机器学习（2025 秋季学期）

课程总结与复习

成绩比例（调整）

- 作业　　(35%)
- 实践　　(15%)
- 期末考试　(50%)

考试题型

- 选择题（约30分）：3分/题×10题
- 判断题（约10分）：2分/题×5题
- 简答题（约10分）：5分/题×2题
- 理论推导题（约15分）：15分/题×1题
- 计算题（约35分）：10~15分/题×3题

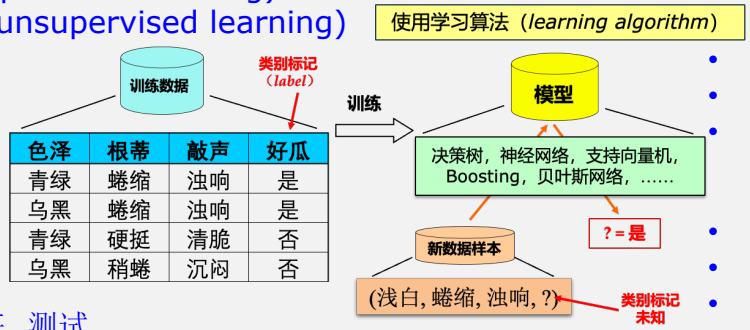
考试内容范围

- 考试范围：第1~11、13、16章
- 不考的章节：第12、14、15章

第1章 绪论

- 基本术语
- 假设空间
- 归纳偏好

- 监督学习(supervised learning)
- 无监督学习(unsupervised learning)



- 假设(hypothesis)
- 真相(ground-truth)
- 学习器(learner)

- 分类，回归
- 二分类，多分类
- 正类，反类

- 数据集；训练，测试
- 示例(instance)，样例(example)
- 样本(sample)
- 属性(attribute)，特征(feature)；属性值
- 属性空间，样本空间，输入空间
- 特征向量(feature vector)
- 标记空间，输出空间

- 未见样本(unseen instance)
- 未知“分布”
- 独立同分布(i.i.d.)
- 泛化**(generalization)**

第2章 模型评估与选择

- 经验误差与过拟合
- 评估方法
- 性能度量
- 比较检验
- 偏差与方差

三个关键问题：

- 如何获得测试结果？ ⇨ 评估方法
- 如何评估性能优劣？ ⇨ 性能度量
- 如何判断实质差别？ ⇨ 比较检验

第3章 线性模型

- 线性回归
  - 最小二乘法（最小化均方误差）
- 二分类任务
  - 对数几率回归
    - 单位阶跃函数
    - 对数几率函数
    - 极大似然法
  - 线性判别分析
    - 最大化广义瑞利商
- 多分类任务
  - 一对一
  - 一对其余
  - 多对多
    - 最大化广义瑞利商
- 类别不平衡问题
  - 基本策略：再缩放

第4章 决策树

- 基本流程
- 划分选择
- 剪枝处理（预剪枝，后剪枝）
- 连续与缺失值
- ~~多变量决策树~~

$$\mathrm{Ent}(D) = -\sum_{k=1}^{|\mathcal{Y}|} p_k \log_2 p_k$$

$$\mathrm{Gain}(D,a) = \mathrm{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|}\mathrm{Ent}(D^v)$$

$$\mathrm{Gain\_ratio}(D,a) = \frac{\mathrm{Gain}(D,a)}{\mathrm{IV}(a)}$$

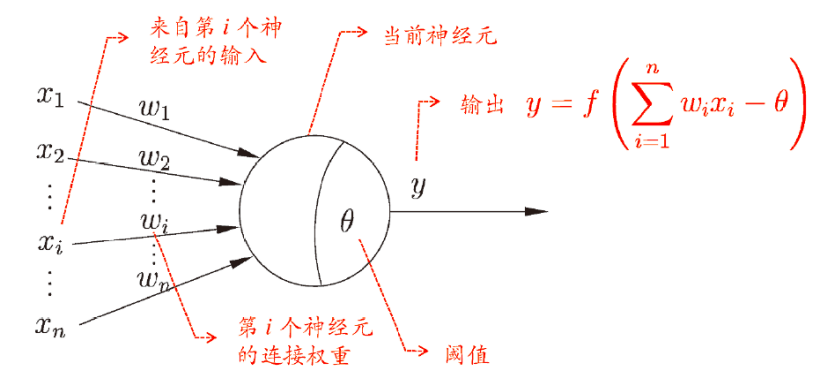
$$\mathrm{IV}(a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

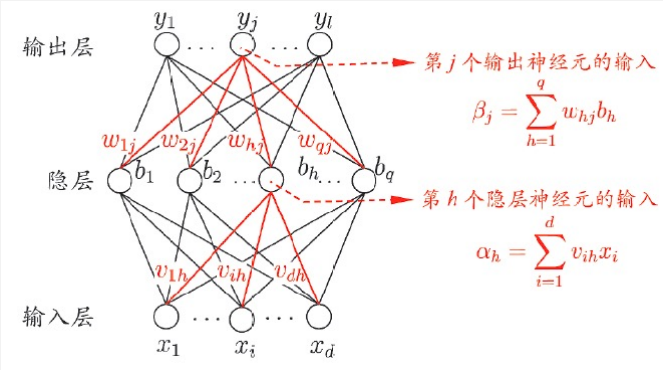
$$\mathrm{Gini}(D) = \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} = 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2 .$$

$$\mathrm{Gini\_index}(D,a) = \sum_{v=1}^{V} \frac{|D^v|}{|D|}\mathrm{Gini}(D^v)$$

第5章 神经网络

- 神经元模型
- 感知机与多层网络
- 误差逆传播算法
- 全局最小与局部最小
- ~~其他常见神经网络~~
- 深度学习~~(补充内容不考)~~





第6章 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

- 支持向量机的“最大间隔”思想
- 对偶问题及其解的稀疏性
- 通过向高维空间映射解决线性不可分的问题
- 引入“软间隔”缓解特征空间中线性不可分的问题
- 将支持向量的思想应用到回归问题上得到支持向量回归
- 将核方法推广到其他学习模型

第7章 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

样本相对于类标记的类条件概率(class-conditional probability),亦称 似然(likelihood)

$$P(c \mid \boldsymbol{x}) = \frac{P(c)\,P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})}$$

先验概率（prior）
样本空间中各类样本所占的比例，可通过各类样本出现的频率估计（大数定律）

证据 (evidence)
因子，与类别无关

第8章 集成学习

- 个体与集成
- Boosting
- Bagging与随机森林
- 结合策略
- ~~多样性~~

- 序列化方法
  - **AdaBoost** [Freund & Schapire, JCSS97]
  - GradientBoost [Friedman, AnnStat01]
  - LPBoost [Demiriz, Bennett, Shawe-Taylor, MLJ06]
  - … …
- 并行化方法
  - **Bagging** [Breiman, MLJ96]
  - Random Forest [Breiman, MLJ01]
  - Random Subspace [Ho, TPAMI98]
  - … …

常用结合方法：

- 投票法
  - 绝对多数投票法
  - 相对多数投票法
  - 加权投票法
- 平均法
  - 简单平均法
  - 加权平均法
- 学习法

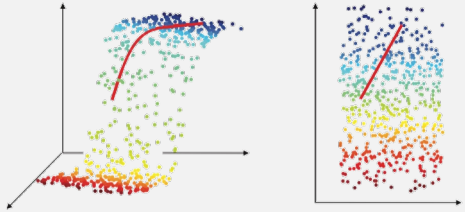
第9章 聚类

- 聚类任务
- 性能度量
- 距离计算
- 原型聚类
- 密度聚类
- 层次聚类

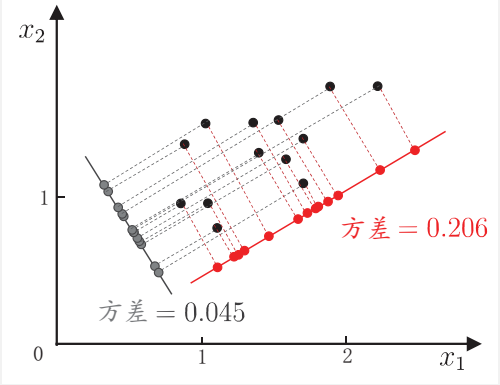
- 原型聚类，亦称“基于原型的聚类”(prototype-based clustering)
  - 假设：聚类结构能通过一组原型刻画
  - 过程：先对原型初始化，然后对原型进行迭代更新求解
  - 代表：k均值聚类，学习向量量化(LVQ)，高斯混合聚类
- 密度聚类，亦称“基于密度的聚类”(density-based clustering)
  - 假设：聚类结构能通过样本分布的紧密程度确定
  - 过程：从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇
  - 代表：DBSCAN, OPTICS, DENCLUE
- 层次聚类(hierarchical clustering)
  - 假设：能够产生不同粒度的聚类结果
  - 过程：在不同层次对数据集进行划分，从而形成树形的聚类结构
  - 代表：AGNES(自底向上)，DIANA (自顶向下)

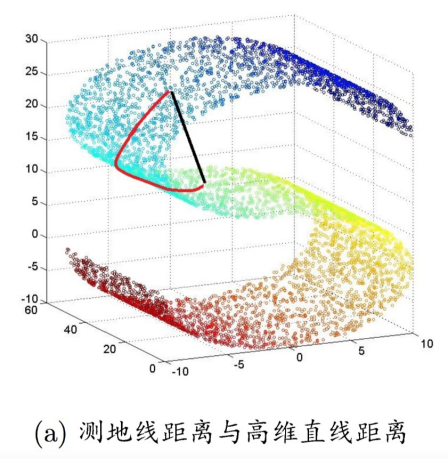
第10章 降维与度量学习

- k近邻学习
- 低维嵌入
- 主成分分析
- 流形学习
- ~~度量学习~~



(a) 三维空间中观察到的样本点　(b) 二维空间中的曲面

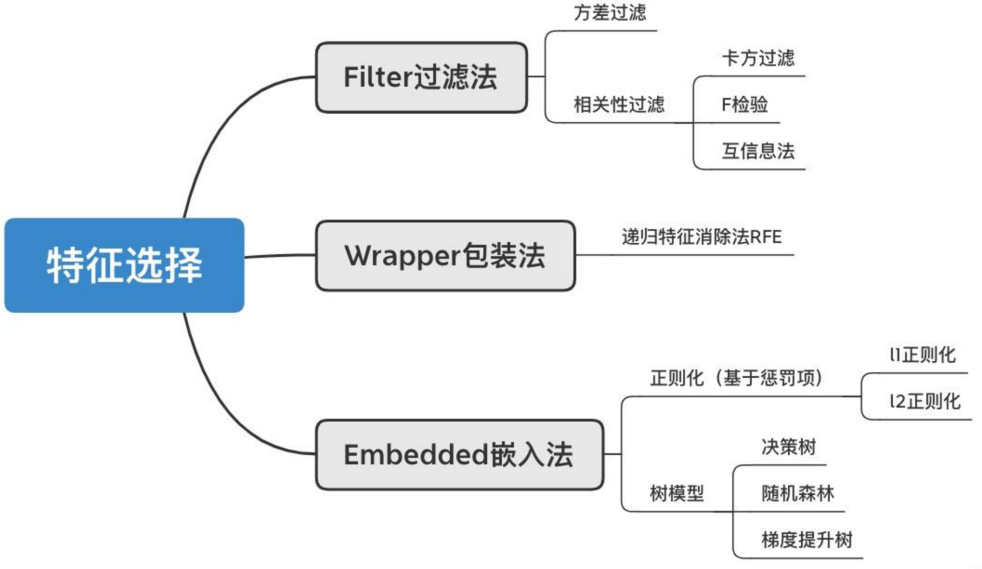




(a) 测地线距离与高维直线距离

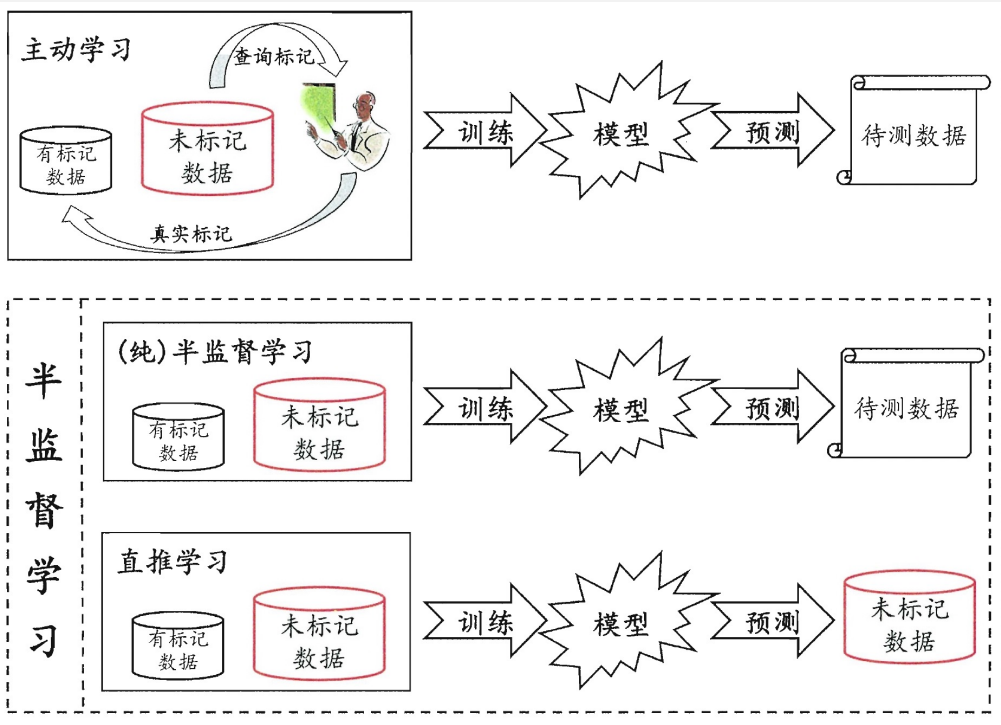
第11章 特征选择与稀疏学习

- 子集搜索与评价
- 过滤式选择
- 包裹式选择
- 嵌入式选择
- 稀疏表示与字典学习
- ~~压缩感知~~



第13章 半监督学习

- 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



第16章 强化学习

- 什么是强化学习?
- K-摇臂赌博机
- 有模型学习
- 免模型学习
- 值函数近似
- 一些方向

- 强化学习：多步决策过程
- 有模型学习
  - 基于动态规划的寻优
- 如何处理环境中的未知因素
  - 蒙特卡罗强化学习
  - 时序差分学习
- 如何处理连续状态空间
  - 值函数近似
- 如何提速强化学习过程
  - 直接模仿学习
  - 逆强化学习