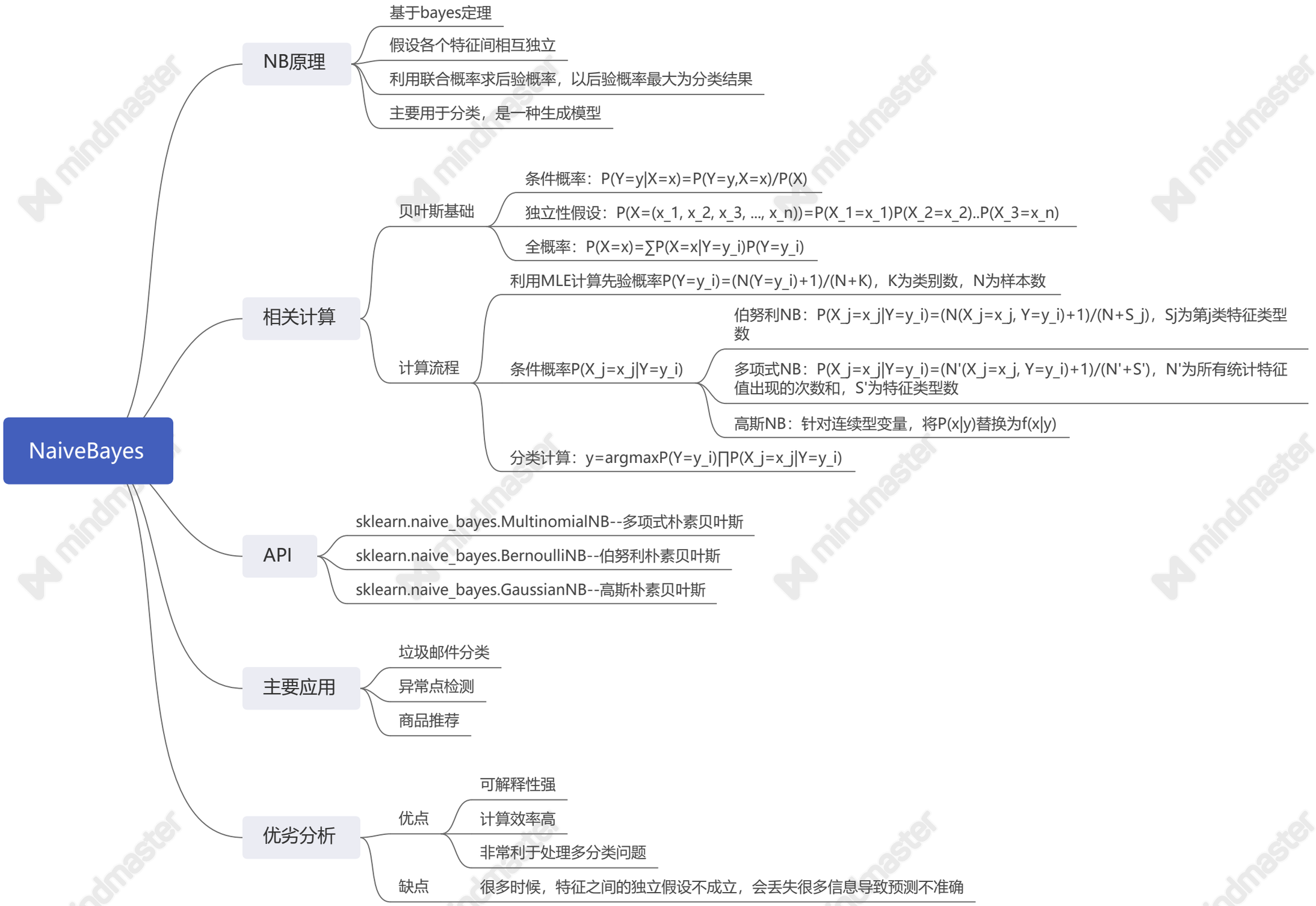
NaiveBayes
NB原理
基于bayes定理
假设各个特征间相互独立
利用联合概率求后验概率，以后验概率最大为分类结果
主要用于分类，是一种生成模型
相关计算
贝叶斯基础
条件概率：P(Y=y|X=x)=P(Y=y,X=x)/P(X)
独立性假设：P(X=(x_1, x_2, x_3, ..., x_n))=P(X_1=x_1)P(X_2=x_2)..P(X_3=x_n)
全概率：P(X=x)=∑P(X=x|Y=y_i)P(Y=y_i)
计算流程
利用MLE计算先验概率P(Y=y_i)=(N(Y=y_i)+1)/(N+K)，K为类别数，N为样本数
条件概率P(X_j=x_j|Y=y_i)
伯努利NB：P(X_j=x_j|Y=y_i)=(N(X_j=x_j, Y=y_i)+1)/(N+S_j)，Sj为第j类特征类型数
多项式NB：P(X_j=x_j|Y=y_i)=(N'(X_j=x_j, Y=y_i)+1)/(N'+S')，N'为所有统计特征值出现的次数和，S'为特征类型数
高斯NB：针对连续型变量，将P(x|y)替换为f(x|y)
分类计算：y=argmaxP(Y=y_i)∏P(X_j=x_j|Y=y_i)
API
sklearn.naive_bayes.MultinomialNB--多项式朴素贝叶斯
sklearn.naive_bayes.BernoulliNB--伯努利朴素贝叶斯
sklearn.naive_bayes.GaussianNB--高斯朴素贝叶斯
主要应用
垃圾邮件分类
异常点检测
商品推荐
优劣分析
优点
可解释性强
计算效率高
非常利于处理多分类问题
缺点
很多时候，特征之间的独立假设不成立，会丢失很多信息导致预测不准确