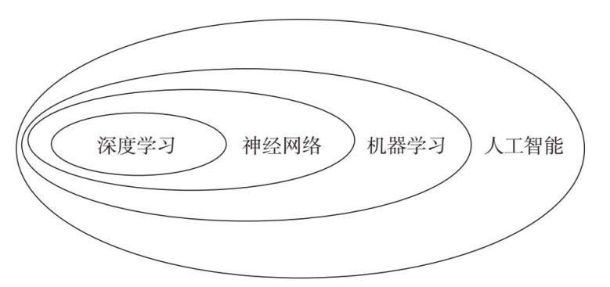
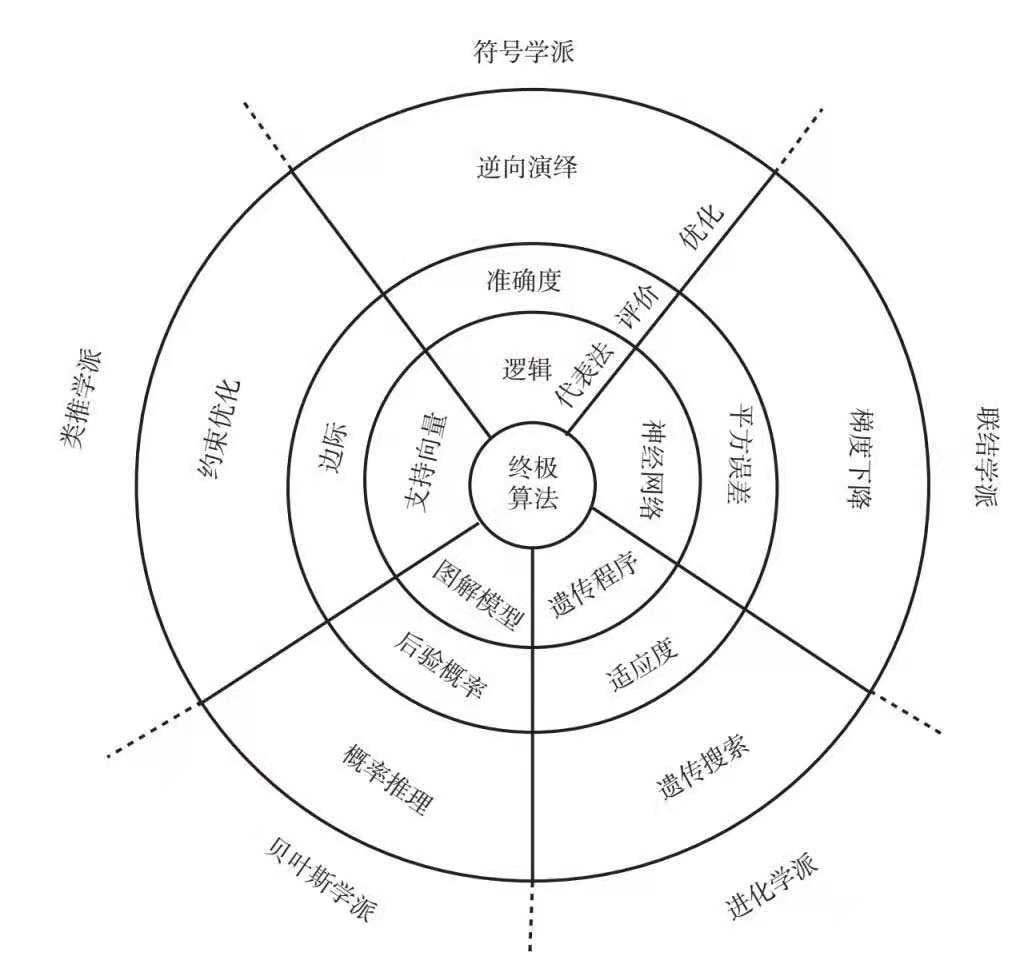
机器学习-笔记

1.机器学习介绍

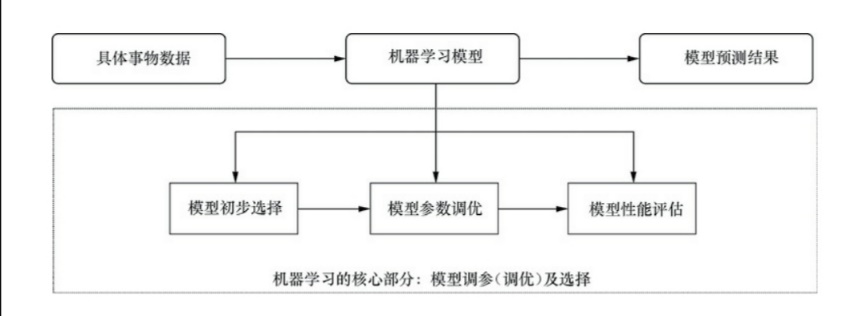
20世纪70-80年代 知识工程、专家系统 → 机器学习



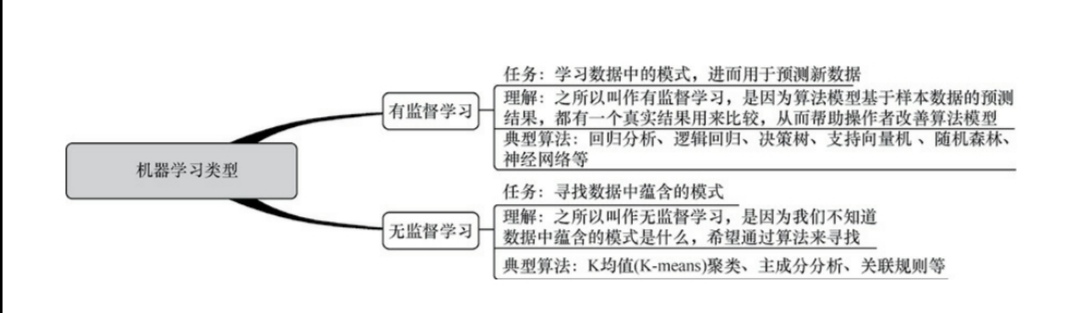
五大学派



机器学习过程：



机器学习类型：



2.机器学习的一些名词；

(X1,X2,..,X3)特征矩阵（行：样本/向量 列：特征）

\*也有 列：向量，行：特征（行数为向量的维度）

X特征向量

Y标签值

3.机器学习的任务：分类和回归

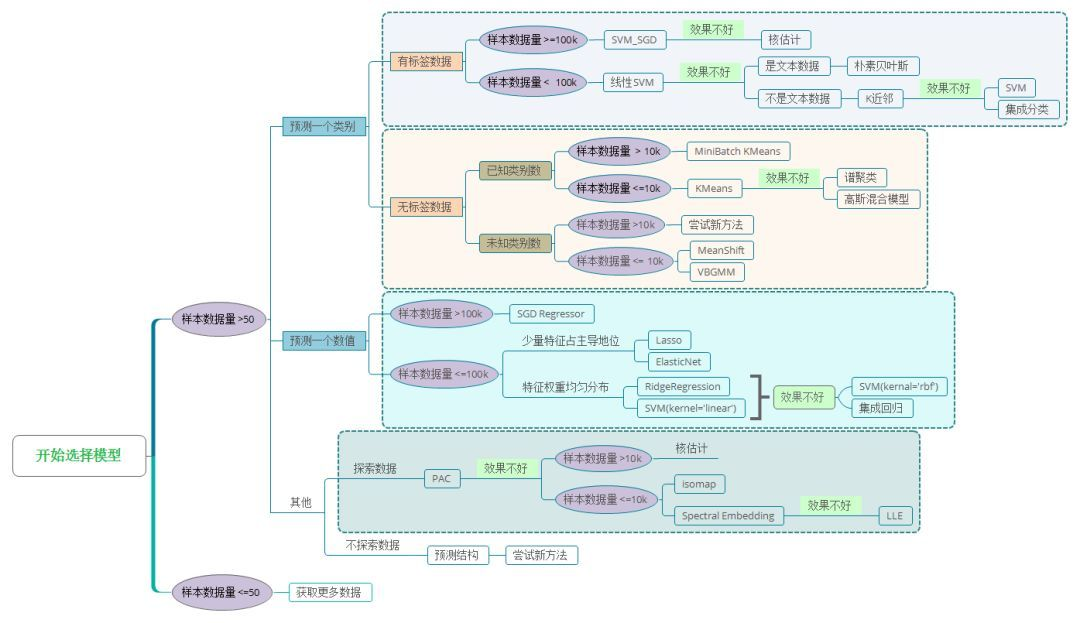
Continuous features连续型特征（数值特征）

Categorical features离散型特征（定性特征）



分类和回归的区别：是否存在“阈值判断”

\*logistic回归：分类，其输出的结果是决策边界，是不连续的



3.机器学习的主要步骤：

特征提取特征处理（特征预处理）特征选择

4.数据集dataset、训练集、测试集

在sklearn中自带的数据集中有以下几个部分：

|  |
| --- |
| data 特征数据数组，样本的特征参数 |
| target标签数组，比如黑猫、白猫 |
| DESCR数据描述 |
| featurenames特征名 |
| targetnames标签名 |

常见错误：

1）没有为模型准备足够的数据

如果你的数据集很小，模型就无法获得足够的用于泛化的辨别特征。这样模型就会过拟合数据，造成训练错误很低但测试错误很高的问题。

解决方案：收集更多数据数据增强

2）数据类的质量很低

解决方法：为你的数据类选择正确的粒度级别

3）数据的质量很低

4）数据类不均衡

解决方法：为未被充分表示的类获取更多数据样本。不过，通常会花费一定的时间和资源对被过度表示的类的数据进行过采样或降采样，在未被充分表示的类中可进行数据增强

5）数据不均衡

解决方法：标准化

6）没有验证或测试

解决方法：将数据集分割为3部分:训练集（60%)、验证集(20%）和测试集（20%)。这样能防止测试集被模型的选择过拟合。

用训练集训练模型

在验证集上测试模型，确保没有出现过拟合

选择表现最好的模型，用测试集对其测试，得到模型的真正准确率

5.经典例子

（1）鸢尾花数据集（4×150）

类型：分类

4个特征(Sepal.Length (花尊长度) 、Sepal.Width (花雩宽度)、Petal.Length(花瓣长度)、Petal.Width(花瓣宽度)），特征值都为正浮点数，单位为厘米。目标值为鸢尾花的分类(lris Setosa (山鸢尾)、lris Versicolour(杂色鸢尾)，lris Virginica(维吉尼亚鸢尾)）。通过决策树算法构建起对其分类的模型。