大数据基础与应用
BIG DATA FOUNDATION AND APPLICATION

数据分析算法

北京理工大学计算机学院　孙新

2019年1月

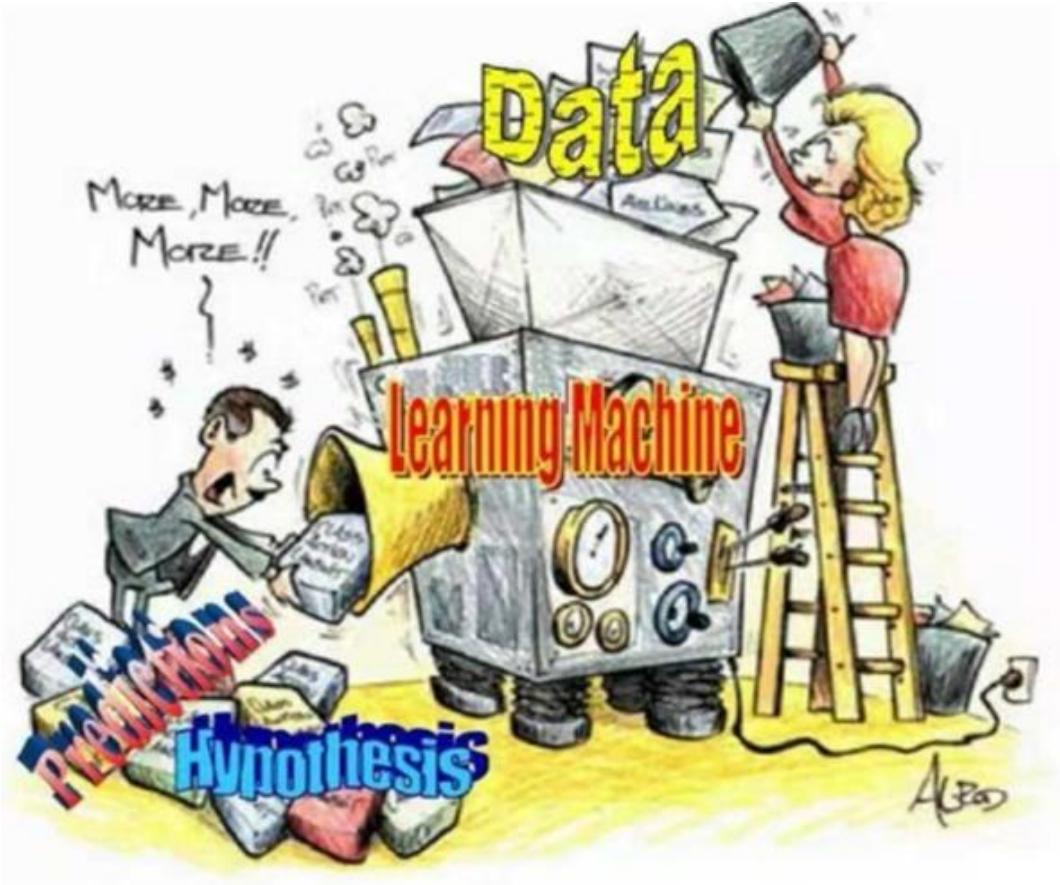
目　录

3.1 什么是机器学习

- 机器学习(Machine Learning, ML)是一门多领域交叉学科，涉及概率论、统计学、算法复杂度理论等多门学科。
- 研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能

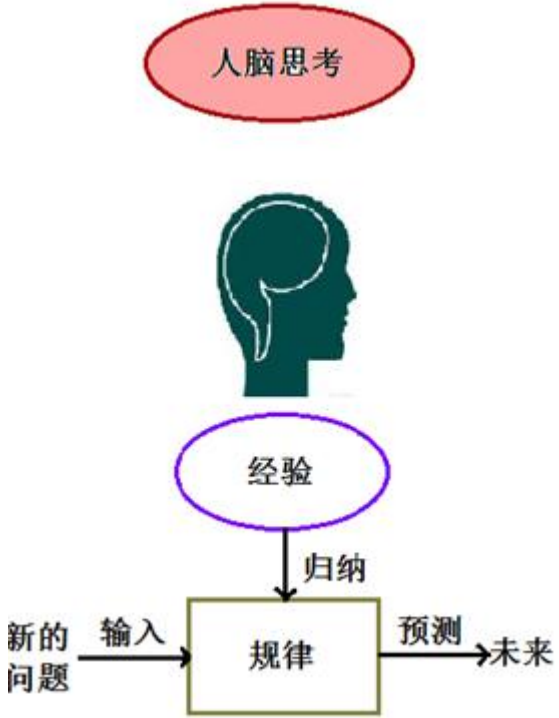
Machine learning is a branch of artificial intelligence,
concerns the construction and study of systems
that can learn from data.
机器学习是人工智能的一个分支，从事构建和研究可以从数据中学习的系统。

3.1 什么是机器学习


Data
More, More, More!!
Learning Machine
Predictions
Hypothesis

人类学习与机器学习

- 从观察中积累经验来获取技能



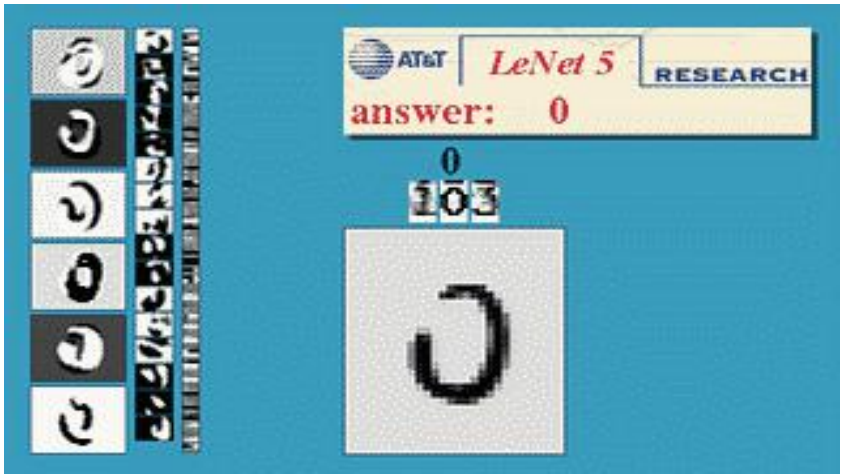
- 从数据中积累或者计算来获取技能

3.2 机器学习的典型任务

- 分类　　Classification
- 回归　　Regression
- 聚类　　Clustering
- 关联　　Association

3.2 机器学习的典型任务——分类

- 定义：按照某种指定的属性特征将数据归类。需要确定类别的概念描述，并找出类判别准则。分类的目的是获得一个分类函数或分类模型（也常常称作分类器），该模型能把数据集合中的数据项映射到某一个给定类别。
  - 例如：
    - 男女性别、
    - 疾病和健康、
    - 垃圾邮件的处理
    - 手写数字的识别

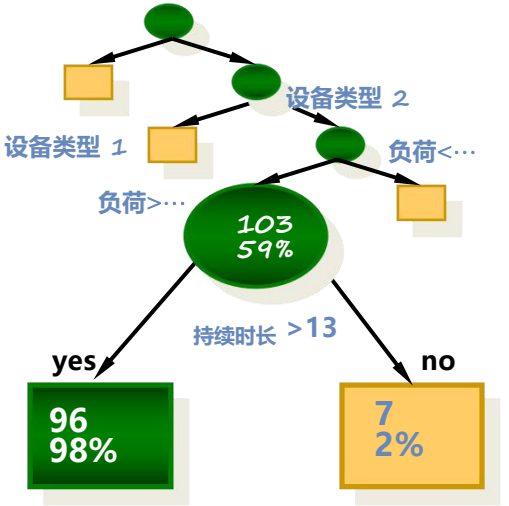


3.2 机器学习的典型任务——分类

- 分类是利用训练数据集通过一定的算法而求得分类规则的。
- 分类可用于提取描述重要数据类的模型或预测未来的趋势

银行根据客户以往贷款记录情况，将客户分为低风险客户和高风险客户，学习得到分类器。对一个新来的申请者，根据分类器计算风险，决定接受或拒绝该申请

分析影响变压器正常运行的因素，预测变压器是否有故障，若有故障，故障为放电故障、过热故障、短路故障等的哪一种。

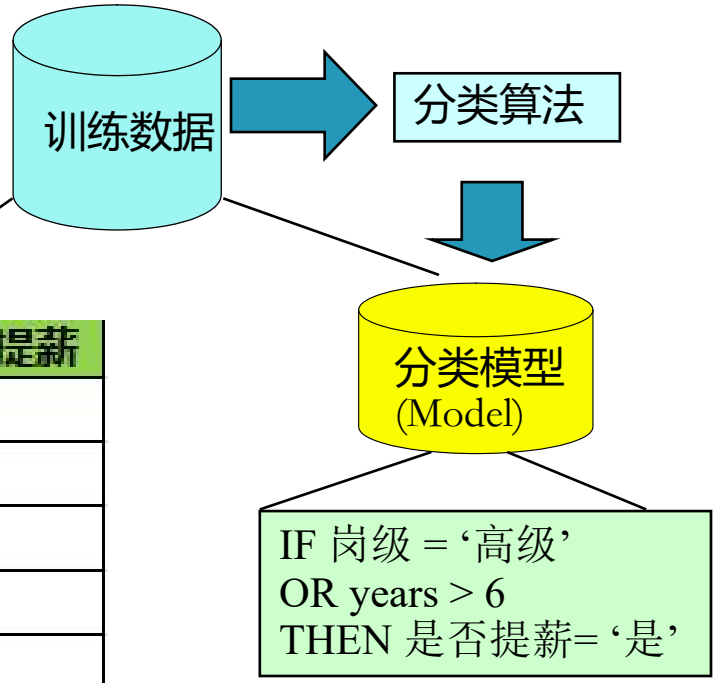


3.2 机器学习的典型任务——分类

- 分类的实现：模型的构建

  - 对每个样本进行类别标记
  - 训练集构成分类模型
  - 分类模型可表示为：分类规则、决策树或数学公式

姓名	岗级	司龄	是否提薪
张伞	中级	3	否
李斯	中级	7	是
王武	高级	2	是
陈柳	中级	7	是
王露	中级	6	否
田超	中级	3	否



3.2 机器学习的典型任务——分类

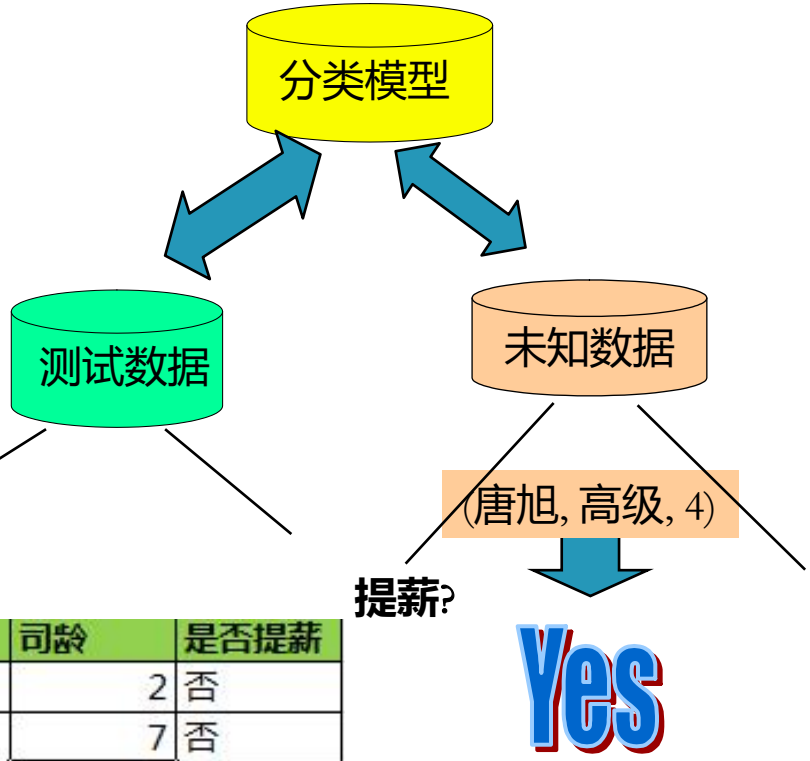
- 分类的实现：

模型的使用

- 识别未知对象的所属类别
- 模型正确性的评价
  - ✓已标记分类的测试样本与模型的实际分类结果进行比较

模型的正确率是指测试集中被正确分类的样本数与样本总数的百分比。

测试集与训练集相分离，否则将出现过拟合（over-fitting）现象



姓名	岗级	司龄	是否提薪
高维	中级	2	否
李文	中级	7	否
王敏	高级	5	是
刘佳	中级	7	是