

INT104W1_机器学习概览

机器学习的定义

机器学习 = Prediction + Decision making (预测与决策)

你可以说AI是一种modern statistics (现代统计学)

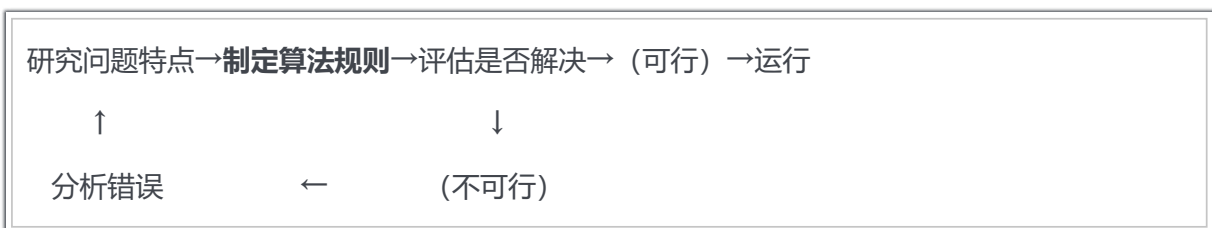
- Feature space特征空间
- label space标记空间

eg: 如果需要通过一个人的身高体重判断性别, 特征空间就是身高体重, 标记空间是性别

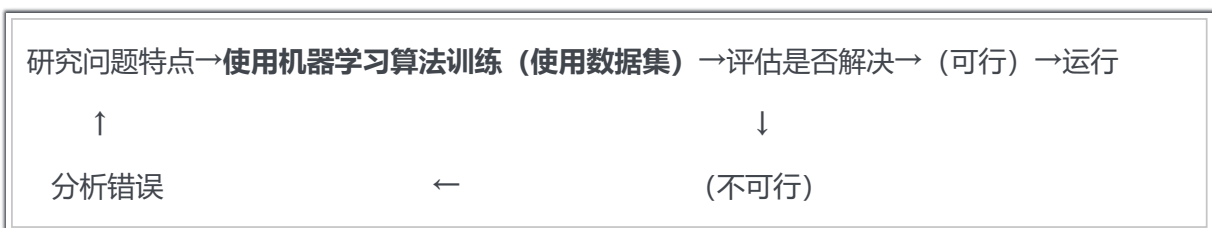
机器学习也可以说是从特征空间到标记空间的映射 (mapping)

对比传统方案

- 传统方案:



- 机器学习 (ML) 方案:



- 机器学习的学习过程 (适应过程) :

运行→更新数据 (使用数据集) →训练ML算法→检查得到的结果→(如需可循环运行)

(↑通过审查结果, 也能得到对问题的更深度理解)

ML可以完成的任务 (你将学习)

- Classification分类
- Regression回归

- Clustering聚类
- Anomaly detection异常检测
- Generation生成
- Modelling建模

■ 监督学习 Supervised Learning

- kNN (k最临近算法)
- 决策树 (Decision tree) 与随机森林 (Random forest)
- 支持向量机 (Support vector machine)

■ 无监督学习 Unsupervised Learning

- k-means (k均值聚类算法)
- DBSCAN(Density-Based Spatial Clustering of Applications with Noise)具有噪声的基于密度的聚类算法
- Hierarchical Cluster Analysis (HCA)层次聚类分析

■ 此外你还将学习

- 半监督学习 Semi-supervised Learning
- 示例学习 Instance Learning
- 强化学习 Reinforcement Learning

■ 模型选择 Model Selection

- 训练数据集 (Training Dataset)
- 测试数据集 (Testing Dataset)
- 验证数据集 (Validation Dataset)

注：训练集不应与验证集重叠（否则抄答案）

过拟合overfitting：模型在面对训练样本

■ ML目前挑战

- 数据不足
- 数据集质量低，不具有代表性

- 不相关特征
- 对训练集过拟合 (Overfitting)
- 欠拟合 (Underfitting)
- 数据不匹配 (数据域)

可能有用的术语:

先验知识: priori knowledge

领域自适应: Domain Adaptation

(非) 语义特征 (non-) semantic feature