人工智能、机器学习和深度学习的关系

阿里云

- **一个目标**
  - 三者统一：给机器赋予人的智能，让机器能够像人一样地思考问题，做出决策；
- **两种途径**
  - 机器学习是实现人工智能的一种途径，让机器使用算法解析数据、从中学习数据特征，并进行归纳判断；
  - 深度学习是机器学习的一类重要方法，采用多层非线性函数（即神经网络）学习数据特征，并进行判断，属于机器学习解决图像、语音、文本等领域问题的一个重要分支。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

机器学习定义与方向

阿里云

- **机器学习**
  - 通过技术的手段，利用已有的数据（经验）开发可以用来对新数据进行预测的模型；
  - 主要研究能产生模型的算法。
- **主要方向**

基于学习方式的划分

根据学习的输入数据是否需要标注进行划分

基于学习策略的划分

根据学习策略是否基于经典数学原理还是模拟人脑感知进行划分

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

基于学习方式的划分

阿里云

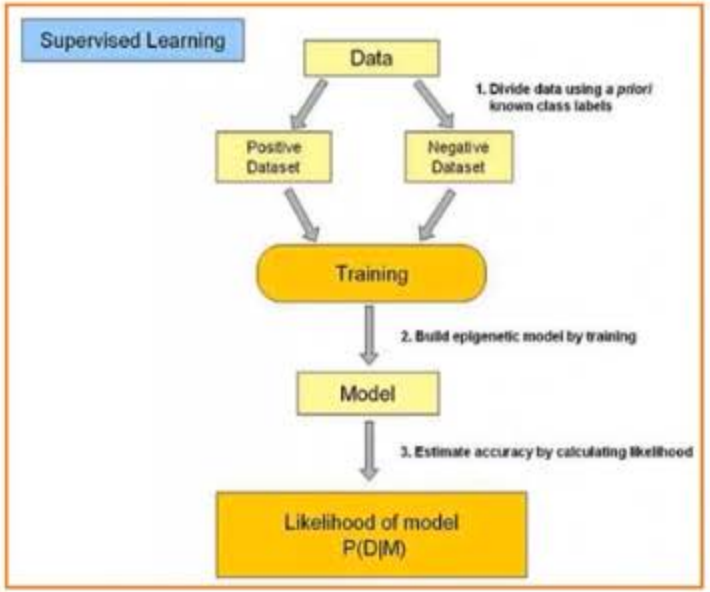
- 机器学习类别

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

有监督学习

阿里云

- **输入数据**
  - 为“训练数据”，每组训练数据有明确标识；
- **学习过程**
  - 将预测结果与“训练数据”的实际结果进行比较，不断调整预测模型，直到模型预测结果达到一个预期的准确率；
- **应用场景**
  - 分类、回归。

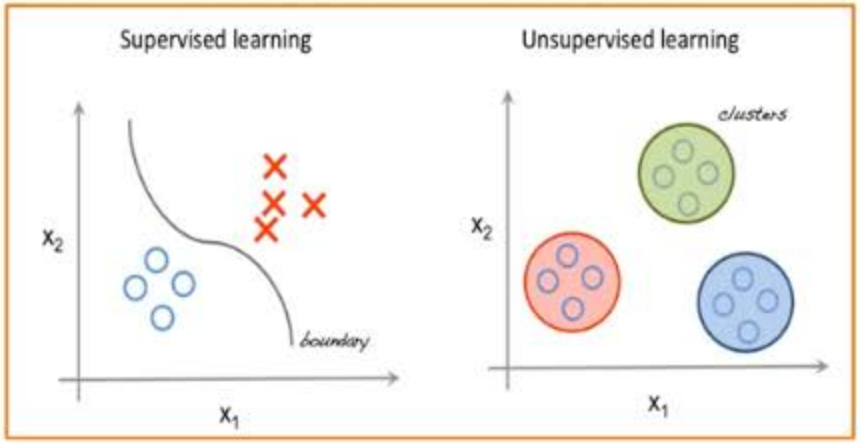


阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

无监督学习

阿里云

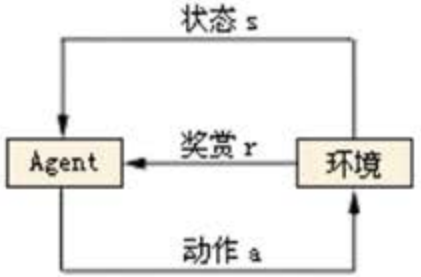
- **输入数据**
  - 数据没有被明确标识；
- **学习过程**
  - 不存在目标变量，基于数据本身去识别变量之间内在的模式和特征；
- **应用场景**
  - 关联分析、聚类。



强化学习

阿里云

- **一种机器学习方式**
  - 以“试错”的方式进行学习，通过与环境进行交互获得的奖赏指导行为，目标是使智能体获得最大的奖赏或实现特定目标；
- **输入数据**
  - 不要求预先给定任何数据，通过接收环境对动作的奖励（反馈）获得学习信息；
- **输出**
  - 模型参数调整；
- **应用领域**
  - 机器人控制、计算机视觉、自然语言处理。



基于学习策略的划分

阿里云

- **传统机器学习**
  - 基于统计、概率、线性代数等数学原理，通过分析输入数据的模式，进行判断与预测。
- **深度学习**
  - 机器学习的一个重要分支；
  - 通过模拟人类大脑感知与组织的工作方式，通过人工神经网络构建，分析输入数据，进行判断与预测。

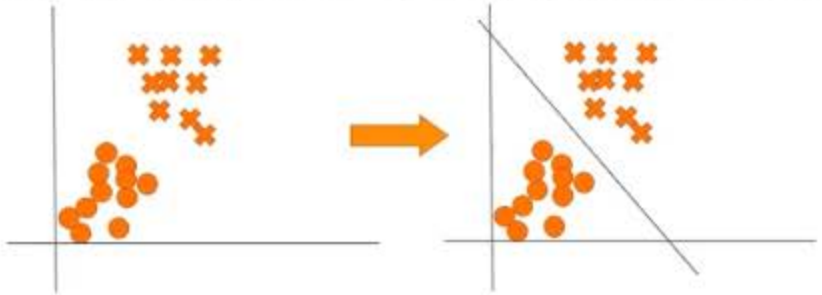
阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分类和回归的概念

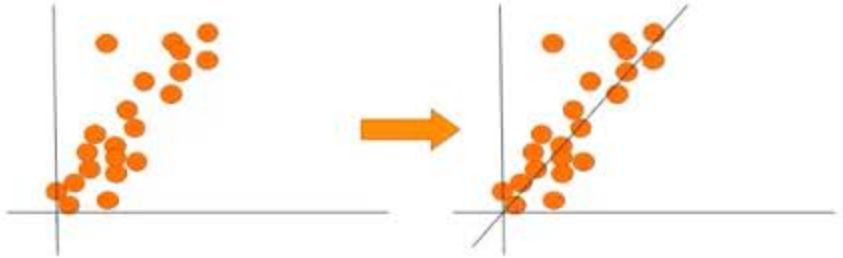
阿里云

- **定义**：分类就是将不同的类别进行分开。而回归则是找到一个空间，使得数据点尽可能的落在空间上。

- 分类：预测变量是离散的。
  - 如：今天天气为晴天



- 回归：预测变量是连续的。
  - 如：明天温度为36.4°



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分类和回归的区别

阿里云

- 分类和回归从输出和目的两方面可以很好的进行区分
- 分类：今天是晴天　回归：明天温度为36度

	分类	回归
输出不同	分类输出类别	回归输出预测值
	分类输出为离散值	回归输出为连续值
	分类输出定性值	回归输出定量值
目的不同	分类目的寻找决策边界	回归目的寻找最优拟合

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

机器学习常见函数

- 机器学习模型训练过程涉及两个重要函数
  - 损失函数
  - 优化函数

损失函数的概念

阿里云

- **定义**
  - 用来估量模型的预测值与真实值的不一致程度，是一个非负实值函数。
- **特点**
  - 损失函数的值越小，说明模型的预测值与真实值越接近；
  - 不同的算法可能使用的损失函数不同。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

优化函数的概念

- 优化函数通过调节参数使误差函数值变小
- 常见的优化函数如下：

批量梯度下降
Batch Gradient Descent

小批量梯度下降法
Min-Batch Gradient Descent

随机梯度下降法
Stochastic Gradient Descent

牛顿法
Newton Method

动量优化法
Momentum

适用性梯度算法
AdaGrad

均方根传播算法
RMSProp

AdaDelta算法
AdaDelta

Adam算法
Adam

优化函数的执行过程

- 以梯度下降法为例：
- 球要运动要到最低点需要知道三个要素
  - 所处位置
  - 移动方向
  - 移动速度

所处位置 → 损失值

移动方向 → 梯度方向

移动速度 → 学习率

3种梯度下降法特点对比

01 批量梯度下降法BGD
易收敛。每次学习使用整个样本集，学习一次的时间长。

02 随机梯度下降法SGD
每次学习使用随机单个样本，学习一次时间短。下降会出现损失函数波动且难收敛。

03 小批量梯度下降法MBGD
每次学习使用小批量样本集，结合了BGD和SGD的优点，弱化了缺点。

机器学习常见评估指标

阿里云

- 评价指标是建立在不同的机器学习任务上的，主要分为三大类：**分类、回归和无监督**

分类任务
- 混淆矩阵
- 准确率
- 精确度和召回率
- F1分数

回归任务
- 均方误差MSE
- 平均绝对误差MAE
- 均方根误差RMSE

无监督任务
- 兰德系数
- 互信息
- 轮廓系数

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分类任务常见评估指标：混淆矩阵

- **定义**：混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用n行n列的矩阵形式来表示;
- 以2分类为例，混淆矩阵如下:

	实际值	
预测值	真阳性（TP） 预测值为真，实际值为真	假阳性（FP） 预测值为真，实际值为假
	假阴性（FN） 预测值为假，实际值为真	真阴性（TN） 预测值为假，实际值为假

混淆矩阵示例

阿里云

实际类别：[0,0,0,0,1,1]

预测类别：[0,0,1,0,1,0]

	实际值	
预测值	真阳性（TP） 预测值为1，实际值为1	假阳性（FP） 预测值为1，实际值为0
	假阴性（FN） 预测值为0，实际值为1	真阴性（TN） 预测值为0，实际值为0

$$混淆矩阵=\begin{bmatrix}1 & 1\\ 1 & 3\end{bmatrix}$$

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

分类任务常见评估指标：其它

准确率

- 衡量分类正确的比例。

精确度

- 又叫查准率，指被预测为正样本的样本中预测正确的占比。

召回率

- 又叫查全率，指被正确检测出来的真实样本占所有真实样本的比例。

F1分数

- 权衡精确度和召回率。

01 02 03 04

	实际值	
预测值	真阳性(TP)=1	假阳性(FP)=1
	假阴性(FN)=1	真阴性(TN)=3

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

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

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

$$F = 2 \cdot \frac{\text{Precision} \cdot Recall}{\text{Precision} + Recall}$$

回归任务常见评估指标

阿里云

均方误差MSE

- 又叫L2范数损失，通过计算真实值和预测值的差值的平方和的均值衡量。

平均绝对误差MAE

- 又叫L1范数损失，通过计算真实值和预测值的差值的绝对值的均值衡量。

均方根误差RMSE

- 表示预测值和真实值差值的样本标准差。

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i)^2 = \frac{1}{2}\left((1.3 - 1.1)^2 + (2.2 - 2.3)^2\right) = 0.025$$

实际值：[1.3, 2.2]

预测值：[1.1, 2,3]

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|f(x_i) - y_i'| = \frac{1}{2}(|1.3 - 1.1| + |2.2 - 2.3|) = 0.15$$

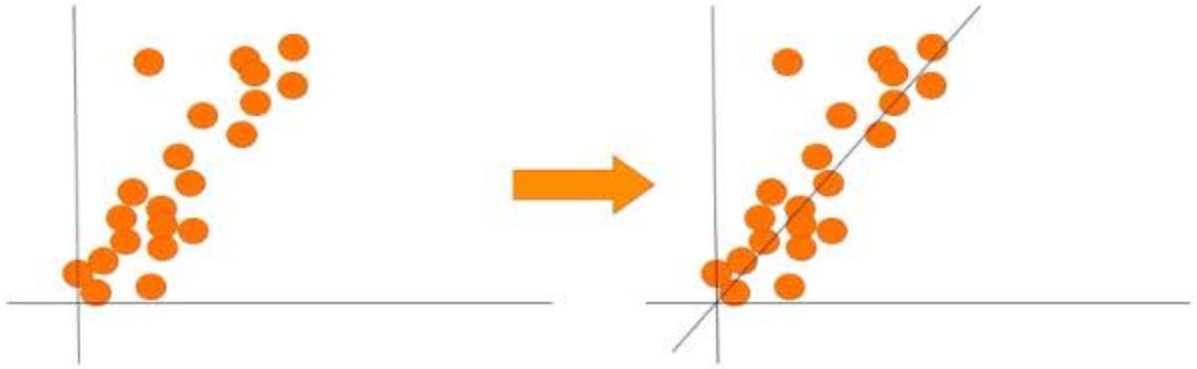
$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(f(x_i) - y_i')^2} = \sqrt{\frac{1}{2}\left((1.3 - 1.1)^2 + (2.2 - 2.3)^2\right)} = 0.5$$

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

线性回归算法定义和任务类型

阿里云

- **定义**：线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法；
- **任务类型**：回归。
- **应用场景**：异常指标监控 农业贷款预测

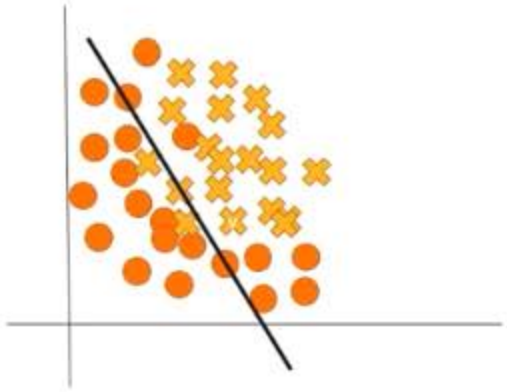


阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

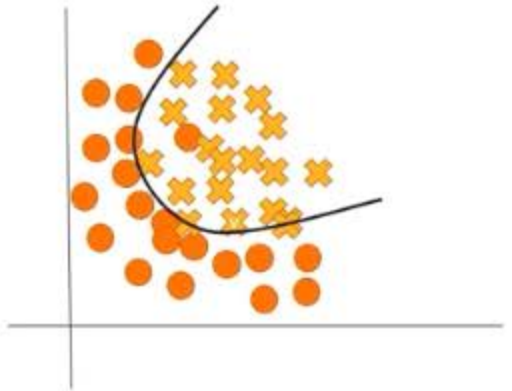
过拟合与欠拟合

阿里云

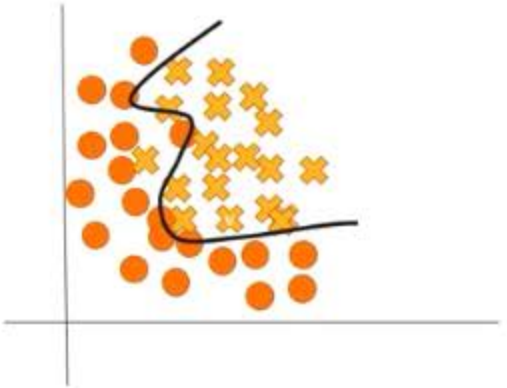
- **定义**：过拟合与欠拟合用来度量模型泛化能力的直观表现；
- **欠拟合**：模型在训练集、测试集上均表现不佳的情况；
- **过拟合**：在训练集上表现很好，到了验证和测试阶段就很差。



欠拟合



正常拟合



过拟合

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

线性回归正则化模型

阿里云

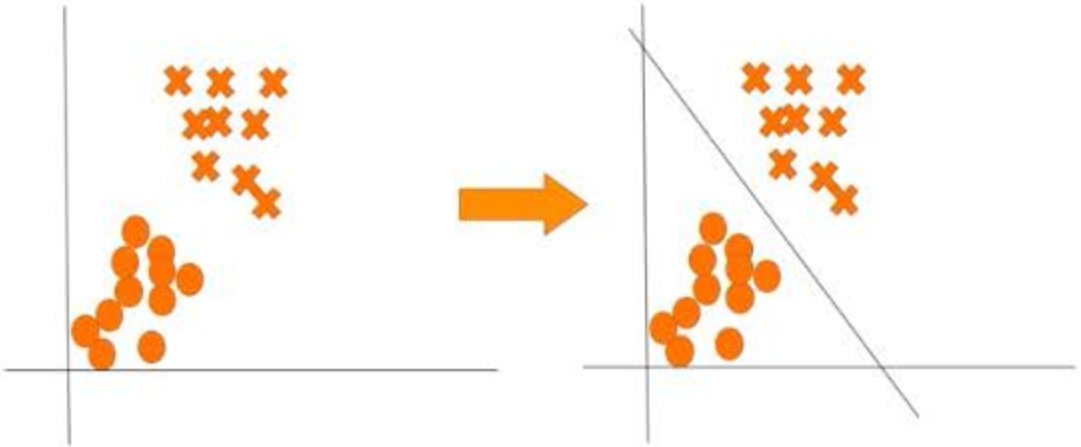
- 正则化能有效的防止过拟合现象
- 根据正则化的选择，线性回归正则化模型有3种

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

逻辑回归算法定义和任务类型

阿里云

- **定义**：逻辑回归是一种广义线性回归，在线性回归的基础上添加非线性变换，使得逻辑回归输出值为离散型
- **任务类型**：分类
- **应用场景**：学生考试成绩预测　雾霾天气预测



逻辑回归多分类应用

阿里云

- 逻辑回归常用于二分类，根据策略不同，可以将逻辑回归用于多分类任务

一对多法
One-Vs-Rest

一对一法
One-Vs-One

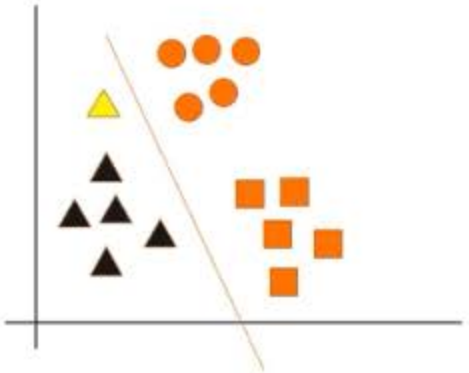
Softmax法
Softmax.

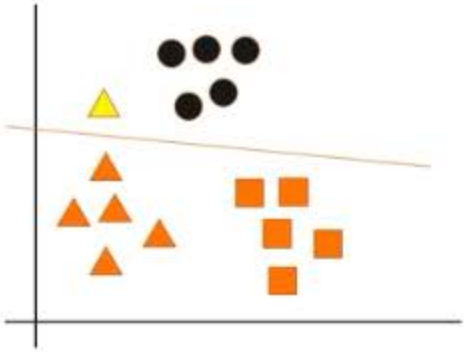
阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

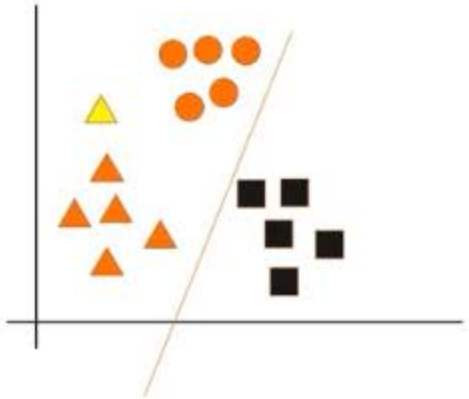
一对多法

- 对K分类，训练时依次把某个类别的样本归为一类，其他剩余的样本归为另一类，得到K个分类器；
- 预测时分别用K个分类器进行预测，选择结果最大的作为分类的结果。

- 优点：普适性还比较广，效率较高；
- 缺点：易造成数据不平衡。



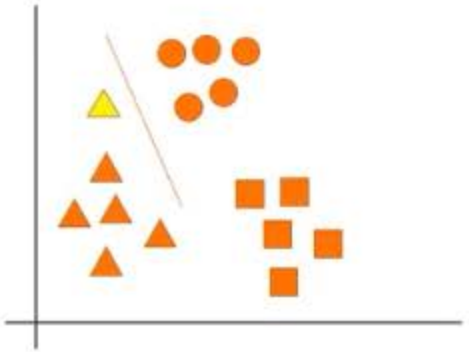


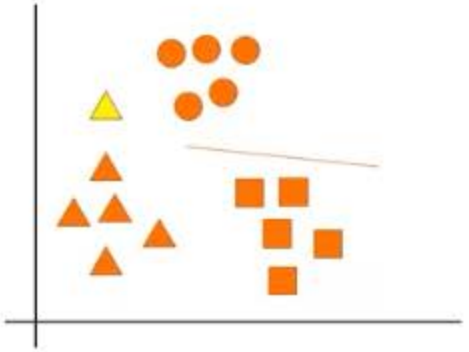


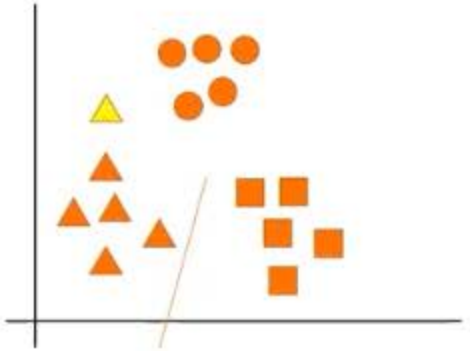
一对一法

- 对K分类，训练时依次让不同类别数据两两组合训练，得到 $\frac{K(K-1)}{2}$ 个二分类模型
- 预测时分别用二分类器进行预测，最后得票最多的类别即为未知样本的类别

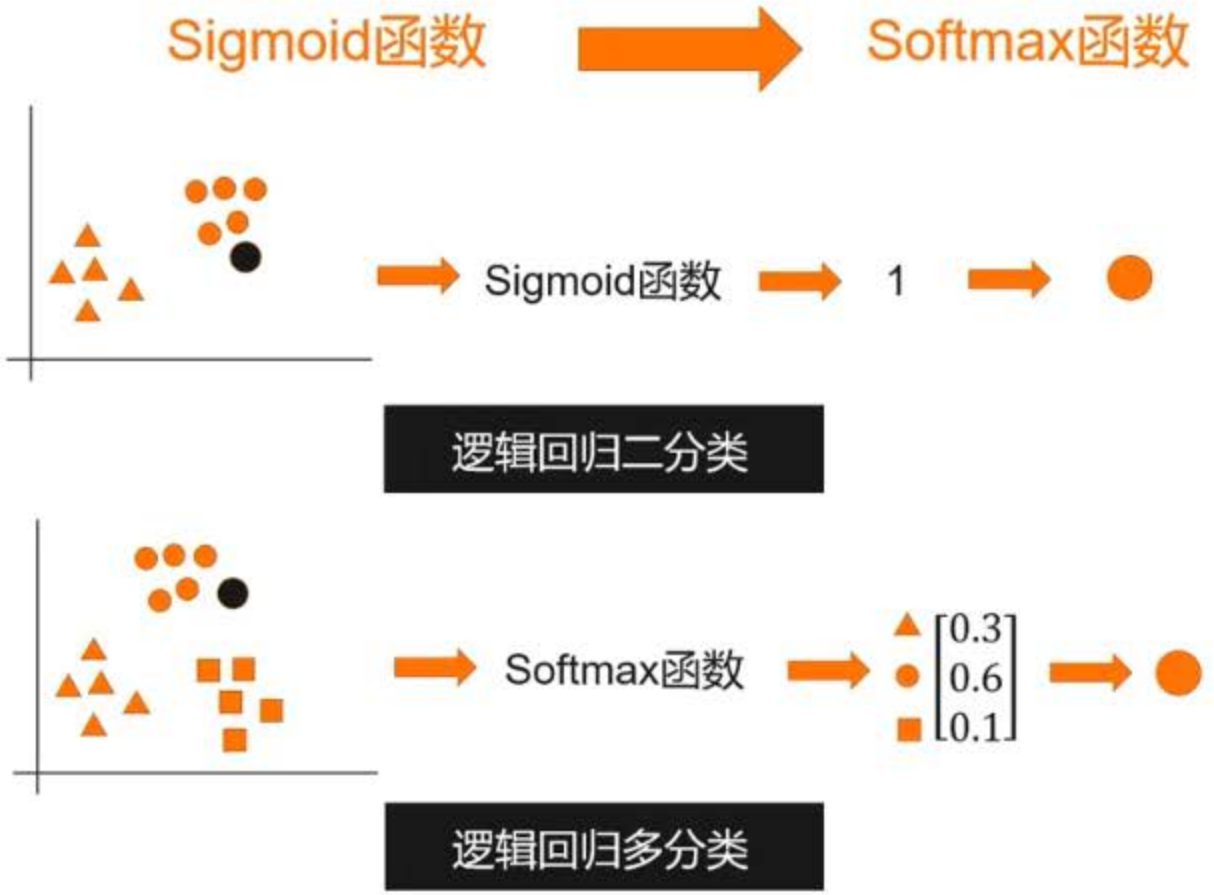
- 优点：一定程度规避数据不平衡情况，性能相对稳定，训练效率提高
- 缺点：训练的二分类模型更多，影响预测时间







Softmax法



朴素贝叶斯算法介绍

阿里云

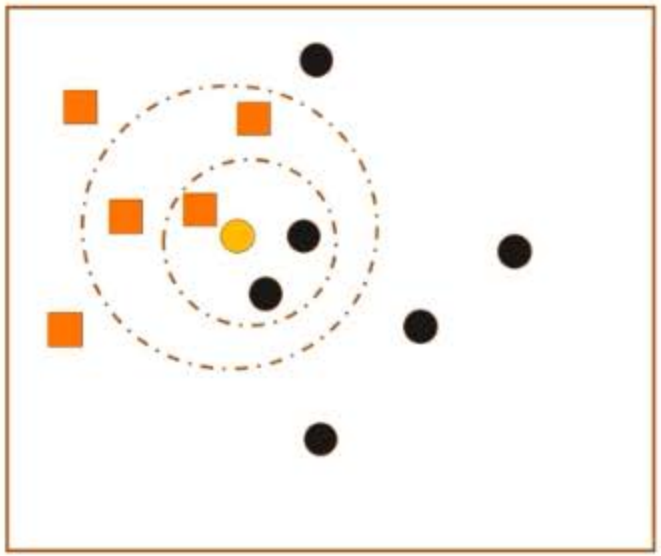
- **定义**：朴素贝叶斯法（Naive Bayes model）是基于贝叶斯定理与特征条件独立假设的分类方法
- **任务类型**：分类
- **应用场景**：垃圾邮件分类 舆情分析

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

K近邻算法介绍

阿里云

- **定义**：K近邻即从训练集中找到与新实例最近的K个实例，根据K个实例来进行预测
- **任务类型**：分类、回归
- **应用场景**：约会匹配 商品推荐



01 K近邻分类

K=3，新实例属于圆

K=5，新实例属于正方形

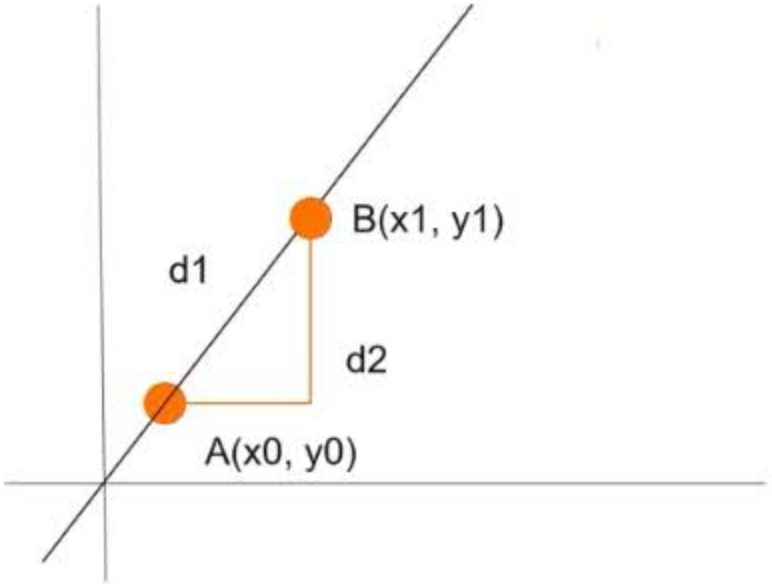
02 K近邻回归

K=3，新实例预测值为最近3个实例的平均值

K=5，新实例预测值为最近5个实例的平均值

K近邻算法：距离度量

- 特征空间中两个实例点之间的距离是二者相似程度的反映；
- K近邻算法通过距离来寻找离新实例最近的K个实例。



- 欧式距离：两个点的直线距离

$$d1 = \sqrt{(x1 - x0)^2 + (y1 - y0)^2}$$

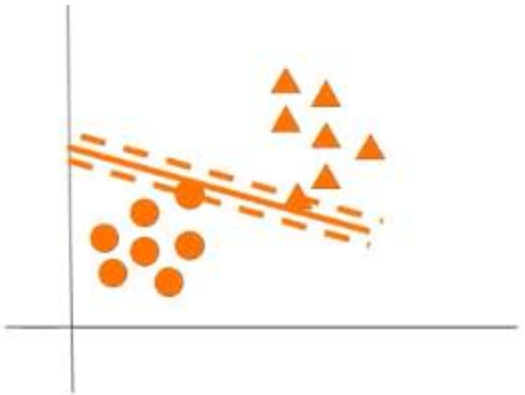
- 曼哈顿距离：两个点在标准坐标系上的绝对轴距之和

$$d2 = |x1 - x0| + |y1 - y0|$$

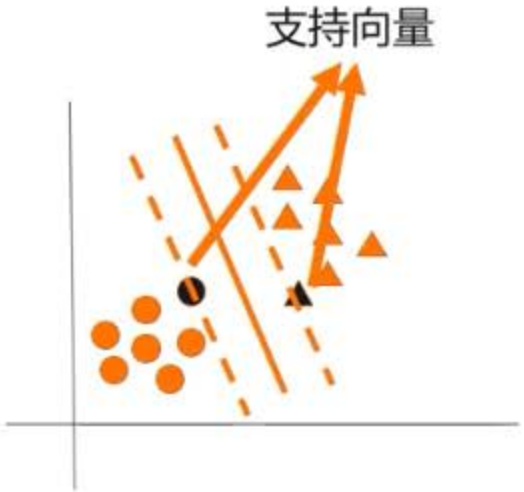
支持向量机算法介绍

阿里云

- **定义**：支持向量机是一类按监督学习方式对数据进行二元分类的广义线性分类器，其决策边界是对学习样本求解的最大边距超平面。
- **应用场景**：心脏病预测 用户窃电识别



决策边界不合理



决策边界最大化

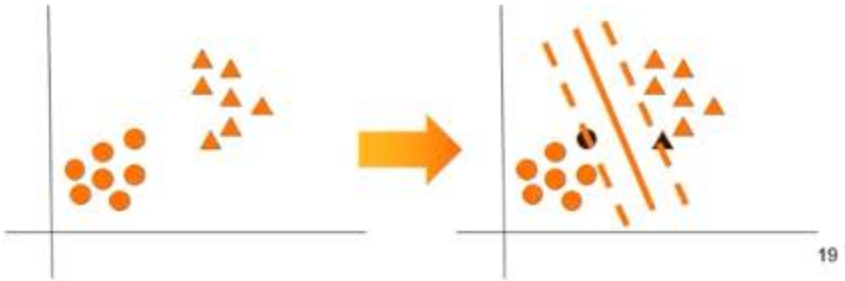
阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

线性可分与线性不可分

阿里云

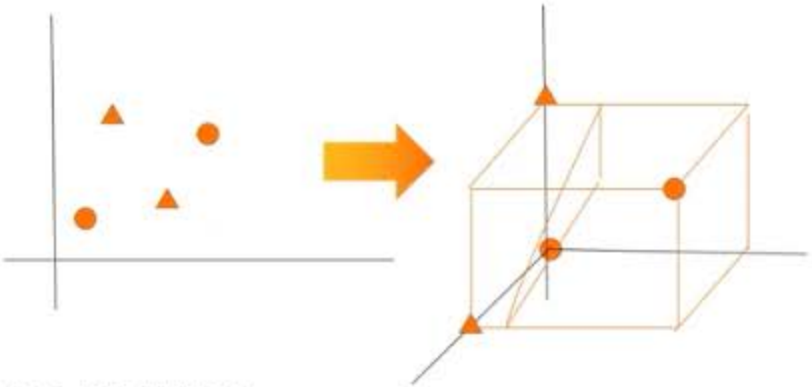
线性可分

如果样本可以直接使用一个线性函数切分，则称样本线性可分。



线性不可分

如果样本不能直接使用一个线性函数切分，则称样本线性不可分。通过升维，将低维度映射到高维度实现线性可分。



注意：为了解决支持向量机线性不可分，引入核函数的概念。

核函数概念与常用核函数介绍

阿里云

- **定义**：将数据样本升维，使低维非线性可分变为高维线性可分；
- 常见的核函数：

01 线性核函数

02 径向基核函数

03 多项式核函数

04 Sigmoid核函数

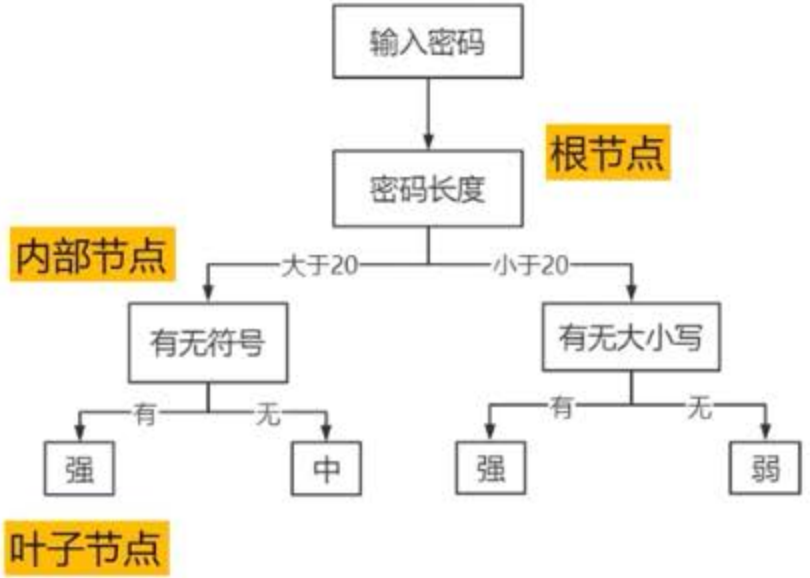
阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

决策树算法介绍

- **定义**：决策树是一种以树结构形式来表达的预测分析模型
- **类别**：分类树和回归树
- **应用场景**：银行贷款预测 动物识别

决策树结构

- 结构：决策树由节点和分支构成

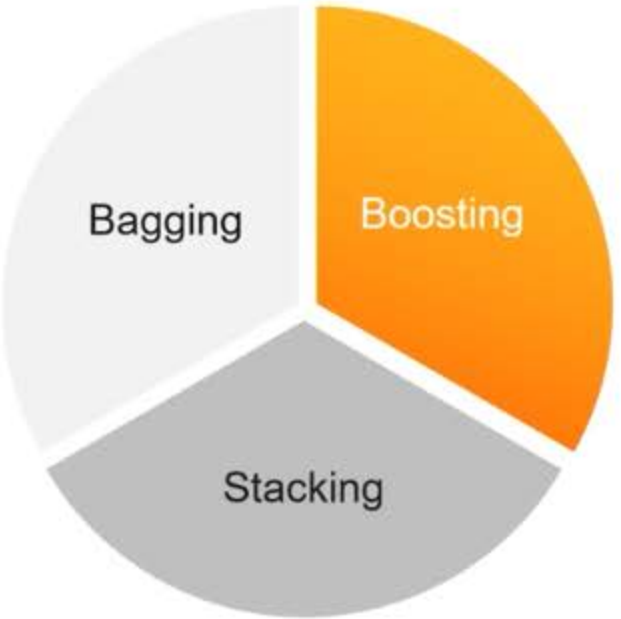


决策树构建步骤

- 决策树的构造通常有三个步骤：特征选择、决策树生成、决策树剪枝。

集成学习算法定义及流派

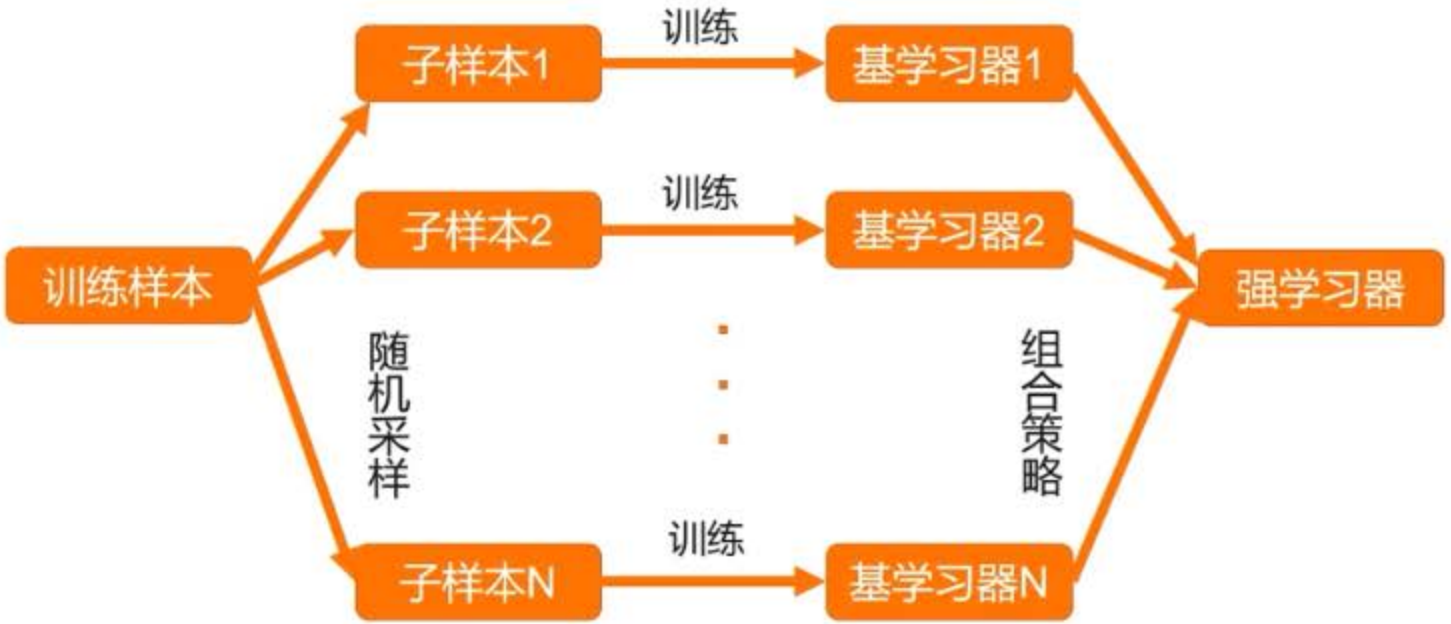
- **定义：** 通过构建和结合多个机器学习算法（基学习器）完成学习任务；
- **重要条件**：基学习器学习结果之间存在差异。
- **应用场景：** 土地覆盖测绘 恶意软件检测
- **三大流派：** Bagging、boosting、stacking



Bagging集成学习算法

阿里云

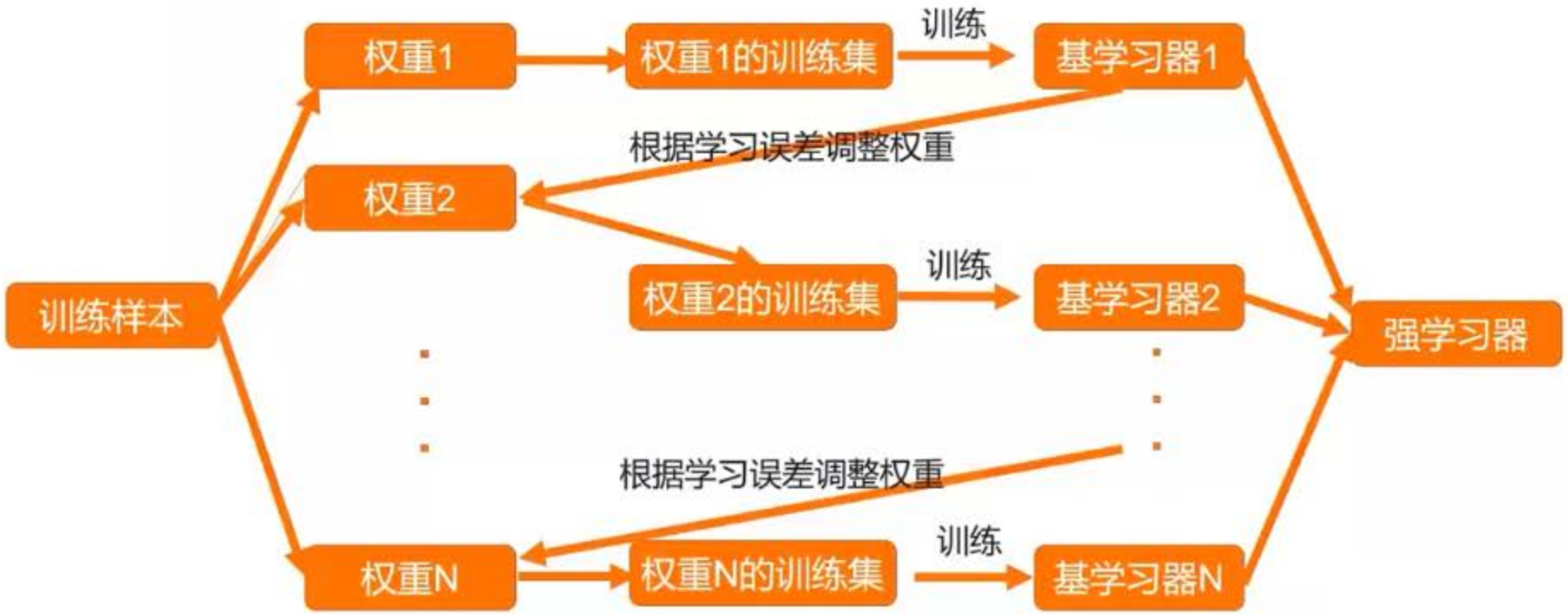
- **定义**：Bagging算法主要对样本训练集合进行随机化抽样，通过反复的抽样训练新的模型，最终在这些模型的基础上取平均。



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

Boosting集成学习算法

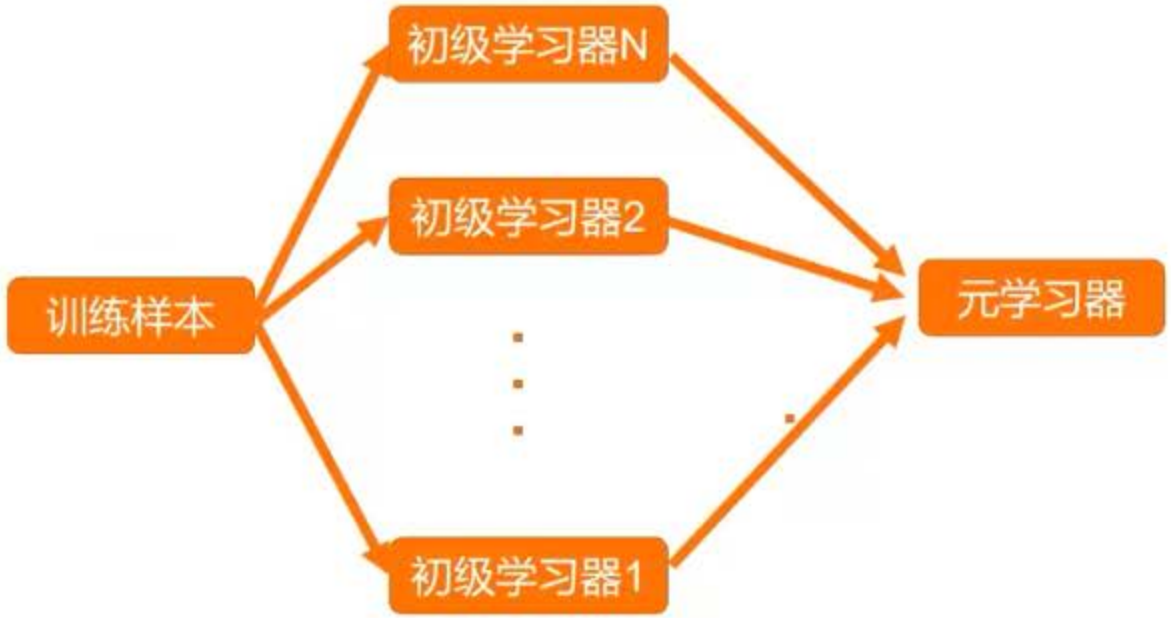
- **定义**：Boosting通过不断地使用一个弱学习器弥补前一个弱学习器的“不足”的过程，来串行地构造一个较强的学习器，这个强学习器能够使目标函数值足够小。



Stacking集成学习算法

阿里云

- **定义**：Stacking是通过一个元分类器或者元回归器来整合多个分类模型或回归模型的集成学习技术。基础模型利用整个训练集做训练，元模型将基础模型的输出作为特征进行训练。

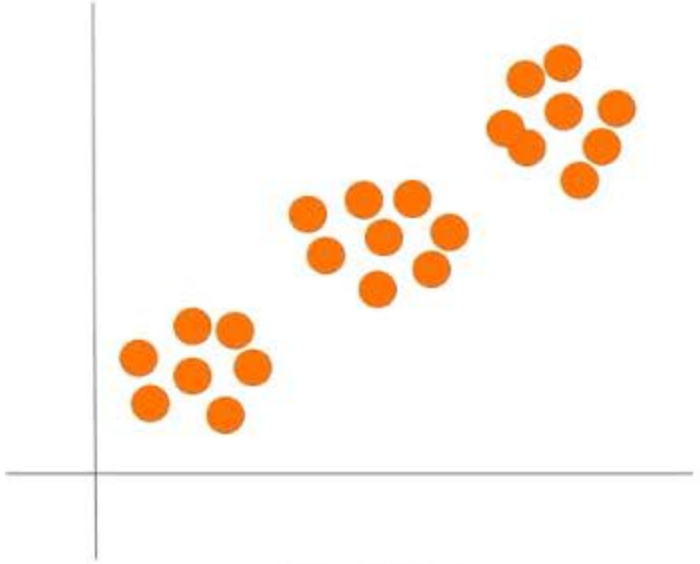


阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

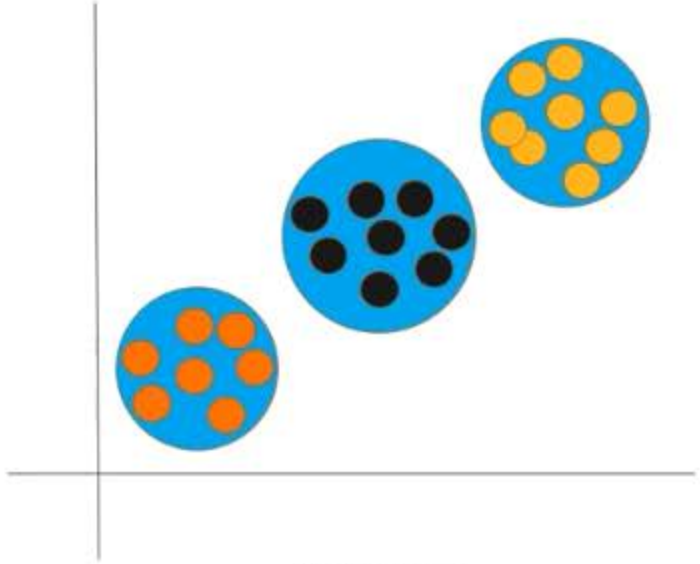
聚类算法介绍与应用

阿里云

- **定义**：聚类属于无监督学习的一种，使同一类的数据尽可能聚集到一起，不同数据尽量分离
- **应用场景**：非人恶意流量识别 新闻主题聚类



原始数据



聚类结果

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER