日期: /

导论 Introduction

一. 机器学习

1. 从经验中得出结论.

2. 本质:

① 数据中存在某种模式
② 无法在数学上精确表示它
③ 有关于这种模式的数据

3. 通过观察数据得到一种理解

4. 为什么研究机器学习

① 设计更好的系统:
② 认识科学: 帮助理解人类学习
③ 时机成熟: 算法, 数据, 计算资源

5 分类:

① 监督学习 (Supervised learning)
- 有标签数据.
- 有直接反馈
- 用于预测结果.

② 无监督学习 (Unsupervised learning)

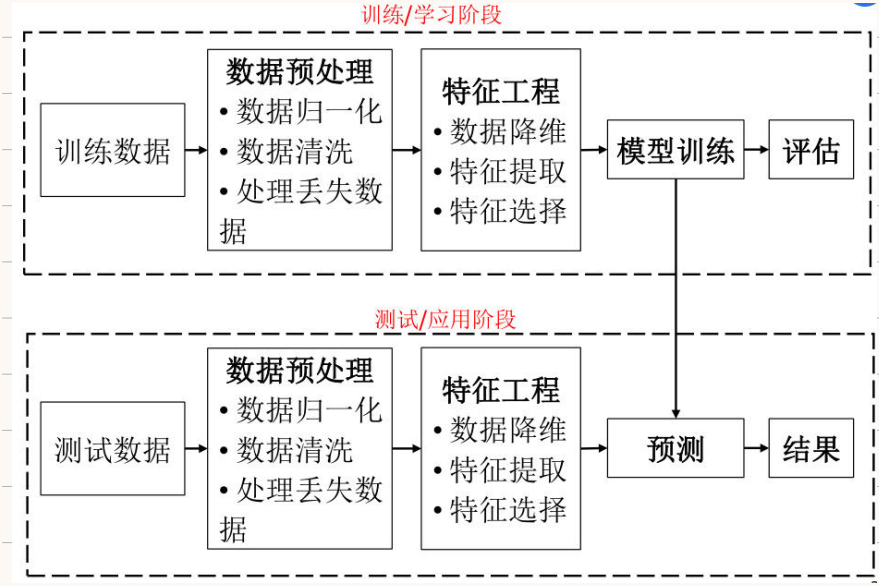
日期: /

- 无标签数据
- 无反馈
- 寻找数据间隐藏的模式

③强化学习 (Reinforcement learning).

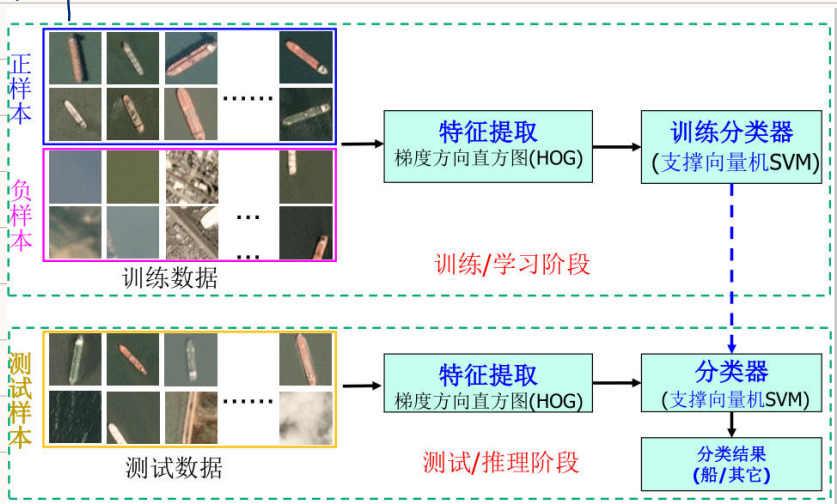
- 决策过程
- 奖励机制
- 从一系列动作中学习

b. 机器学习基本流程



日期: /

举例



二. 机器学习具体流程

1. 训练和测试集划分

① 收集数据集，标记数据

② 拆分数据集

- 训练集：用于训练模型，确定网络参数
- 验证集：调超参数，选择特征，调整学习算法
- 测试集：评估算法的性能

③ 数据集比例划分问题

没有一般标准，确保训练集，验证集，测试集比例各占多少是合适的

[依据于训练样本容量和数据信噪比]

日期: /

- 当样本数量不多（小于1万）的时候，通常将训练集/验证集/测试集的比例设为60%:20%:20%
- 在没有验证集的情况下，训练集/测试集的比例设为70%:30%
- 当样本数量很大（百万级别）的时候，通常将相应的训练集/验证集/测试集比例设为98%:1%:1%或者99%:1%(训练集/测试集)

指明多少数据是足够的
它依赖于基础函数信噪比和拟合数据的模型复杂度

- 验证集的规模应该尽可能大，至少要能够区分出你所尝试的不同算法之间的性能差异。通常来说，验证集的规模应该在1000 到 10000 个样本数据之间

2. 训练和测试：
① 训练模型
② 评价

三. 数据挖掘概述
1. 什么是数据挖掘？

- 在大型数据存储中，自动地发现有用信息的过程
  - 探查大型数据集，发现先前**未知**的**有用**信息
  - 或是预测未来观测结果

日期: /

- 更严谨的表述
  - **数据挖掘**就是**从大量的、不完全的、有噪声的、模糊的、随机的**实际应用数据中，提取**隐含在其中的、人们事先不知道的**、但又是**潜在有用**的信息和知识的过程。

2. 举例

- 非数据挖掘
  - 从电话簿查找电话号码
  - 从Web中查找信息“数据 挖掘”
  - 获得职工的平均资薪
- 数据挖掘
  - 某插班生应该读几年级?
  - 买哪只股票更可能挣钱?
  - 怎么才能多卖化妆品?
  - 海量文档该如何归类?
  - 行驶车辆如何预警?
  - 广告如何派送更好?

3. 数据挖掘的核心任务是知识发现

数据: 原始的, 未解释的信号或者符号, 如: 1

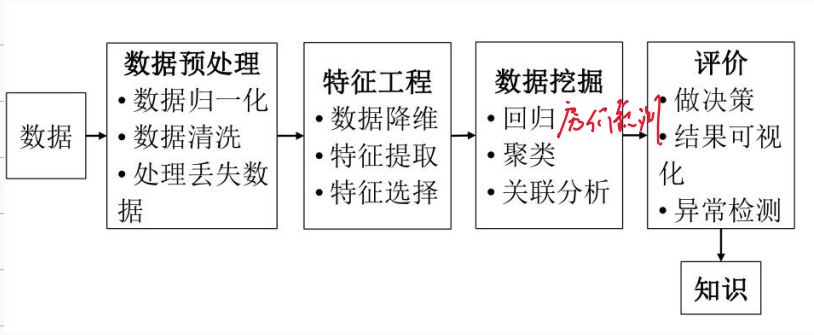
信息:有一定解释或意义的数据, 如: S.O.S

知识: 综合信息形成的观点和普适性的理论

智慧: 能够综合知识和经验用以生存计划的人类思维的结晶

日期: /

4. 数据挖掘基本流程



5. 机器学习与数据挖掘区别

- **机器学习**是人工智能的一个分支，旨在使系统从提供的数据中自动学习，并随着时间的推移改进它们的学习，而无需明确编程。它被用作一种数据挖掘技术。
- **数据挖掘（Data Mining or Data Science）侧重于分析数据并**从中提取知识和/或未知的有趣模式。 目标是了解数据中的模式以解释某些现象，而不是开发一个可以预测未知/新数据结果的复杂模型。
  提取的知识可以进一步用于商业应用，例如，可以对现有数据使用数据挖掘来了解公司的销售趋势，然后构建机器学习模型以从该数据中学习，找到相关性并适应新数据。
- **两者的相似性**
  - **机器学习**通常被视为更接近人工智能。
  - **数据挖掘**通常被视为更接近软件工程。
- **深度学习**是机器学习的一个子领域，其中模型是神经网络

6. 数据挖掘任务类型

日期: /

- 预测问题：预测对象的未知特性 → 预测任务（分类和回归）
- 聚类问题：获取数据中未知模式
- 关联分析：获取未知的关联关系
- 异常检测：获取未知的数据异常

（聚类问题、关联分析、异常检测）→ 描述任务

四. 数据挖掘任务（4个）

1. 分类问题

① 问题定义：

(1) 训练集：给定一组对象，对象用一组属性(attr)和一个类别(class)进行标记进行描述

(2) 分类模型：寻找一个能够描述 attr 和 class 关系的函数 [f(attr) = class]

(3) 目标：对于已知 attr，未知 class 的对象尽可能准确的 class 估计

- 引入测试集对模型性能进行评估
- 分类问题 → 划分 trainset / testset

② 核心思想

日期: /

寻找一个模型使其对特征与类别之间的关系进行描述

③ 应用:

1. 垃圾邮件过滤

- 对象: 邮件
- 特征: 邮件单词 (文本向量化 $\{num(word_i)\}$)
- 训练样本: 是/否为垃圾邮件
- 输出: 一封邮件为垃圾邮件概念

2. 图像识别

④ 分类技术:

- 基本分类模型
  - 决策树 Decision Tree based Methods
  - 规则学习 Rule-based Methods
  - 最近邻 Nearest-neighbor
  - 神经网络 Neural Networks
  - 贝叶斯方法 Naïve Bayes and Bayesian Belief Networks
  - 支持向量机 Support Vector Machines
- 集成方法
  - 提升方法Boosting, Bagging, 随机森林Random Forests

2. 回归问题

问题定义:

① 训练集: 给定一组对象, 这些对象可用一组特征属

日期: /

性和一个被预测属性（连续变量）进行描述。

② 回归模型：寻找一个能描述特征属性和被预测属性关系的函数，f（特征属性）=被预测~

③ 目标：最小化 模型预测值与真实预测属性之差[均方误差]

☆ 与分类问题区别在于 被预测属性是否连续

3. 聚类问题

① 问题定义

(1) 给定一组数据（数据间可通过一种距离度量）

(2) 目标：寻找一组 数据点使得

- 同一 cluster 内部之间距离尽可能小
- 不同 cluster 之间的点的距离尽可能大

(3) 度量距离的设计

- 每个点：一组属性描述。
- 数据属性为连续值 → 可用欧式距离
- 不同物品 → 不同距离度量方法

② 核心思想：

寻找分组方式，使得
- 组内 相似
- 组间 差异大

③ 功能：

日期: /

{ 凸显数据特征.
降低数据分析难度

④ 不使聚类的情况.

- 依据简单规则的数据对象划分
  - 例如入大学之后对给位同学进行分班，在世界杯根据抽签结果对参赛队伍进行分组等。
  - 没有考虑到个体之间距离的因素。
- 根据外部属性进行简单划分
  - 例如根据籍贯、民族对人口进行族群划分，根据年龄将人分为少年、中年、老年等。
  - 缺少必要的数据聚类建模过程。
- 从外部标签学习获得的分类模型(classification)
  - 数据集本身具有明确的类型标签，根据标签训练模型对数据进行划分。
  - 数据类型划分标准是通过外部信息获得的，而非数据集本身。

⑧ 聚类特点.

{ 根据聚类对象在特征空间的距离
对数据进行聚类建模.
将数据无监督的划分为若干组或簇(cluster)

4. 关联规则.

① 问题定义:

日期: /

(item)

① 给定一个记录的集合，每一条记录包括若干项
② 从集合中找出由一个 item / item set 预测另一个 item / item set 同时出现的规则.

④ 表述形式 X => Y.
① 满足X中条件的数据库元素，在一定程度上也满足Y中的条件.
② X为前项，Y被称为后项

5. 偏离/异常检测
① 从正常的行为中检测重要的偏离
② 应用：电信欺诈检测
网络入侵.

③ 挑战：

- 离群点的数量是未知的
- 分析过程可能是**无监督**的
  - 无监督模型的一个难点是对于分析结果非常难以验证（这一 点和聚类问题相同）
- 分类过程如果是**有监督**的
  - “正常”样本的数量是远远多于“异常”样本
  - 异常检测往往可以等价于一种**非对称**的分类问题
- 数据不平衡问题(海底捞针问题)：机器学习分类器从大量负类（不感兴趣的）中找到少数正类（感兴趣，或故障）
  - 1. 每年大约有2%的信用卡账户被欺骗。（大多数欺诈检测领域严重不平衡）
  - 2. 工厂生产故障率通常约0.1%。