有监督学习方法

回归

- 回归的结果是得到一个具体的值，且值是连续的
- 回归的目的是定量
- 回归的目的是为了找到最优拟合，通过回归算法得到是一个最优拟合线，这个线条可以最好的接近数据集中的各个点。

模型

- 线性回归
  - 建立一个线性模型来预测连续型变量的值
  - 计算简单，易于解释，对于线性关系适用。但对于非线性数据表现差。
- Lasso回归
  - 通过加入正则化项来防止过拟合
  - 能够处理多重共线性，防止过拟合。但需要调节正则化参数。
- 决策树回归
  - 使用决策树来对连续型变量进行预测
  - 易于理解和解释，能够处理非线性数据。但容易过拟合，对噪声和异常值敏感。
- 支持向量机回归
  - 寻找一个最优的超平面来拟合数据，并用于预测连续型变量的值
  - 在高维空间中效果好，对于小样本数据表现良好。但对于大规模数据和非线性问题需要调整参数。

分类

- 分类的结果是得到一个类别，是离散的
- 分类的目的是定性，把回归的结果用某个二值函数处理就可以得到分类结果
- 分类的目的是为了寻找决策边界，即分类算法得到是一个决策面，用于对数据集中的数据进行分类。

模型

- 决策树
  - 根据特征对数据进行分割，从而将数据划分为不同的类别
  - 易于理解和解释，对缺失值不敏感，能够处理非线性数据。但容易过拟合，对噪声和异常值敏感
- 逻辑回归
  - 预测观测结果的概率，并基于一个阈值进行分类
  - 计算效率高，输出结果有概率解释，适用于线性可分问题。但对于非线性数据表现不佳
- 支持向量机（SVM）
  - 找到一个最优的超平面来对数据进行分割
  - 在高维空间中效果好，对于小样本数据表现良好。但对于大规模数据和非线性问题需要调整参数
- 朴素贝叶斯
  - 基于特征之间的条件独立性，计算每个类别的概率，并选择具有最高概率的类别
  - 计算简单，适用于大规模数据集，对于多类别问题也有效。但假设特征之间条件独立，可能导致精度降低
- K最近邻算法（KNN）
  - 根据最近邻的类别对数据进行分类
  - 简单易用，对于多分类问题适用，对于异常值不敏感。但计算量大，对于维度高的数据不适用
- 随机森林
  - 基于多个决策树的集成学习方法来进行分类
  - 对于高维数据和大规模数据表现良好，能够处理非线性关系，对于缺失值和异常值不敏感。但模型复杂度高，训练时间长

Presented with xmind