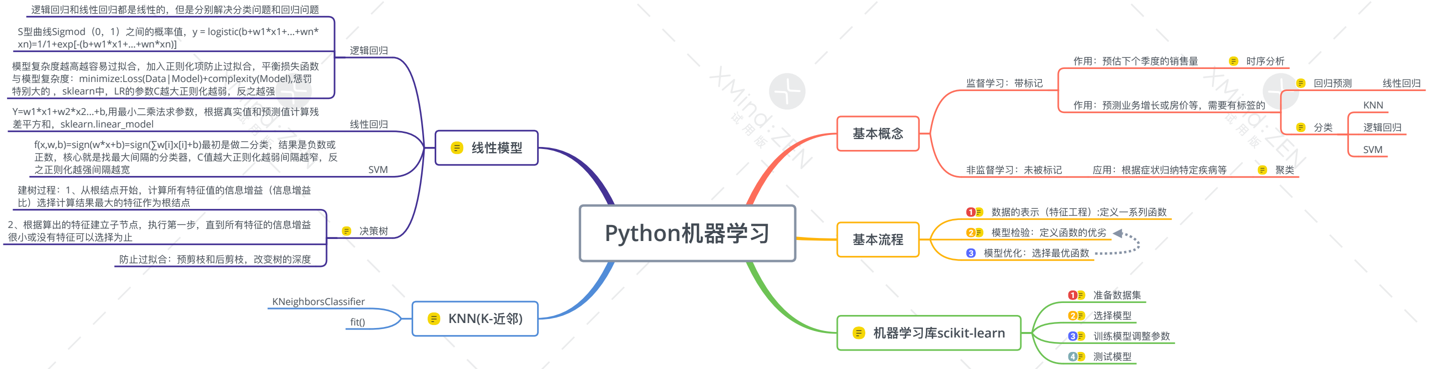

Python机器学习

- **基本概念**
  - 监督学习：带标记
    - 作用：预估下个季度的销售量
      - 时序分析
    - 作用：预测业务增长或房价等，需要有标签的
      - 回归预测
        - 线性回归
      - 分类
        - KNN
        - 逻辑回归
        - SVM
  - 非监督学习：未被标记
    - 应用：根据症状归纳特定疾病等
      - 聚类
- **基本流程**
  - ① 数据的表示（特征工程）:定义一系列函数
  - ② 模型检验：定义函数的优劣
  - ③ 模型优化：选择最优函数
- **机器学习库scikit-learn**
  - ① 准备数据集
  - ② 选择模型
  - ③ 训练模型调整参数
  - ④ 测试模型
- **线性模型**
  - 逻辑回归
    - 逻辑回归和线性回归都是线性的，但是分别解决分类问题和回归问题
    - S型曲线Sigmod（0，1）之间的概率值，y = logistic(b+w1*x1+...+wn*xn)=1/1+exp[-(b+w1*x1+...+wn*xn)]
    - 模型复杂度越高越容易过拟合，加入正则化项防止过拟合，平衡损失函数与模型复杂度：minimize:Loss(Data|Model)+complexity(Model),惩罚特别大的， sklearn中，LR的参数C越大正则化越弱，反之越强
  - 线性回归
    - Y=w1*x1+w2*x2...+b,用最小二乘法求参数，根据真实值和预测值计算残差平方和，sklearn.linear_model
  - SVM
    - f(x,w,b)=sign(w*x+b)=sign(∑w[i]x[i]+b)最初是做二分类，结果是负数或正数，核心就是找最大间隔的分类器，C值越大正则化越弱间隔越窄，反之正则化越强间隔越宽
  - 决策树
    - 建树过程：1、从根结点开始，计算所有特征值的信息增益（信息增益比）选择计算结果最大的特征作为根结点
    - 2、根据算出的特征建立子节点，执行第一步，直到所有特征的信息增益很小或没有特征可以选择为止
    - 防止过拟合：预剪枝和后剪枝，改变树的深度
- **KNN(K-近邻)**
  - KNeighborsClassifier
  - fit()

试用版 XMind:ZEN 试用版