KNN

- 模型
  - 获得一组已知特征值和类别的样本数据
  - 当输入一个无标签数据，统计该数据最近的K个数据，以数量最多的类别为分类结果
  - 是一种判别模型
- 策略
  - 模型唯一的参数为K值，我们选择分类误差最小的K值
  - 风险函数：$\frac{1}{N}\sum_{i=1}^{N} I(f(x_i;K)! = y_i)$
    - N为选择的验证集样本数
- 算法
  - K值选择
    - K值说明
      - K=N时，总体多数类为预测结果，无意义
      - K=1时，相当于以最近的1个样本为预测结果，噪音大
    - 交叉验证，选择最优K值
    - 一般k<=sqrt(n)
  - 距离计算
    - 欧氏距离
    - 曼哈顿距离
    - 核函数：将低维数据映射到高维空间中使之线性可分
  - 一般计算流程
    - 获取原始数据
    - 特征值归一化
    - 计算已知数据与当前预测数据的距离
    - 将计算结果从小到大排序
    - 选择前K个结果，以类别数最多的作为预测值(分类)；或计算前K个样本的均值(回归)
- API
  - sklearn.neighbors.KNeighborsClassifier()
  - sklearn.neighbors.KNeighborsRegressor()
- 主要应用
  - 分类：K个近邻中类别数量最多的一类
  - 回归：K个近邻的均值
  - 异常值检测：K个近邻距离均值大于设定阈值则可视为异常值
- 优劣分析
  - 优点
    - 没有前提假设，适应性强
    - 适用于高维度数据分类
    - 原理简单，可解释性强
    - 适用于异常值检测
  - 缺点
    - 计算量大，特别在遍历计算K值时比较费时