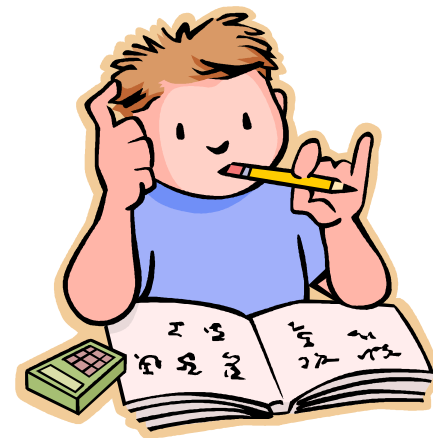




- 数据挖掘技术的产生
- 数据挖掘概念
- 数据挖掘技术的发展趋势
- 数据挖掘技术的分类问题
- 数据挖掘常用的知识表示模式与方法
- 不同数据存储形式下的数据挖掘问题
- 粗糙集方法及其在数据挖掘中的应用
- 数据挖掘的应用分析



作业：基于卷积网络的手写数字体识别



MNIST 数据集可在 <http://yann.lecun.com/exdb/mnist/> 获取, 它包含了四个部分:

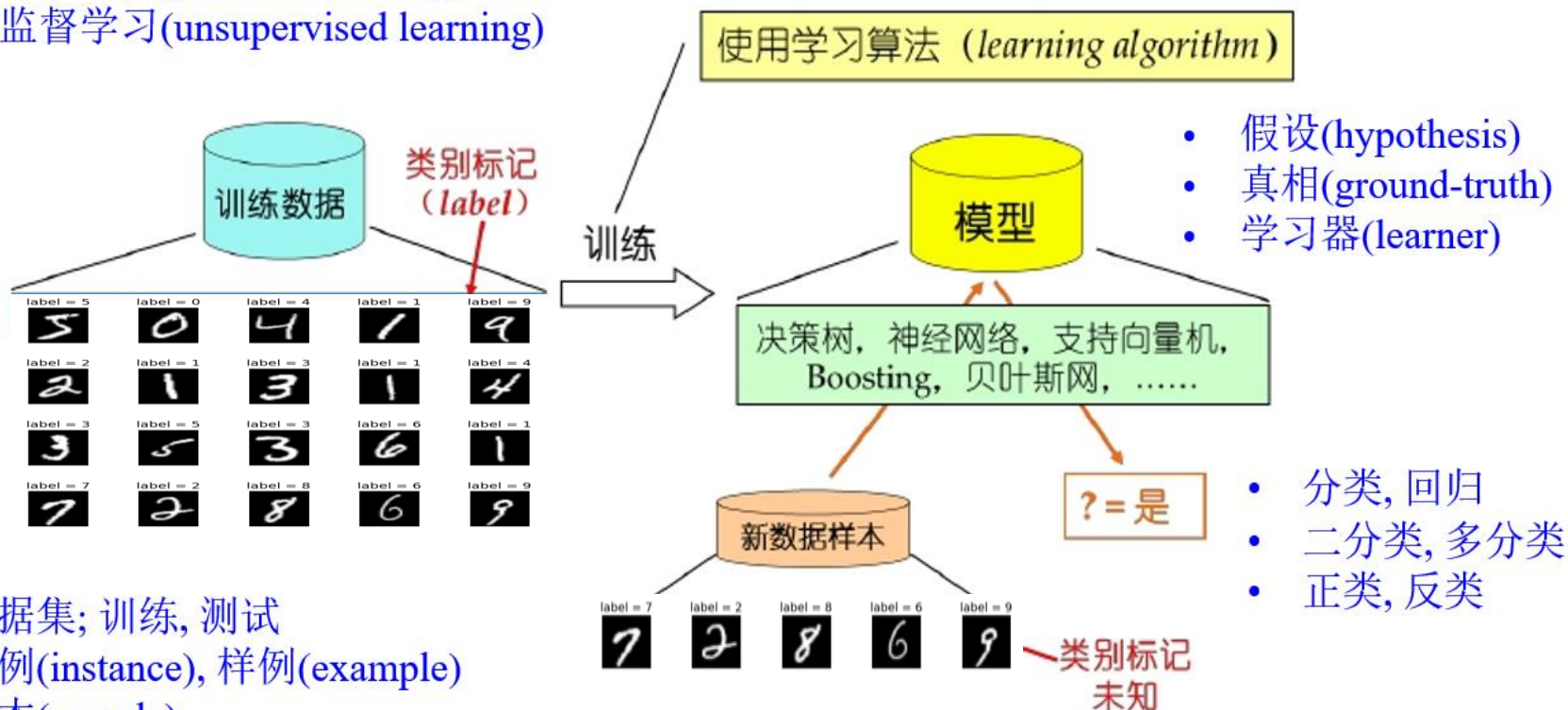
- Training set images: train-images-idx3-ubyte.gz (9.9 MB, 解压后 47 MB, 包含 60,000 个样本)
- Training set labels: train-labels-idx1-ubyte.gz (29 KB, 解压后 60 KB, 包含 60,000 个标签)
- Test set images: t10k-images-idx3-ubyte.gz (1.6 MB, 解压后 7.8 MB, 包含 10,000 个样本)
- Test set labels: t10k-labels-idx1-ubyte.gz (5KB, 解压后 10 KB, 包含 10,000 个标签)

MNIST 数据集来自美国国家标准与技术研究所, National Institute of Standards and Technology (NIST). 训练集 (training set) 由来自 250 个不同人手写的数字构成, 其中 50% 是高中学生, 50% 来自人口普查局 (the Census Bureau) 的工作人员. 测试集 (test set) 也是同样比例的手写数字数据.

基本术语



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




- 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)



表 1.1 西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

(色泽=?) \wedge (根蒂=?) \wedge (敲声=?)  好瓜

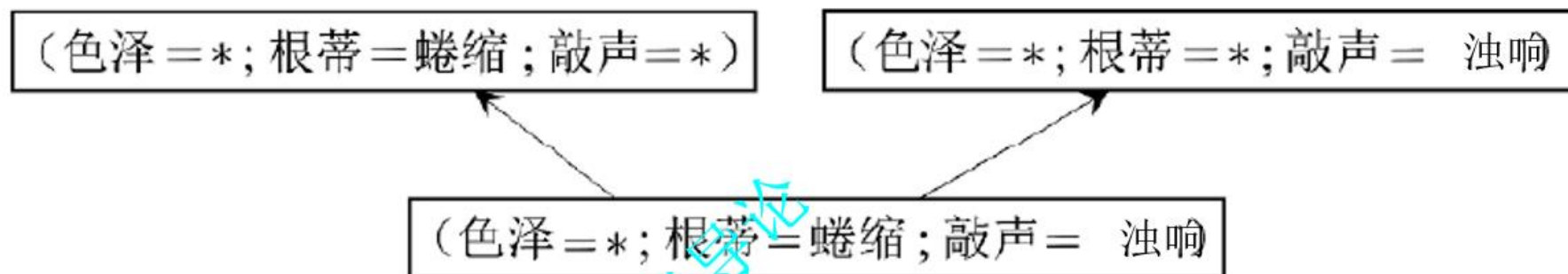
学习过程 \rightarrow 在所有假设(hypothesis)组成的空间中进行搜索的过程

目标: 找到与训练集“匹配”(fit)的假设

假设空间的大小: $n_1 \times n_2 \times n_3 + 1$



版本空间(version space): 与训练集一致的假设集合



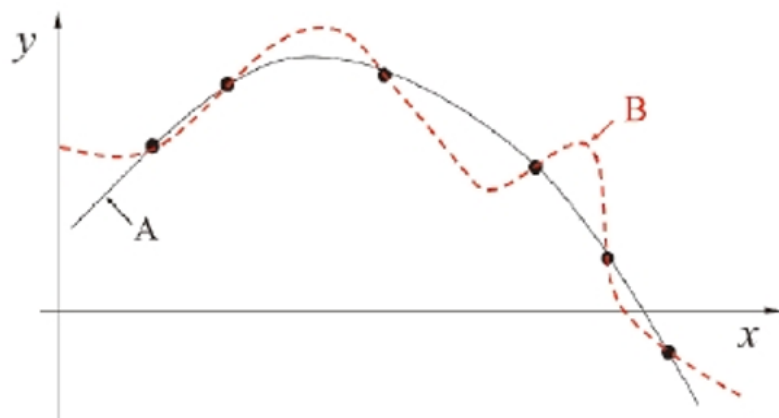
在面临新样本时, 会产生不同的输出

应该采用哪一个
模型(假设)?

例如: (青绿; 蜷缩; 沉闷)



机器学习算法在学习过程中对某种类型假设的偏好



A更好?
B更好?

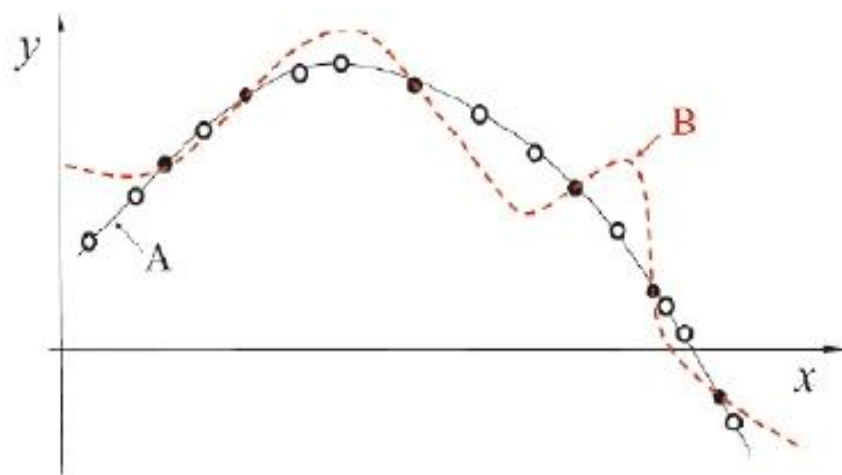
一般原则:



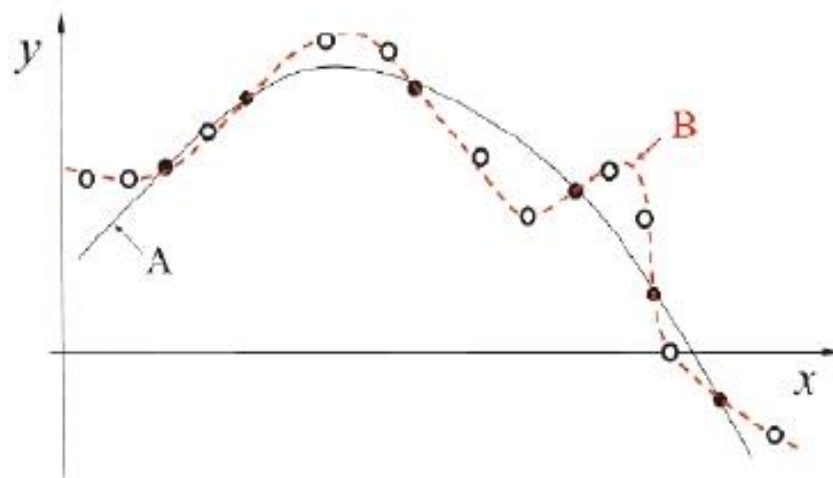
任何一个有效的机器学习算法必有其偏好

学习算法的归纳偏好是否与问题本身匹配，
大多数时候直接决定了算法能否取得好的性能！

没有免费的午餐定理(No Free Lunch Theorem)



(a) A 优于 B



(b) B 优于 A

图 1.4 没有免费的午餐. (黑点: 训练样本; 白点: 测试样本)

NFL定理: 一个算法 \mathcal{L}_a 若在某些问题上比另一个算法 \mathcal{L}_b 好, 必存在另一些问题, \mathcal{L}_b 比 \mathcal{L}_a 好。



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





泛化误差：在“未来”样本上的误差

经验误差：在训练集上的误差，亦称“训练误差”

- 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)

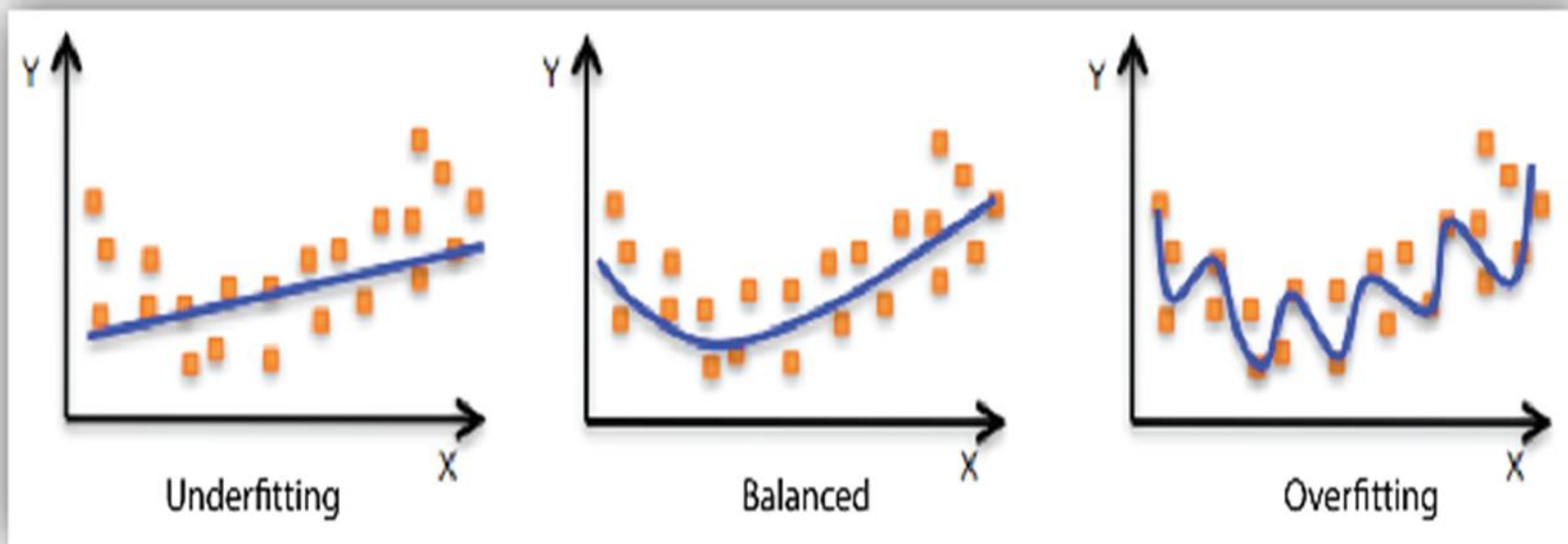
拟合: fitting



过拟合: overfitting

拟合: fitting

欠拟合: underfitting



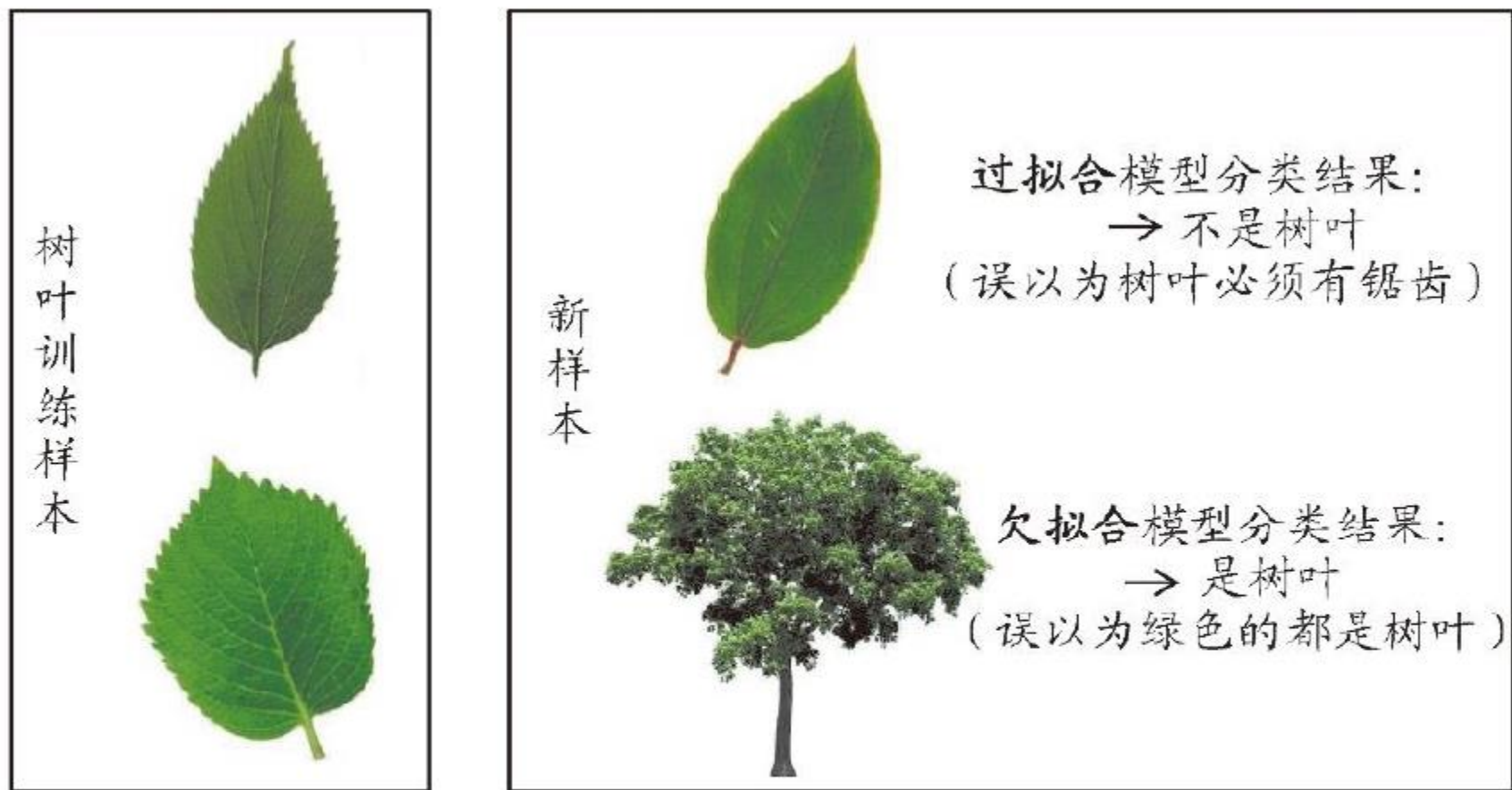


图 2.1 过拟合、欠拟合的直观类比



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

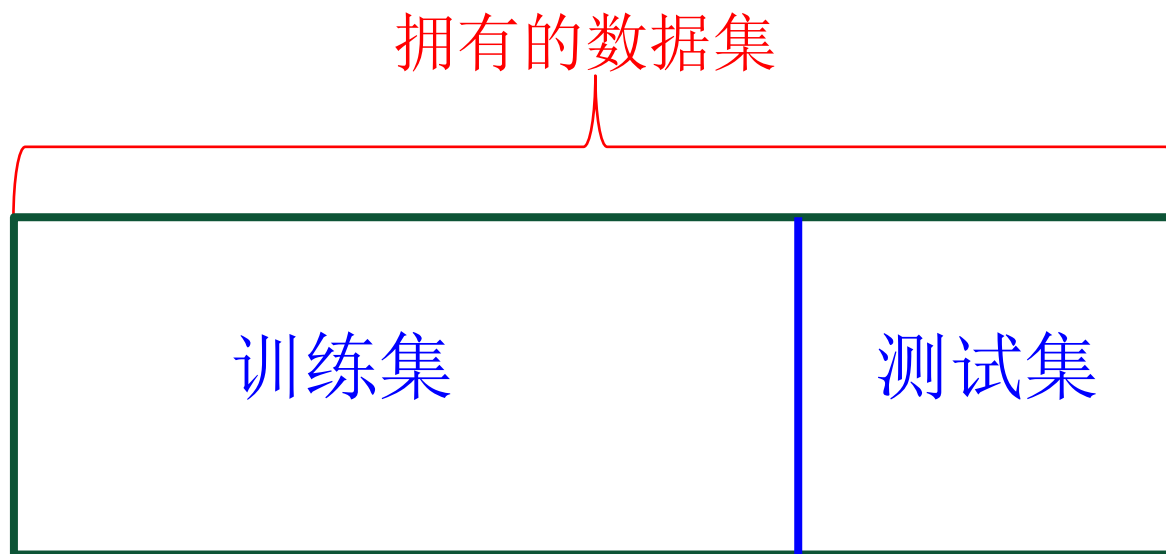


首先：怎么获得“测试集” (test set) ？

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)



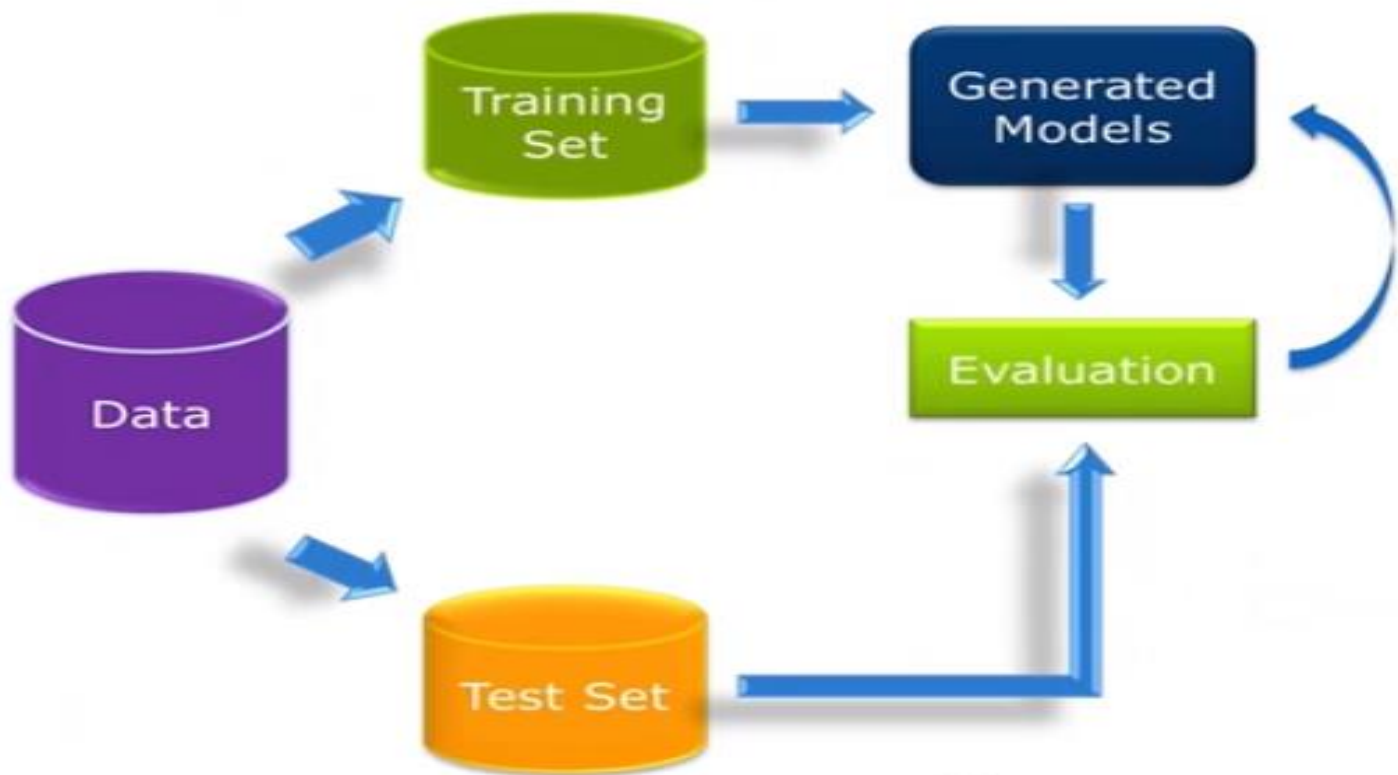
注意：

- 保持数据分布一致性（例如：分层采样）
- 多次重复划分（例如：100次随机划分）
- 测试集不能太大、不能太小（例如：1/5~1/3）

交叉验证 (Cross-validation)



- 在给定的建模样本中，拿出大部分样本进行建模型，留小部分样本用刚建立的模型进行预报，并求这小部分样本的预报误差，记录它们的平方加和。



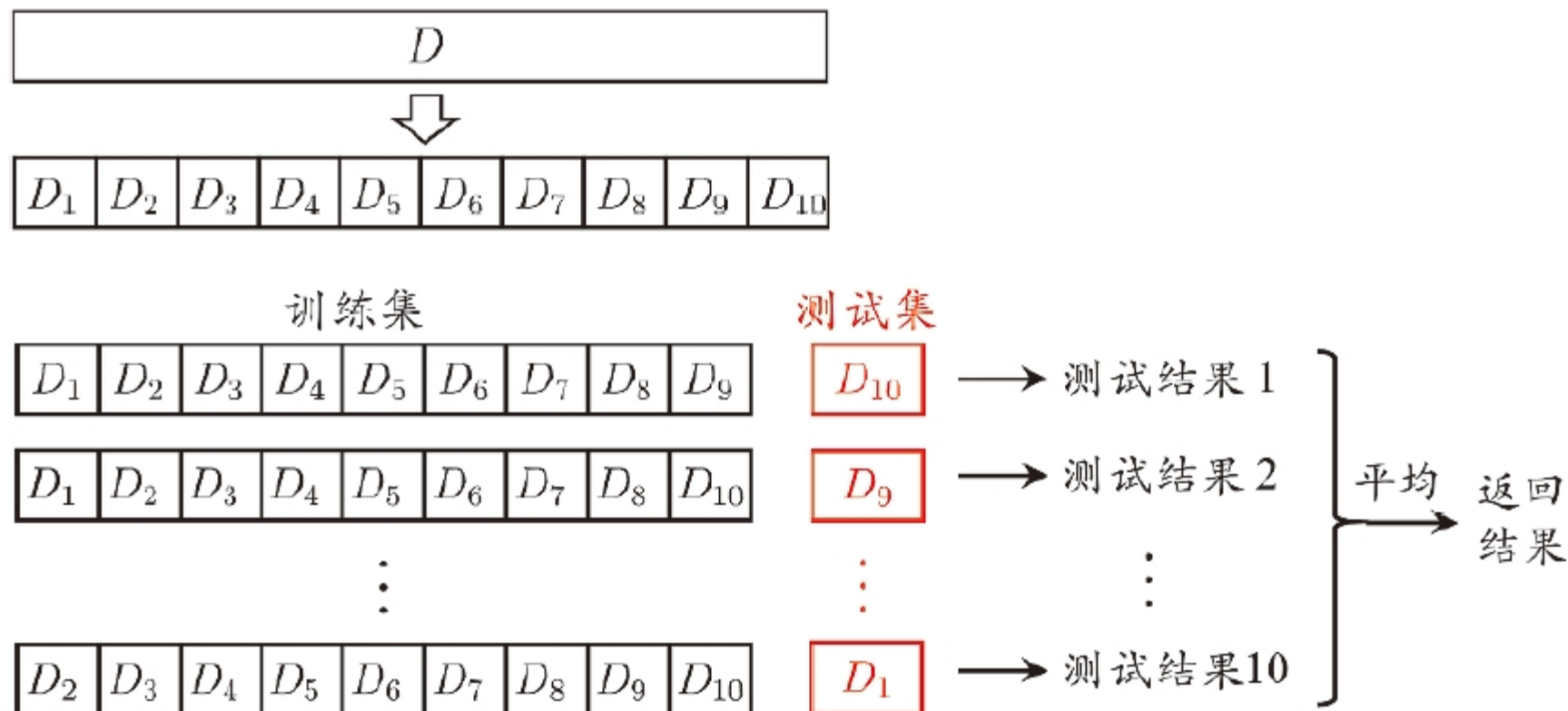
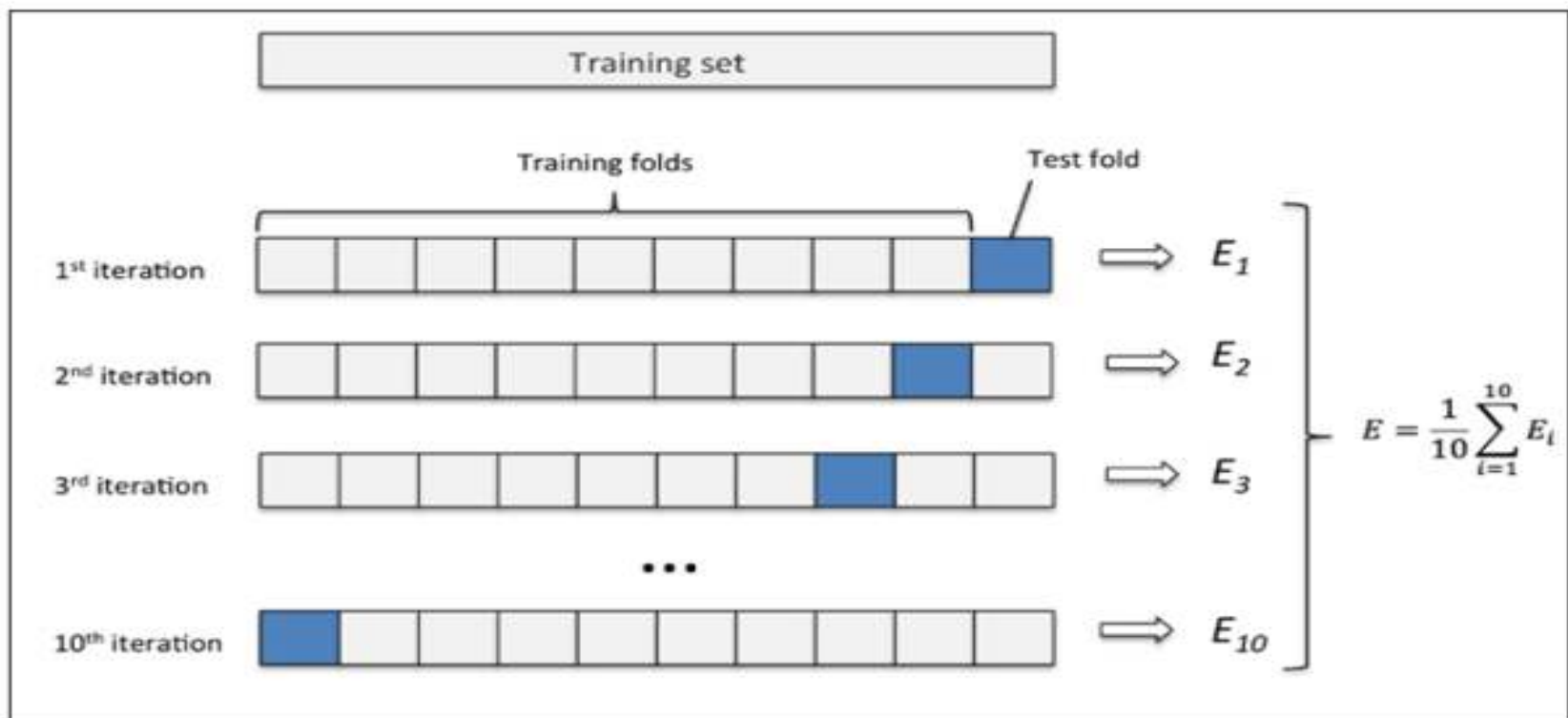


图 2.2 10 折交叉验证示意图

交叉验证 (Cross-validation)



K折交叉验证，初始采样分割成K个子样本，一个单独的子样本被保留作为验证模型的数据，其他K-1个样本用来训练。



交叉验证 (Cross-validation)



比如有数据: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6]

分为K=3组后: Fold1: [0.5, 0.2]

Fold2: [0.1, 0.3]

Fold3: [0.4, 0.6]

交叉验证时分别进行训练和测试, 每个测试集误差MSE加和平均就得到了交叉验证的总评分

Model1: Trained on Fold1 + Fold2, Tested on Fold3

Model2: Trained on Fold2 + Fold3, Tested on Fold1

Model3: Trained on Fold1 + Fold3, Tested on Fold2



10折交叉验证(10-fold cross validation), 将数据集分成**10**份, 轮流将其中**9**份做训练**1**份做验证, **10**次结果的均值作为对算法精度的估计, 一般还需要进行多次**10**折交叉验证求均值, 例如: **10**次**10**折交叉验证, 以求更精确一点。

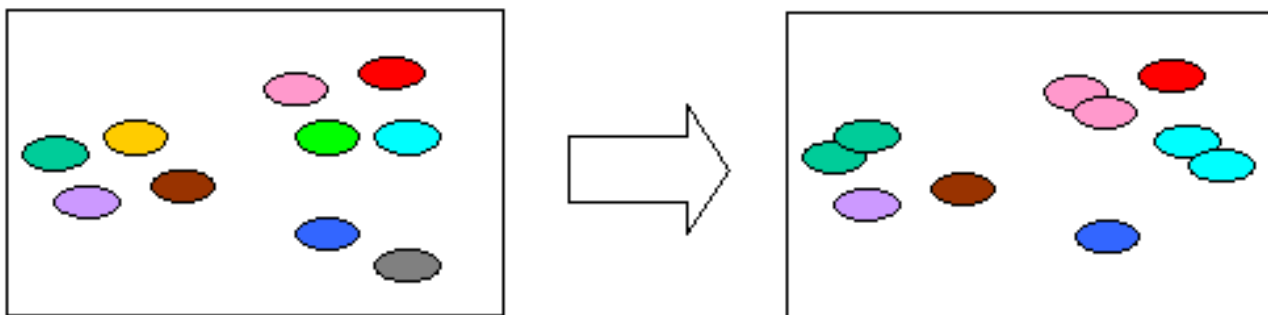
交叉验证重复**K**次, 每个子样本验证一次, 平均**K**次的结果或者使用其它结合方式, 最终得到一个单一估测。

这个方法的优势在于, 同时重复运用随机产生的子样本进行训练和验证, 每次的结果验证一次。

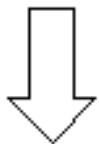


基于“自助采样” (bootstrap sampling)

亦称“有放回采样”、“可重复采样”



约有 36.8% 的样本不出现



“包外估计” (out-of-bag estimation)

➤ 训练集与原样本集同规模

➤ 数据分布有所改变



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(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

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

□ 回归(regression) 任务常用均方误差：

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



- 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。



- 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



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		真实值	
		猫	不是猫
预测值	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 https://blog.csdn.net/Orange_Spotty_Cat



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

Precision (猫) = $10/18 = 55.6\%$

Recall (猫) = $10/13 = 76.9\%$

Specificity (猫) = $45/53 = 84.9\%$



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

F1分数（F1 Score），F1分数（Score），又称平衡F1分数（balanced F Score），它被定义为精确率和召回率的调和平均数。



F_1 分数（**Score**），又称平衡 F_1 分数（**balanced F Score**），它被定义为精确率和召回率的调和平均数。

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

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。

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

混淆矩阵



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

对猫而言，

$$\text{F1-Score} = (2 * 0.769 * 0.556) / (0.769 + 0.556) \\ = 64.54\%$$



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。

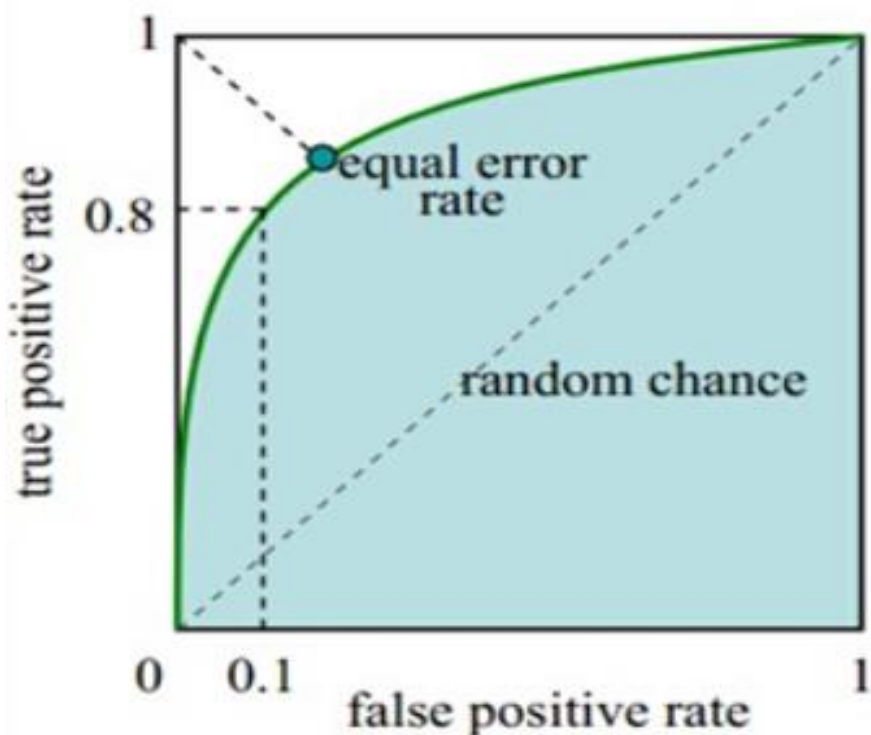
课后作业：用你熟悉的编程语言，编写一个2*2的混淆矩阵计算器，并计算图中猪的F1分数。



- 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。



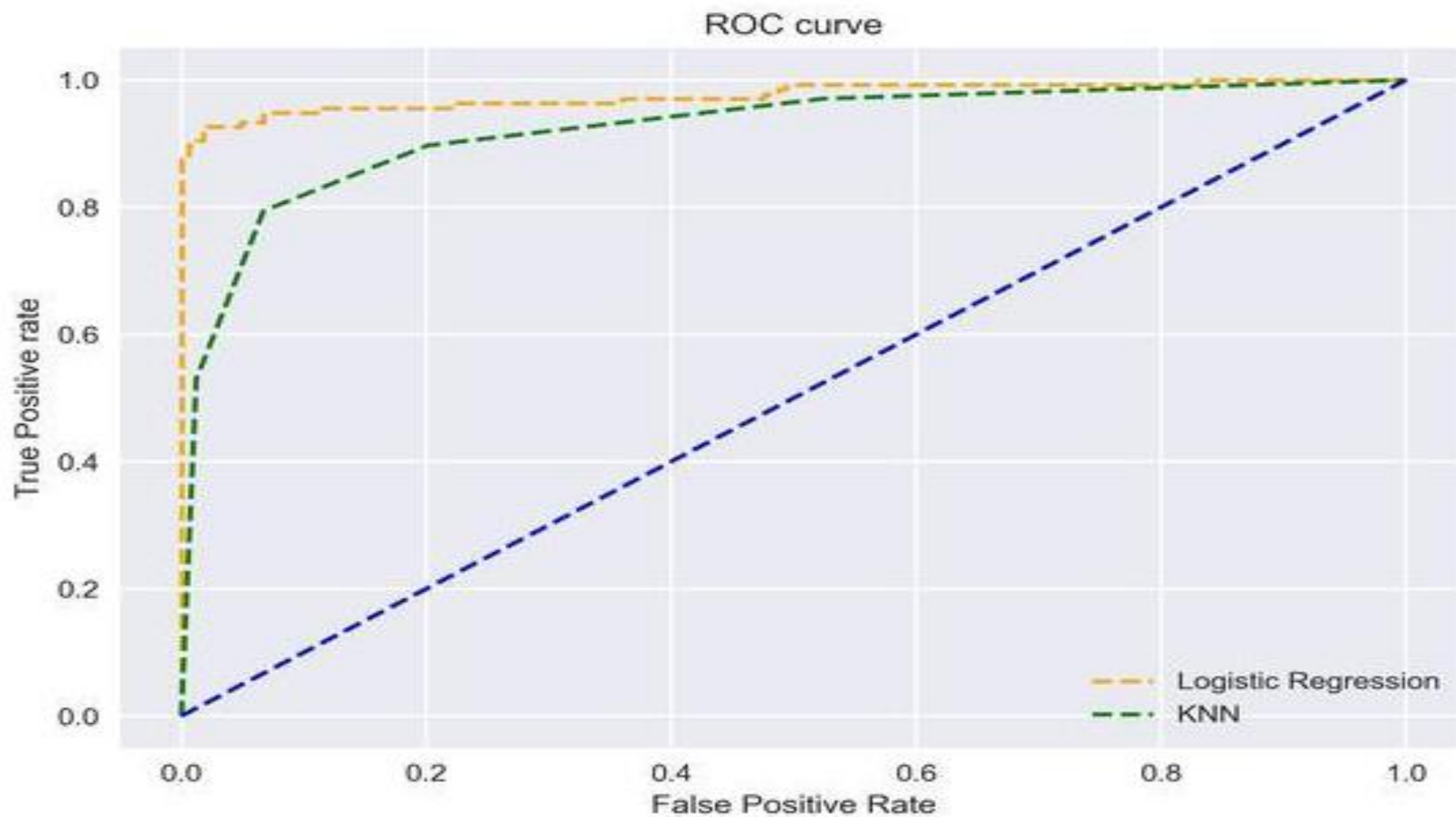
ROC曲线：是一个概率曲线，在不同的阈值下绘制TPR与FPR的关系图，从本质上把“信号”与“噪声”分开。



纵轴TPR: recall(正类覆盖率), TPR越大, 预测正类中实际正类越多。

横轴FPR: $1 - \text{TNR}$, $1 - \text{recall}$, FPR越大, 预测正类中实际负类越多。

第一章 绪论





- 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。



- 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。

谢谢！