对立概念辨析：名词解释或简答题

机器学习（绪论）

- 机器学习典型过程（方案题）
- 分类vs回归
  - 哪些模型、性能指标（如MAE, MSE, R^2可做回归指标）
- 数据标注（联系半监督：三个假设、自训练）
- 数据变换

  标准化（实属、图片、视频、文本等）
- 特征工程
  - 传统ML vs DL
  - 表格数据、文本特征……

模型评估

- 作用（找到泛化能力强的模型 适应未见过的数据）
- 经验误差vs泛化误差
- 过拟合vs欠拟合
- 模型评估/选择 三个关键问题（前两个重要）
  - **交叉验证**
- 性能度量
  - **混淆矩阵**算查准率、查全率、F1
  - ROC，AUC的区别
  - 比较检验

线性模型

- 可分类与回归
- 离散属性处理（有序/无序）
- 线性回归公式推导（简单、多元）
- 广义线性模型（联系函数）
- **对率回归**（推导）
- 线性判别分析（做分类任务）（推导）
- 多分类（要知道怎么算）
- 类别不平衡

决策树

（会出计算题，可参考课件例子）

- 要知道根结点用什么属性
- 不同决策树模型

- 剪枝
  - 预剪枝、后剪枝异同点（时间开销、过拟合/欠拟合）
- ~~缺失值~~

SVM

- 明白求解原理
- 核函数（思想、有哪些、性质）
- ~~软间隔~~
- 正则化（定义、常用正则化项）
- ~~支持向量回归~~

贝叶斯分类器

- 理论框架：贝叶斯决策论（没有训练过程）
- 判别式vs生成式
- 朴素（假定属性独立）、半朴素（对属性条件独立性假设放松）的贝叶斯分类器区分
  - 半朴素分类器常用2种模型
- 计算过程
- ~~拉普拉斯修正~~
- ~~贝叶斯网络、EM~~

集成学习（考的比较简单）

- ~~理解“多样性”~~
- 序列化方法vs并行化方法 区别（概念、计算过程）
- ~~bagging，多层堆叠~~

聚类

- 分类vs聚类（概念、区别）
- **距离计算**
  - 距离度量的基本性质
  - 常用距离形式（闵可夫斯基距离->欧氏/曼哈顿距离）
  - 无序属性（VDM）、混合属性（MinkovDM）
- 常见3种聚类方法（前两种重要）
- k-means计算步骤、适用条件
- DBSCAN适用条件
- 重要概念

几个概念

- **Eps-邻域** (Eps-neighborhood of a point)
  点p的Eps-邻域，记为$N_{Eps}(p)$，定义为$N_{Eps}(p) = \{q \in D | dist(p,q) \leq Eps\}$.
- **核心点** (Core points)
  如果给定**点p**的Eps-邻域内的样本点数大于等于MinPts，则称**点p**为核心点。
- **密度直达**(directly density-reachable)
  若：1）$p \in N_{Eps}(q)$ 2）$|N_{Eps}(q)| \geq MinPts$ 则称点**p**由核心点**q**密度直达。
- **密度可达**（density-reachable)
  如果存在样本序列$p_1, p_2, \ldots, p_n$；如果满足$p_1 = q$, $p_n = p$。若p_{i+1}是由p_i密度直达的，则称**p**是由**q**密度可达的。
- **密度相连**(density-connected)
  对于点**p**和点**q**，若点**p**点**q**都是从点**o**密度可达的，则称点**p**和点**q**密度相连。
- 簇(cluster)
  对于数据集D，若C是其中一个簇，C中的点需要满足以下两个条件：
  1）$\forall p, q$，如果$p \in C$且q是由p密度可达的，则$q \in C$。 2）$\forall p, q \in C$，p和q是密度相连的。
- **噪音**(noise)不属于任何簇的点为噪音数据。

- ~~聚类过程~~
- 优缺点

降维

- 概念
- KNN
- ~~MDS~~
- **PCA**
  - 最近重构性、最大可分性
  - ~~KPCA~~
- 流形学习（概念）
- 其他方法（了解）
  - 必连约束、勿连约束

特征选择

- 特征选择vs降维（异同）
- 过滤式选择（relief概念原理）
- ~~LVW（包裹式选择）~~
- **嵌入式选择与正则化**（正则化必考）
  - L1正则化：L1（起特征选择效果），L2范数
- ~~字典学习以后都不用看~~

神经网络

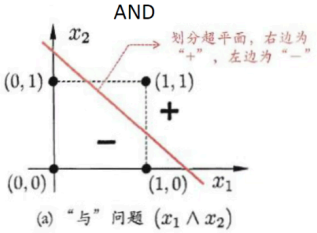
- 激活函数（定义、种类及其图形、性质）
- **感知机**（与或非、异或问题）可能出计算题

感知机与异或问题

感知机只有输出层神经元进行激活函数处理，即只有一层功能神经元，学习能力非常有限

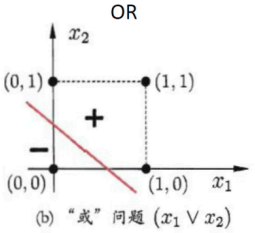
与或非问题都是线性可分的异或这样的非线性可分问题不能解决

AND

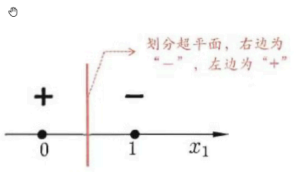


(a) “与”问题 $(x_1 \wedge x_2)$

OR

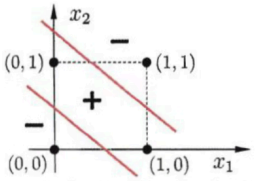


(b) “或”问题 $(x_1 \vee x_2)$



(c) “非”问题 $(\neg x_1)$

NOT



(d) “异或”问题 $(x_1 \oplus x_2)$

XOR

- 多层前馈网络（3个结构特点、前馈指拓扑结构不存在环或者回路）
- BP神经网络（梯度下降可能出计算题）
- 缓解过拟合策略（早停、验证集、正则化）
- 全局最小vs局部最小
- 随机梯度下降
  - 对比梯度下降
- 池化层（**作用**、平均池化与最大池化）
- ~~后面不看~~

半监督学习

- ~~概念~~
- 主动学习
- 数据分布假设
  - 使用无label数据需要对数据分布做假设
  - ✓连续性（continuity）假设：具有相似特征的样本很可能拥有相同的label
  - ✓聚类（cluster）假设：数据有内在的cluster结构，相同cluster内的样本具有相同的label
  - ✓流形（manifold）假设：数据在低维流形上分布，数据内在复杂性远比数据原始维度要低，可通过降维观察
- 自训练、active learning + self-training
- ~~后面都不用看~~