# 1、机器学习概览

2019年1月5日　　20:37

# 第一章 机器学习概览

- 机器学习的种类
    - 监督式/无监督式学习

      根据训练期间接受的监督数量和监督类型，可以将机器学习系统分为以下四个类别：监督式学习、无监督式学习、半监督事学习和强化学习
        - 监督式学习
        - 无监督式学习
        - 半监督式学习

          有些算法可以处理部分标记的训练数据——通常是大量未标记数据和少量的标记数据，此即半监督式学习。有些照片托管服务就是半监督式学习。一旦你将所有的家庭照片上传到服务器以后，它会自动识别出任务A出现在照片1、5和11中，人物B出现在照片2、5和7中。这是算法的无监督部分（聚类）。现在系统需要你做的知识，告诉它这些人都是谁。给每个人一个标签以后，它就可以对每张照片中的每个人命名。大多数半监督式学习算法式监督学习和无监督学习的结合
            - 深度信念网络
        - 强化学习
    - 批量学习和在线学习

      根据系统是否可以从传入的数据流中进行增量学习进行分类。
        - 批量学习

          系统无法进行增量学习——即必须使用所有可用数据进行训练。先训练系统，然后将其投入生产环境，这时学习过程停止，它只是将学到的应用出来。
        - 在线学习

          在在线学习中，可以循序渐进地给系统提供训练数据，逐步积累学习成果。这种提供数据的方式可以是单独的（就是一次给一个example），也可以采用小批量的小组数据来进行训练。系统可以根据飞速写入的最新数据进行学习
    - 基于实例和基于模型的学习
        - 基于实例的学习

          系统先完全记住学习实例，然后通过某种相似度度量方式将其泛化到新的实例
        - 基于模型的学习

          首先根据实例构建模型，然后使用该模型进行预测。
- 机器学习的主要挑战
    - 来自于数据的挑战
        - 训练数据量不足
        - 训练数据不具有代表性
        - 质量差的数据（比如有很多缺失值）
        - 数据中有很多无关特征
    - 来自算法的挑战
        - 过拟合
            - 解决方法

        - 简化模型（正则化）
        - 收集更多训练数据
        - 减少数据中的噪声
    - 欠拟合
        - 解决方法
            - 选择更复杂的模型
            - 给算法提供更好的特征集
            - 减少模型的约束
- 测试与验证