一、基础知识：

Naïve Bayes，即朴素贝叶斯分类器，有坚实的理论基础——贝叶斯定理。贝叶斯定理基于条件概率，条件概率 P(A|B)表示在事件 B 已经发生的前提下，事件 A 发生的概率，即$P(A|B)=\frac{P(AB)}{P(B)}$，贝叶斯定理通过 P(A|B)来求 P(B|A)：$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$，其中 P(A)由全概率公式可分解为：$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i)$。

假设给定训练数据集(X, Y)，其中每个样本 x 都包括 n 维特征，即$x=(x_1,x_2,\dots,x_n)$，类标记集合含有 k 中类别，即$y=(y_1,y_2,\dots,y_n)$。对于测试集样本 x，为判断其类别，从概率的角度来看，就是 x 属于 k 个类别中哪个概率最大，问题就变成找出$P(y_1|x),P(y_2|x),\dots,P(y_k|x)$中最大的项，即求出后验概率最大的输出：$\arg\max_{y_k}P(y_k|x)$。由贝叶斯定理可知：$P(y_k|x)=\frac{P(x|y_k)P(y_k)}{\sum_{k=1}^{n}P(x|y_k)P(y_k)}$。

分子中的$P(y_k)$是先验概率，可直接根据训练集数据计算得出，而条件概率$P(x|y_k)$有指数级数量的参数，假设第 j 维特征x_j可取值有S_j个，$j=1,2,3,\dots,n$，y 可取值有 K 个，那么参数个数为$K\prod_{j=1}^{n}S_j$。

朴素贝叶斯对条件概率作了条件独立性假设，即各个维度的特征$x_1,x_2,\dots,x_n$相互独立，在这个假设下，条件概率：$P(x|y_k)=P(x_1,x_2,\dots,x_n|y_k)=\prod_{i=1}^{n}P(x_i|y_k)$，如此，参数规模降为$\sum_{i=1}^{n}S_iK$，那么$P(y_k|x)=\frac{P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}$，于是朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_{y_k}P(y_k|x)=arg\max_{y_k}\frac{P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}$$

在计算先验概率和条件概率时，需要做平滑处理：

$$P(y_k)=\frac{N_{y_k}+a}{N+ka}$$

$$P(x_i|y_k)=\frac{N_{y_k,x_i}+a}{N_{y_k}+na}$$

其中，N 为总样本个数，k 为总类别个数，N_{y_k}是类别为y_k的样本个数，a 为平滑值，n 为特征的维数，N_{y_k,x_i}是类别为y_k的样本中，第 i 维特征的值是x_i的样本个数。

在实际实现的过程中，考虑到$P(y_k|x)$中分母都为$P(x)$，所以在比较时可以忽略分母而只考虑分子。考虑到大量的概率浮点数乘法运算，为避免 floating-point

underflow 问题，将乘法转化为取 log 再相加的运算：

$$y = f(x) = arg\max_{y_k} P(y_k|x) = arg\max_{y_k}(\log P(y_k) + \sum_{i=1}^{n} \log P(x_i|y_i))$$