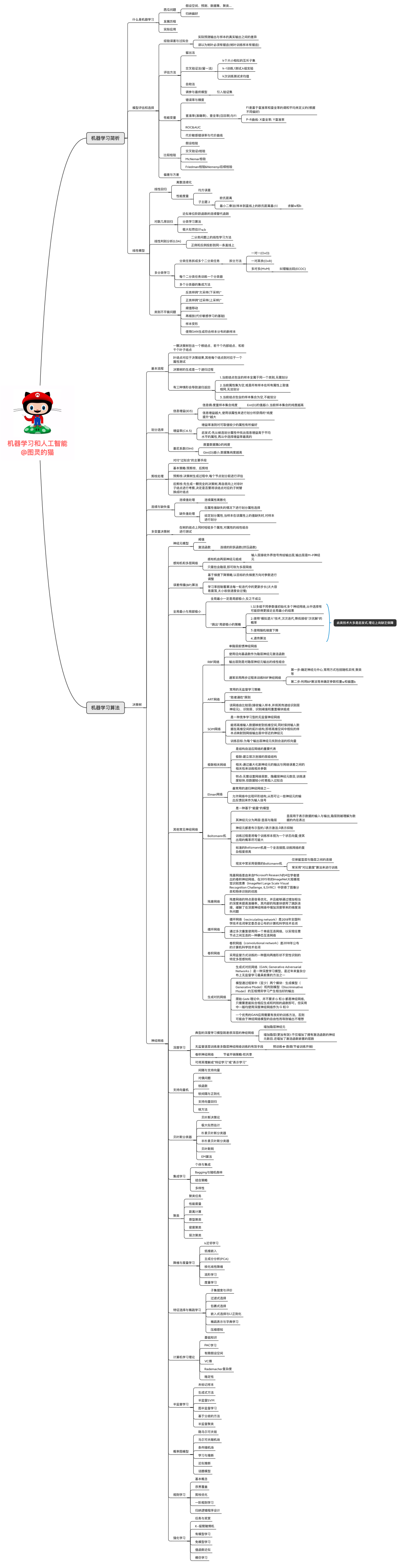

机器学习和人工智能
@图灵的猫

- 机器学习简析
  - 什么是机器学习
    - 西瓜问题
      - 假设空间、预测、数据集、聚类...
      - 归纳偏好
    - 发展历程
    - 实际应用
  - 模型评估和选择
    - 经验误差与过拟合
      - 实际预测输出与样本的真实输出之间的差异
      - 误以为树叶必须有锯齿(树叶训练样本有锯齿)
    - 评估方法
      - 留出法
      - 交叉验证法(留一法)
        - k个大小相似的互斥子集
        - k-1训练,1测试,k组实验
        - k次训练测试求均值
      - 自助法
      - 调参与最终模型
        - 引入验证集
    - 性能变量
      - 错误率与精度
      - 查准率(准确率)、查全率(召回率)与F1
        - F1是基于查准率和查全率的调和平均来定义的(根据不同偏好)
        - P-R曲线: X查全率; Y查准率
      - ROC&AUC
      - 代价敏感错误率与代价曲线
    - 比较检验
      - 假设检验
      - 交叉验证t检验
      - McNemar检验
      - Friedman检验&Nemenyi后续检验
    - 偏差与方差
  - 线性模型
    - 线性回归
      - 离散连续化
      - 性能度量
        - 均方误差
        - 子主题 2
          - 欧氏距离
          - 最小二乘法(样本到直线上的欧氏距离最小)
            - 求解w和b
    - 对数几率回归
      - 近似单位阶跃函数的连续替代函数
      - 分类学习算法
      - 极大似然估计w,b
    - 线性判别分析(LDA)
      - 二分类问题上的线性学习方法
      - 正例和反例投影到同一条直线上
    - 多分类学习
      - 分类任务拆成多个二分类任务
        - 拆分方法
          - 一对一(OvO)
          - 一对其余(OvR)
          - 多对多(MvM)
            - 纠错输出码(ECOC)
      - 每个二分类任务训练一个分类器
      - 多个分类器的集成方法
    - 类别不平衡问题
      - 反类样例"欠采样(下采样)"
      - 正类样例"过采样(上采样)"
      - 阈值移动
      - 再缩放(代价敏感学习的基础)
      - 样本变形
      - 使用GAN生成符合样本分布的新样本
- 机器学习算法
  - 决策树
    - 基本流程
      - 一颗决策树包含一个根结点、若干个内部结点、和若干个叶子结点
      - 叶结点对应于决策结果,其他每个结点则对应于一个属性测试
      - 决策树的生成是一个递归过程
      - 有三种情形会导致递归返回
        - 1.当前结点包含的样本全属于同一个类别,无需划分
        - 2.当前属性集为空,或是所有样本在所有属性上取值相同,无法划分
        - 3.当前结点包含的样本集合为空,不能划分
    - 划分选择
      - 信息增益(ID3)
        - 信息熵:度量样本集合纯度
          - Ent(D)的值越小,当前样本集合的纯度越高
        - 信息增益越大,使用该属性来进行划分所获得的"纯度提升"越大
      - 增益率(C4.5)
        - 增益率准则对可取值较少的属性有所偏好
        - 启发式:先从候选划分属性中找出信息增益高于平均水平的属性,再从中选择增益率最高的
      - 基尼系数(Gini)
        - 度量数据集D的纯度
        - Gini(D)越小,数据集纯度越高
    - 剪枝处理
      - 对付"过拟合"的主要手段
      - 基本策略:预剪枝、后剪枝
      - 预剪枝:决策树生成过程中,每个节点划分前进行评估
      - 后剪枝:先生成一颗完全的决策树,再自底向上对非叶子结点进行考察,决定是否要将该结点对应的子树替换成叶结点
    - 连续与缺失值
      - 连续值处理
        - 连续属性离散化
      - 缺失值处理
        - 在属性值缺失的情况下进行划分属性选择
        - 给定划分属性,当样本在该属性上的值缺失时,对样本进行划分
    - 多变量决策树
      - 在树的结点上同时检验多个属性,对属性的线性组合进行测试
    - 神经网络
      - 神经元模型
        - 阈值
        - 激活函数
          - 连续的阶跃函数(挤压函数)
      - 感知机和多层网络
        - 感知机由两层神经元组成
          - 输入层接收外界信号传给输出层,输出层是M-P神经元
        - 只需包含隐层,即可称为多层网络
      - 误差传播(BP)算法
        - 基于梯度下降策略,以目标的负梯度方向对参数进行调整
        - 学习率控制着算法每一轮迭代中的更新步长(太大容易震荡,太小收敛速度会过慢)
      - 全局最小与局部极小
        - 全局最小一定是局部极小,反之不成立
        - "跳出"局部极小的策略
          - 1.以多组不同参数值初始化多个神经网络,从中选择有可能获得更接近全局最小的结果
          - 2.使用"模拟退火"技术,次次迭代,降低接收"次优解"的概率
          - 3.使用随机梯度下降
          - 4.遗传算法
          - 此类技术大多是启发式,理论上尚缺乏保障
      - 其他常见神经网络
        - RBF网络
          - 单隐层前馈神经网络
          - 使用径向基函数作为隐层神经元激活函数
          - 输出层则是对隐层神经元输出的线性组合
          - 通常采用两步过程来训练RBF神经网络
            - 第一步:确定神经元中心,常用方式包括随机采样,聚类等
            - 第二步:利用BP算法等来确定参数权重w和偏置b
        - ART网络
          - 常用的无监督学习策略
          - "胜者通吃"原则
          - 该网络由比较层(接收输入样本,并将其传递给识别层神经元)、识别层、识别阈值和重置模块组成
        - SOM网络
          - 是一种竞争学习型的无监督神经网络
          - 能将高维输入数据映射到低维空间,同时保持输入数据在高维空间的拓扑结构,即将高维空间中相似的样本点映射到网络输出层中邻近的神经元
          - 训练目标:为每个输出层神经元找到合适的权向量
        - 极联相关网络
          - 是结构自适应网络的重要代表
          - 极联:建立层次连接的层级结构
          - 相关:通过最大化新神经元的输出与网络误差之间的相关性来训练相关参数
          - 特点:无需设置网络层数、隐藏层神经元数目,训练速度较快,但数据较小时易陷入过拟合
        - Elman网络
          - 最常用的递归神经网络之一
          - 允许网络中出现环形结构,从而可让一些神经元的输出反馈回来作为输入信号
        - Boltzmann机
          - 是一种基于"能量"的模型
          - 其神经元分为两层:显层与隐层
            - 显层用于表示数据的输入与输出,隐层则被理解为数据的内在表达
          - 神经元都是布尔型的,1表示激活,0表示抑制
          - 训练过程是将每个训练样本视为一个状态向量,使其出现的概率尽可能大
          - 标准的Boltzmann机是一个全连接图,训练网络的复杂程度很高
          - 现实中常采用受限的Boltzmann机
            - 仅保留显层与隐层之间的连接
            - 常采用"对比散度"算法来进行训练
        - 残差网络
          - 残差网络是由来自Microsoft Research的4位学者提出的卷积神经网络，在2015年的ImageNet大规模视觉识别竞赛（ImageNet Large Scale Visual Recognition Challenge, ILSVRC）中获得了图像分类和物体识别的优胜
          - 残差网络的特点是容易优化，并且能够通过增加相当的深度来提高准确率。其内部的残差块使用了跳跃连接，缓解了在深度神经网络中增加深度带来的梯度消失问题
        - 循环网络
          - 循环网络（recirculating network）是2018年全国科学技术名词审定委员会公布的计算机科学技术名词
          - 通过多次重复使用同一个单级互连网络，以实现任意节点之间互连的一种静态互连网络
        - 卷积网络
          - 卷积网络（convolutional network）是2018年公布的计算机科学技术名词
          - 采用监督方式训练的一种面向两维形状不变性识别的特定多层感知机
        - 生成对抗网络
          - 生成式对抗网络（GAN, Generative Adversarial Networks ）是一种深度学习模型，是近年来复杂分布上无监督学习最具前景的方法之一
          - 模型通过框架中（至少）两个模块：生成模型（Generative Model）和判别模型（Discriminative Model）的互相博弈学习产生相当好的输出
          - 原始 GAN 理论中，并不要求 G 和 D 都是神经网络，只需要是能拟合相应生成和判别的函数即可。但实用中一般均使用深度神经网络作为 G 和 D
          - 一个优秀的GAN应用需要有良好的训练方法，否则可能由于神经网络模型的自由性而导致输出不理想
      - 深度学习
        - 典型的深度学习模型就是很深层的神经网络
          - 增加隐层神经元
          - 增加隐层(更加有效):不仅增加了拥有激活函数的神经元数目,还增加了激活函数嵌套的层数
        - 无监督逐层训练是多隐层神经网络训练的有效手段
          - 预训练+微调(节省训练开销)
        - 卷积神经网络
          - 节省开销策略:权共享
        - 可将其理解成"特征学习"或"表示学习"
      - 支持向量机
        - 间隔与支持向量
        - 对偶问题
        - 核函数
        - 软间隔与正则化
        - 支持向量回归
        - 核方法
      - 贝叶斯分类器
        - 贝叶斯决策论
        - 极大似然估计
        - 朴素贝叶斯分类器
        - 半朴素贝叶斯分类器
        - 贝叶斯网
        - EM算法
      - 集成学习
        - 个体与集成
        - Bagging与随机森林
        - 结合策略
        - 多样性
      - 聚类
        - 聚类任务
        - 性能度量
        - 距离计算
        - 原型聚类
        - 密度聚类
        - 层次聚类
      - 降维与度量学习
        - k近邻学习
        - 低维嵌入
        - 主成分分析(PCA)
        - 核化线性降维
        - 流形学习
        - 度量学习
      - 特征选择与稀疏学习
        - 子集搜索与评价
        - 过滤式选择
        - 包裹式选择
        - 嵌入式选择与L1正则化
        - 稀疏表示与字典学习
        - 压缩感知
      - 计算机学习理论
        - 基础知识
        - PAC学习
        - 有限假设空间
        - VC维
        - Rademacher复杂度
        - 稳定性
      - 半监督学习
        - 未标记样本
        - 生成式方法
        - 半监督SVM
        - 图半监督学习
        - 基于分歧的方法
        - 半监督聚类
      - 概率图模型
        - 隐马尔可夫链
        - 马尔可夫随机场
        - 条件随机场
        - 学习与推断
        - 近似推断
        - 话题模型
      - 规则学习
        - 基本概念
        - 序贯覆盖
        - 剪枝优化
        - 一阶规则学习
        - 归纳逻辑程序设计
      - 强化学习
        - 任务与奖赏
        - K-摇臂赌博机
        - 有模型学习
        - 免模型学习
        - 值函数近似
        - 模仿学习