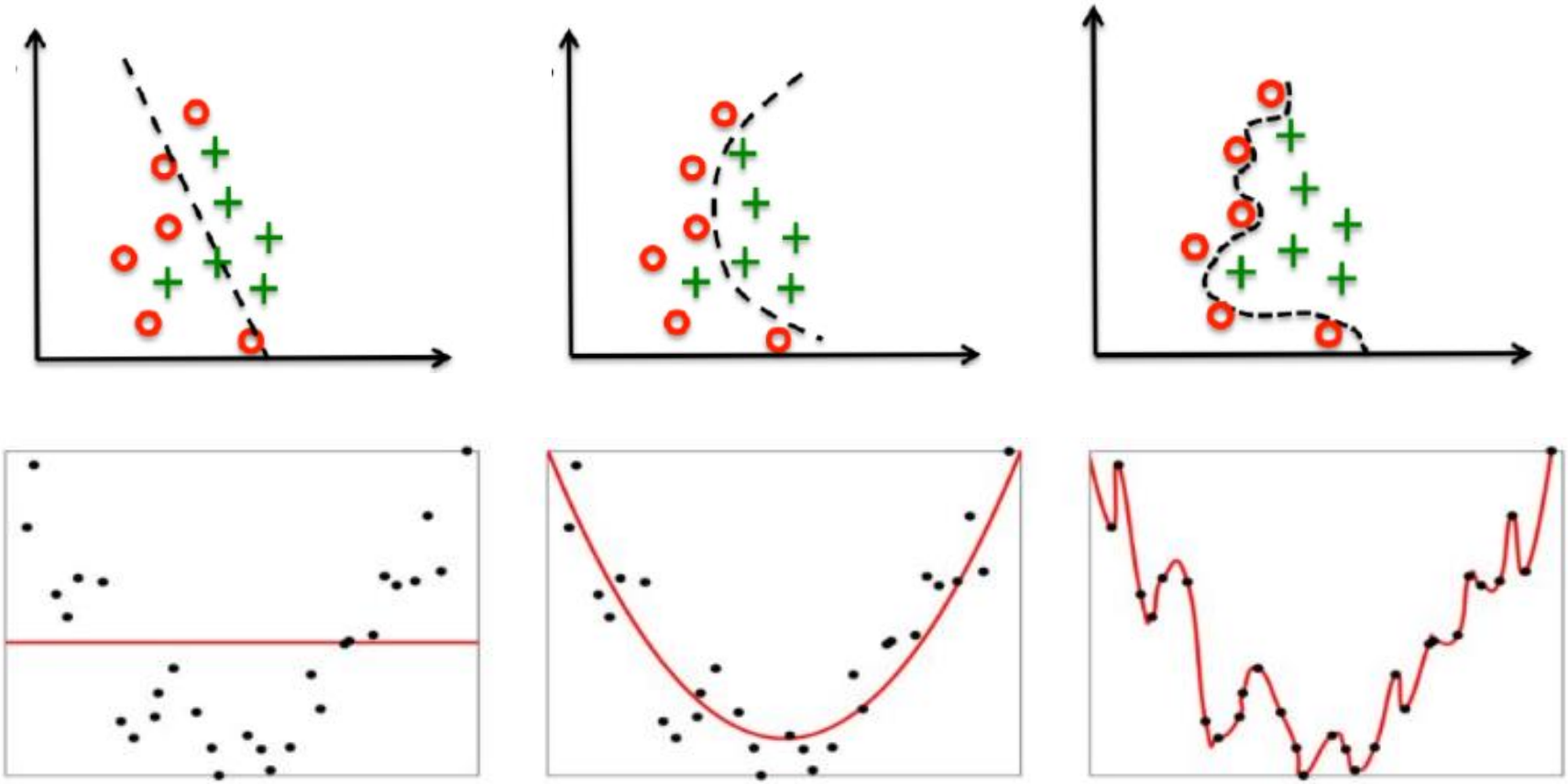
例如，错误率低、精度高

- 误差：样本真实输出与预测输出之间的差异
  - 训练(经验)误差：训练集上
  - 测试误差：测试集
  - 泛化误差：除训练集外所有样本

  - 泛化误差越小越好
  - 经验误差是否越小越好？

NO! 因为会出现“过拟合”(overfitting)

- 欠拟合：对训练样本的一般性质尚未学好
- 过拟合: 学习器把训练样本学习的“太好”，将训练样本本身的特点当做所有样本的一般性质，导致泛化性能下降



欠拟合　　拟合　　过拟合

三个关键问题:

- 如何获得测试结果？ 评估方法
- 如何评估性能优劣？ 性能度量
- 如何判断实质差别？ 比较检验

关键：怎么获得“测试集”(test set)？

我们假设测试集是从样本真实分布中独立采样获得，将测试集上的“测试误差”作为泛化误差的近似，测试集应该与训练集“互斥”

常见方法：

- 留出法 (hold-out)
- **交叉验证法 (cross validation)**
- 自助法 (bootstrap)

通常将包含个m样本的数据集$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
拆分成训练集S和测试集T:

- 留出法:
  - 直接将数据集划分为两个互斥集合
  - 训练/测试集划分要尽可能保持数据分布的一致性
  - 一般若干次随机划分、重复实验取平均值
  - 训练/测试样本比例通常为2:1~4:1

k-折交叉验证法

- *k*-fold cross validation
- 将数据集分层采样划分为k个大小相似的互斥子集，每次用k-1个子集的并集作为训练集，余下的子集作为测试集，最终返回k个测试结果的均值，k最常用的取值是10.

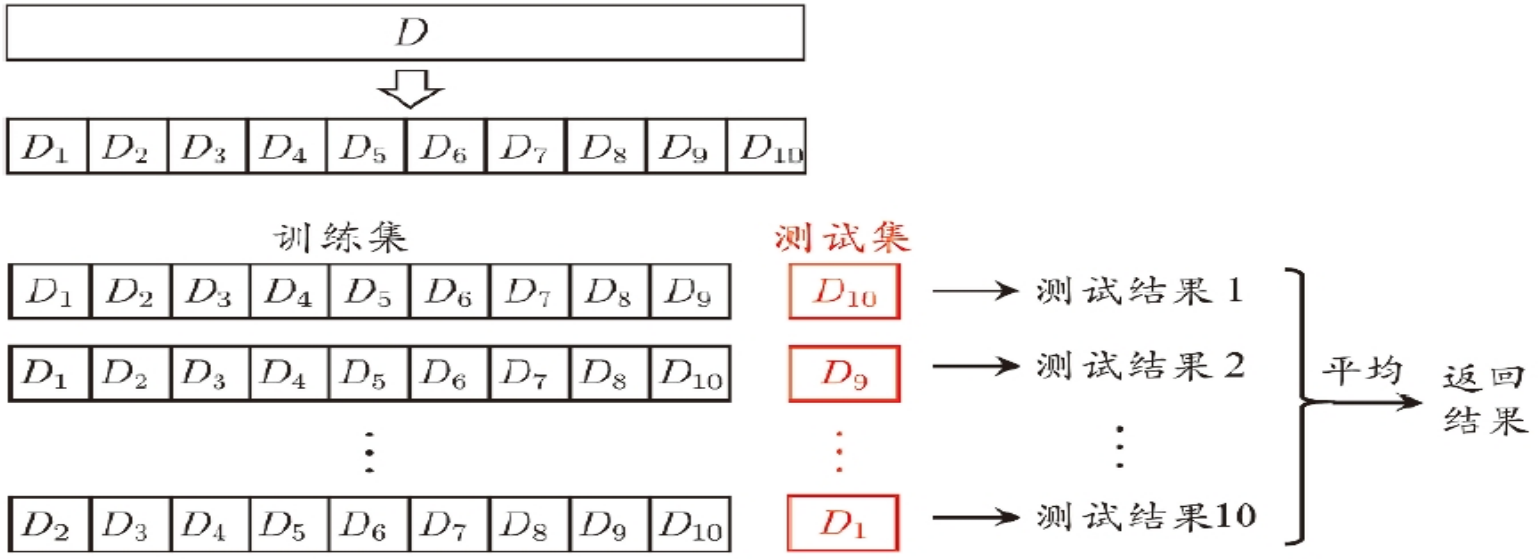


图 2.2 10 折交叉验证示意图

k-折交叉验证法

与留出法类似，将数据集D划分为k个子集同样存在多种划分方式，为了减小因样本划分不同而引入的差别，k折交叉验证通常随机使用不同的划分重复p次，最终的评估结果是这p次k折交叉验证结果的均值，例如常见的“10次10折交叉验证”

假设数据集D包含m个样本，若令 $k = m$ ，则得到留一法：

- 不受随机样本划分方式的影响
- 结果往往比较准确
- 当数据集比较大时，计算开销难以忍受

(leave-one-out, LOO)

自助法（Bootstrapping）：

以自助采样法（即有放回的采样或重复采样）为基础，对数据集 D 有放回采样 m 次得到训练集D'， $D \setminus D'$用做测试集。

- 实际模型与预期模型都使用 m 个训练样本
- 约有1/3的样本没在训练集中出现 (如果m趋于无穷，应为e^{-1}.why?)
- 从初始数据集中产生多个不同的训练集，对集成学习有很大的好处
- 自助法在数据集较小、难以有效划分训练/测试集时很有用；
- 由于改变了数据集分布可能引入估计偏差，所以在数据量足够时，留出法和交叉验证法更常用。

算法的参数：一般由人工设定，亦称“超参数”

模型的参数：一般由学习确定

调参过程相似：先产生若干模型，然后基于某种评估方法进行选择

参数调得好不好对性能往往对最终性能有关键影响

区别：训练集 vs. 测试集 vs. 验证集 (validation set)

算法参数选定后，要用“训练集+验证集”重新训练最终模型

三个关键问题:

- 如何获得测试结果？ 评估方法
- 如何评估性能优劣？ 性能度量
- 如何判断实质差别？ 比较检验

性能度量

- 在预测任务中，给定样例集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$
- 评估学习器的性能f 也即把预测结果$f(\boldsymbol{x})$和真实标记比较.

- 性能度量(performance measure)是衡量模型泛化能力的评价标准，反映了任务需求
- 使用不同的性能度量往往会导致不同的评判结果

对于回归(regression) 任务常用均方误差：

$$E(f; D) = \frac{1}{m} \sum_{i=1}^{m} (f(\boldsymbol{x}_i) - y_i)^2$$

什么样的模型是“好”的，不仅取决于算法和数据，还取决于任务需求

对于分类任务,错误率和精度是最常用的两种性能度量:

- 错误率：分错样本占样本总数的比例
- 精度：分对样本占样本总数的比率

- 错误率：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)\neq y_i\right)$$

- 精度：

$$\begin{aligned}\mathrm{acc}(f;D) &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f\left(\boldsymbol{x}_i\right)=y_i\right)\\ &= 1-E(f;D)\ .\end{aligned}$$

什么样的模型是“好”的，不仅取决于算法和数据，还取决于任务需求

信息检索、Web搜索等场景中经常需要衡量正例被预测出来的比率或者预测出来的正例中正确的比率，此时查准率和查全率比错误率和精度更适合。

统计真实标记和预测结果的组合可以得到“混淆矩阵”

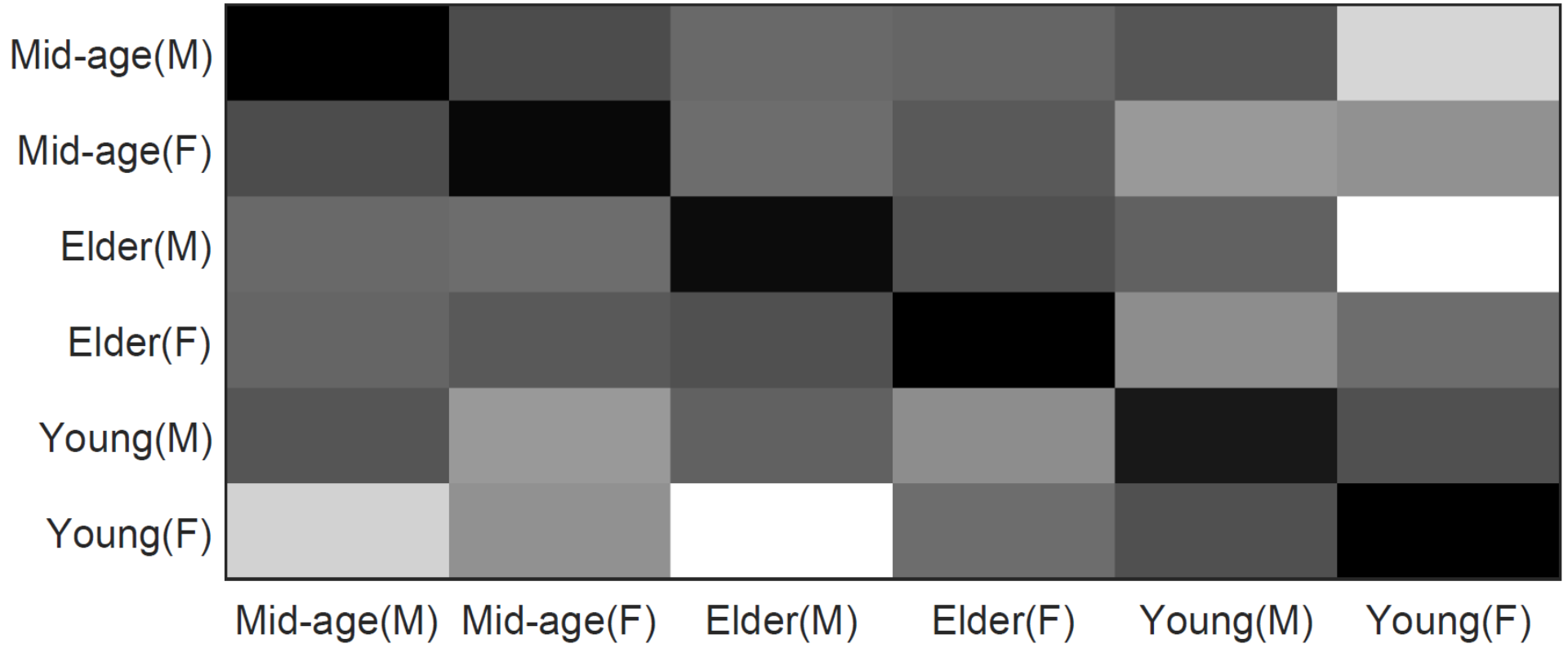
分类结果混淆矩阵

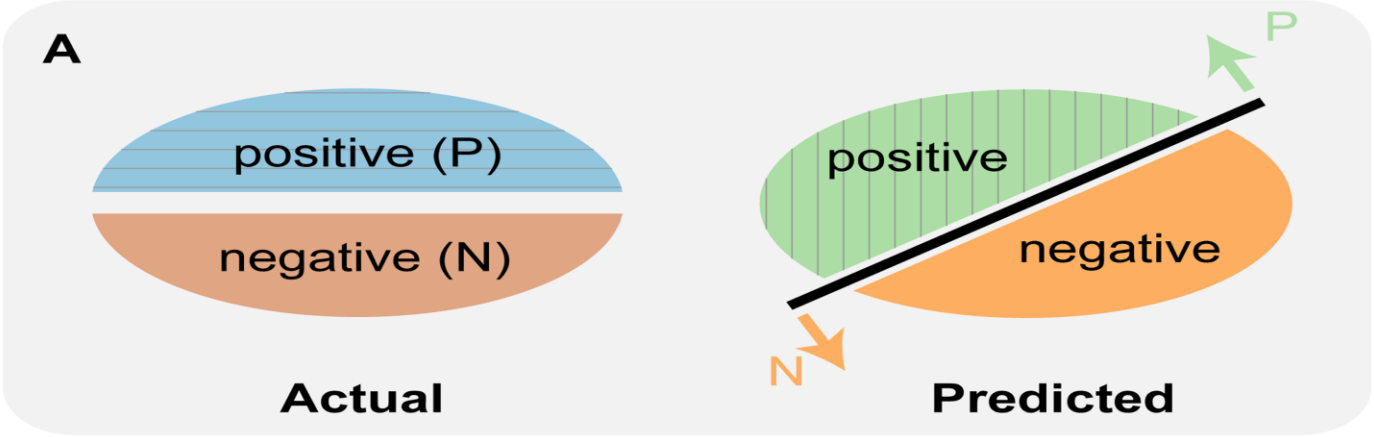
真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

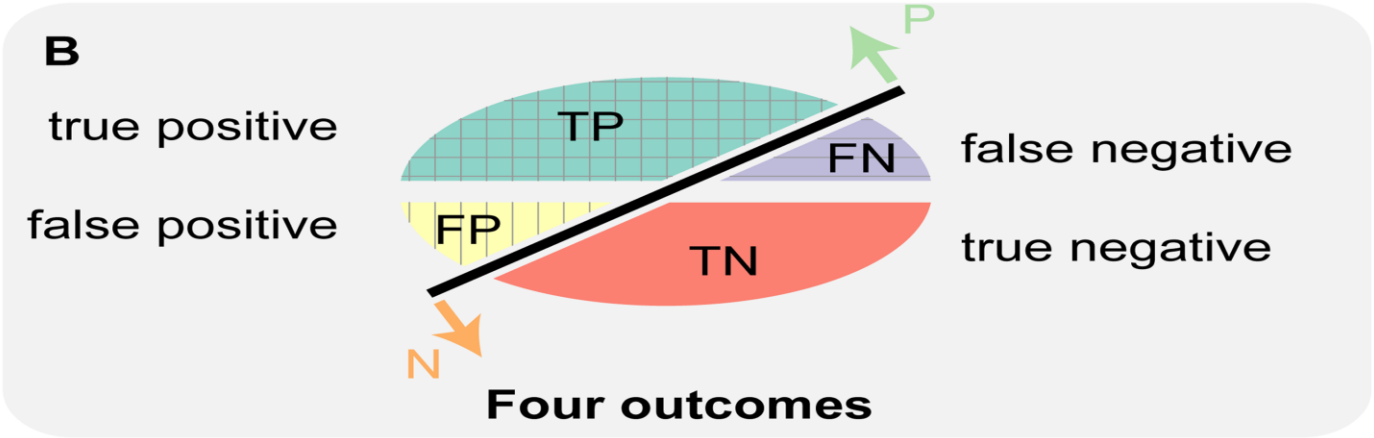
查准率 $P = \frac{TP}{TP + FP}$

查全率 $R = \frac{TP}{TP + FN}$

性能度量


Mid-age(M)
Mid-age(F)
Elder(M)
Elder(F)
Young(M)
Young(F)
Mid-age(M)
Mid-age(F)
Elder(M)
Elder(F)
Young(M)
Young(F)





- 查准率（precision）：

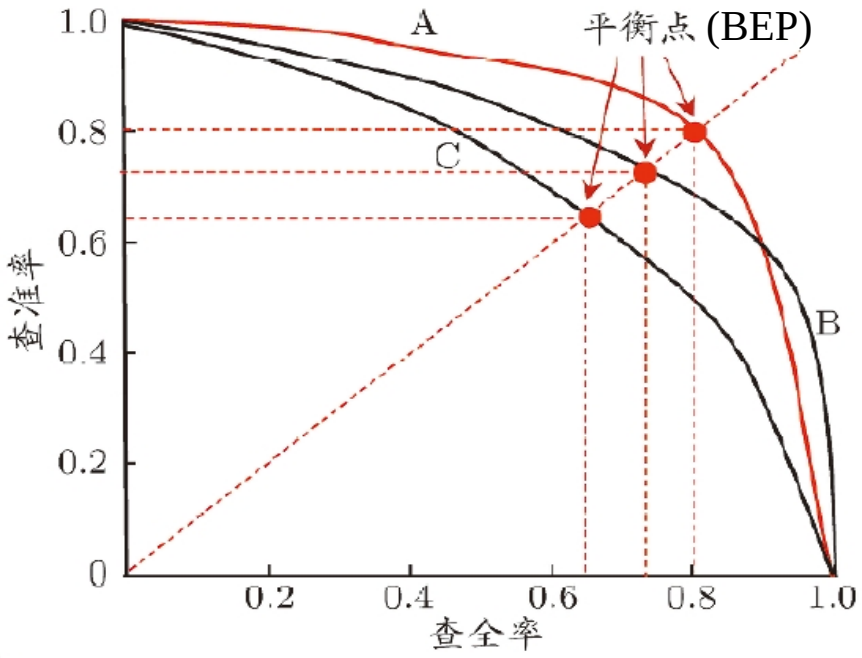
$$P = \frac{TP}{TP + FP}$$

- 查全率（Recall）：

$$R = \frac{TP}{TP + FN}$$

性能度量：P-R曲线、BEP

- 根据学习器的预测结果按正例可能性大小对样例进行排序，并逐个把样本作为正例进行预测，则可以得到查准率-查全率曲线，简称“P-R曲线”。
- 平衡点是曲线上“查准率=查全率”时的取值，可用来用于度量P-R曲线有交叉的分类器性能高低。



P-R曲线：

- 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

性能度量：F-score

比 BEP 更常用的 F1 度量： F1度量的由来是加权调和平均

$$\frac{1}{F_\beta}=\frac{1}{1+\beta^2}\left(\frac{1}{P}+\frac{\beta^2}{R}\right)$$

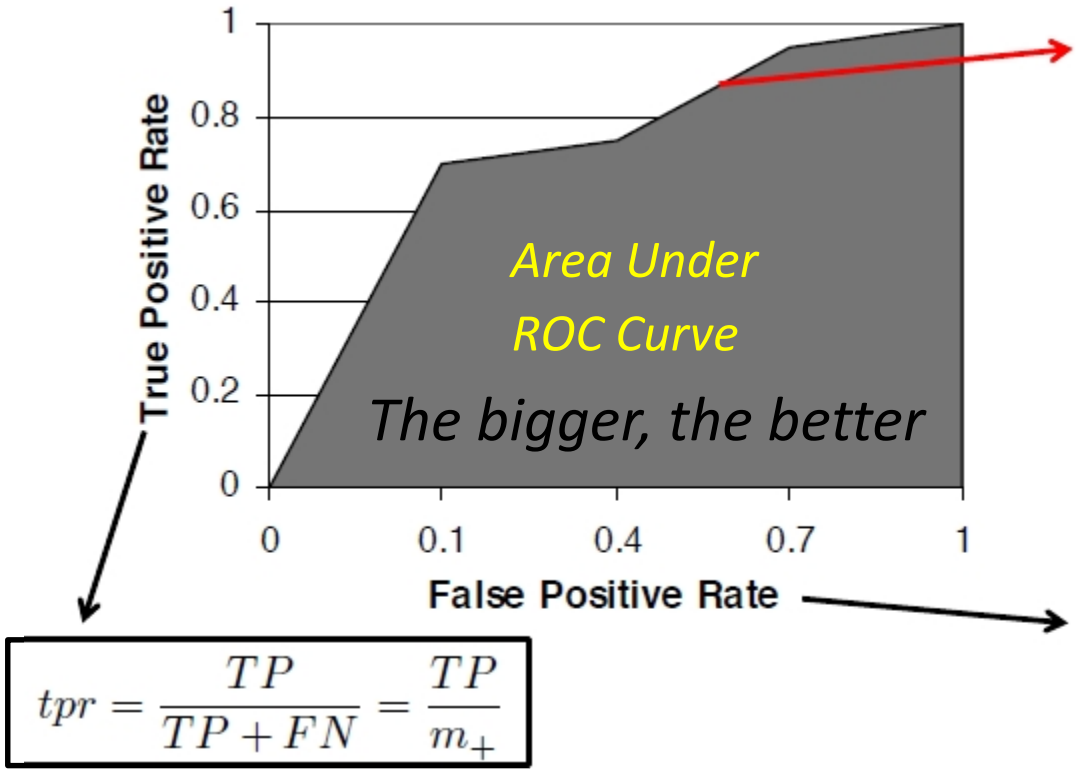
$$F_\beta=\frac{(1+\beta^2)\times P\times R}{(\beta^2\times P)+R}$$

$\beta=1$：标准F1
$\beta>1$：偏重查全率(逃犯信息检索)
$\beta<1$：偏重查准率(商品推荐系统)

调和平均数主要特点：易受极端值的影响，且受极小值的影响比受极大值的影响更大。

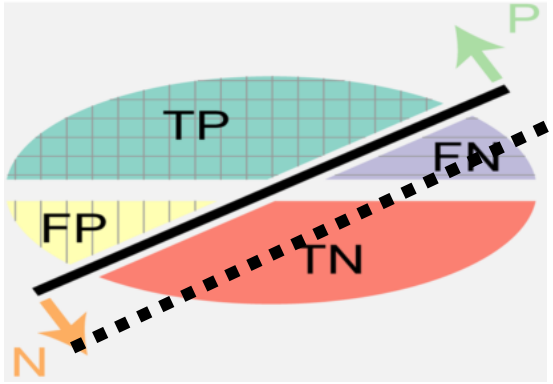
参考文献：https://blog.csdn.net/program_developer/article/details/79937291

AUC: **A**rea **U**nder the ROC **C**urve



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

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



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

- 更多性能指标讨论：http://charleshm.github.io/2016/03/Model-Performance/

- 关于性能比较：
  - 测试性能并不等于泛化性能
  - 测试性能随着测试集的变化而变化
  - 很多机器学习算法本身有一定的随机性

直接选取相应评估方法在相应度量下比大小的方法不可取！

假设检验为学习器性能比较提供了重要依据，基于其结果我们可以推断出若在测试集上观察到学习器A比B好，则A的泛化性能是否在统计意义上优于B，以及这个结论的把握有多大。

参考《机器学习》周志华 第2.4节

“误差”包含了哪些因素？

换言之，从机器学习的角度看，

“误差”从何而来?

对回归任务，泛化误差可通过“偏差-方差分解”拆解为：

$$E(f;D) = bias^2(\boldsymbol{x}) + var(\boldsymbol{x}) + \varepsilon^2$$

期望输出与真实输出的差别

$$bias^2(\boldsymbol{x}) = \left(\bar{f}(\boldsymbol{x}) - y\right)^2$$

同样大小的训练集的变动，所导致的性能变化

$$var(\boldsymbol{x}) = \mathbb{E}_D\left[\left(f(\boldsymbol{x};D) - \bar{f}(\boldsymbol{x})\right)^2\right]$$

训练样本的标记与真实标记有区别

表达了当前任务上任何学习算法所能达到的期望泛化误差下界

$$\varepsilon^2 = \mathbb{E}_D\left[(y_D - y)^2\right]$$

泛化性能是由学习算法的能力、算法在不同数据集的稳定性以及学习任务本身的难度共同决定

一般而言，偏差与方差存在冲突：

- 训练不足时，学习器拟合能力不强，偏差主导
- 随着训练程度加深，学习器拟合能力逐渐增强，方差逐渐主导
- 训练充足后，学习器的拟合能力很强，方差主导

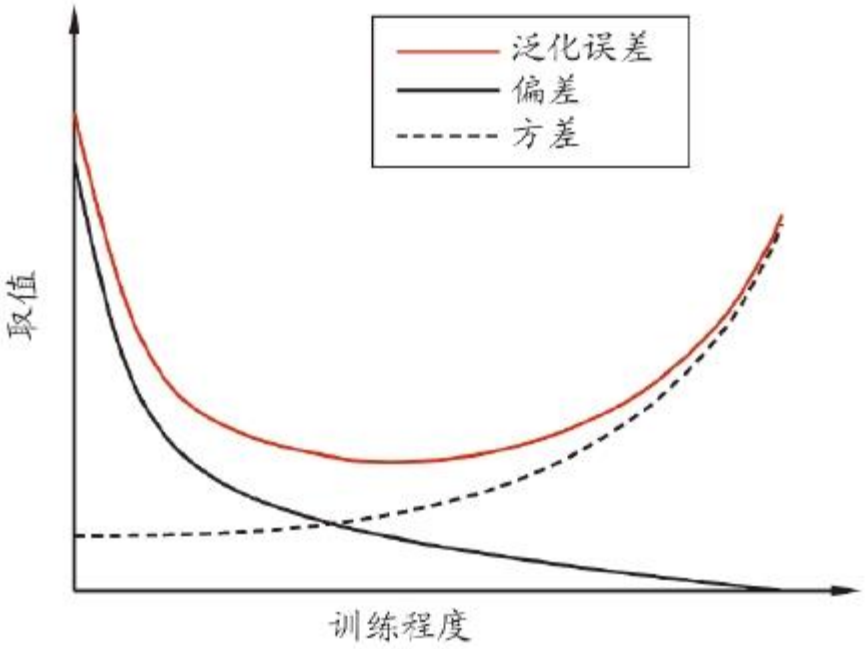


图 2.9 泛化误差与偏差、方差的关系示意图

➢ 更多偏差-方差的讨论：https://zhuanlan.zhihu.com/p/38853908

小结


机器学习
数据集
每条记录是关于一个事件或对象的描述，称为一个“示例”或“样本”
反映事件或对象在某方面的表现或性质的事项，称为“属性”或“特征”
属性张成的空间称为“属性空间”、“样本空间”或“输入空间”
训练数据集（训练样本）、测试数据集（测试样本）、标签
模型
从数据中学得的结果。从数据中学习的模型的过程称为“学习”或“训练”
学得模型后，使用其进行预测的过程称为“测试”
机器学习的目标是使学得的模型能很好地适用于“新样本”，而不是仅仅在训练样本上工作得很好。学得模型适用于新样本的能力，称为“泛化能力”
算法
监督学习：训练数据有标签
分类（K-NN, SVM, 决策树）
回归（线性回归）
非监督学习：训练数据没有标签
聚类
降维（主成分分析）
不依赖标签

三个关键问题:

- 如何获得测试结果？

  评估方法：留出法、交叉验证法、自助法

- 如何评估性能优劣？

  性能度量：均方误差、错误率、精度、查准率、查全率、ROC曲线等

- 如何判断实质差别？

  比较检验

04

PART FOUR

参考资源

Resource

- 在线课程：

  吴恩达（Andrew Ng)：机器学习

  https://www.bilibili.com/video/av50747658?from=search&seid=9857444623866219885

  李宏毅（Hung-Yi Lee）：机器学习

  https://www.bilibili.com/video/av10590361?from=search&seid=10930215693466580647

- 公众号：
  - VALSE
  - 机器之心
  - 极市平台
  - 人工智能前沿讲习班
  - 深度学习大讲堂
  - SIGAI
  - …