# 朴素贝叶斯（native Bayes）

$$P(Y|X) = \frac{P(X,Y)}{P(X)} = \frac{P(Y)P(X|Y)}{\sum_Y P(Y)P(X|Y)}$$

将输入 x 分到后延概率最大的类 y

$$y = argmax_{c_k} P(Y = c_k) \prod_{j=1}^{n} P(X_j = x^{(j)}|Y = c_k)$$

**注意点：**

1 朴素贝叶斯对条件概率分布做了条件独立性的假设。

$$P(X = x|Y = c_k) = P\left(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)} \middle| Y = c_k\right)$$

$$= \prod_{j=1}^{n} P(X^{(j)} = x^{(j)}|Y = c_k)$$

**相关知识点：**

1 先验概率与后验概率

事情还没有发生,要求这件事情发生的可能性的大小,是先验概率.

事情已经发生,要求这件事情发生的原因是由某个因素引起的可能性的大小,是后验概率.

先验概率是指根据以往经验和分析得到的概率，如全概率公式，它往往作为“由因求果”问题中的“因”出现。后验概率是指在得到“结果”的信息后重新修正的概率，如贝叶斯公式中的，是“执果寻因”问题中的“因”。先验概率与后验概率有不可分割的联系，后验概率的计算要以先验概率为基础。

2 利用极大似然估计计算的时候可能出现概率值为 0 的情况，会影响到后验概率的结果，使分类产生偏差。因此采取增加$\lambda$的方式，称为拉普拉斯平滑。

**优点：**

1 因为独立性的假设，模型包含的条件概率的数量大为减少，因此高效，且易于实现。

2 对小规模的数据表现很好，适合多分类任务，适合增量式训练。

**缺点：**

1 因为独立性的假设，牺牲了一定的分类准确率，分类的性能不一定高。

2 对输入数据的表达形式很敏感。

# KNN(k-nearest neighbor)

**注意点：**

1 K 值的减小意味着整体模型变得复杂，容易发生过拟合。K 值变大意味着整体模型变得简单。 K 值通常选取一个比较小的数据，采用交叉验证法来选取最优的 K 值。常用的分类居然侧规则是对数表决，对应于经验风险最小化。
2 KNN 的基本做法是：对给定的训练实例点和输入实例点，首先确定输入实例点的 k 个最近邻训练实例点，然后利用这 k 个训练实例点的类的多数来预测输入实例点的类。
3 KNN 模型对应于基于训练数据集对特征空间的一个划分。KNN 中，当训练集，距离度量，k 值和分类决策规则确定后，其结果唯一确定。
4 KNN 的实现需要考虑如何快速搜索 k 个最近邻点。kd 树是一种便于对 k 维空间中的数据进行快速检索的数据结构。kd 树是二叉树，表示对 k 维空间的一个划分，其每个节点对应于 k 维空间划分的一个超矩形区域。利用 kd 树可以省去对大部分数据点的搜索，从而减少搜索的计算量。kd 树的平均计算复杂度是 O(logN).

**相关知识点：**

**优点：**

1 思想简单，理论成熟，既可以用来做分类也可以用来做回归。
2 可用于非线性分类。
3 训练时间复杂度为 O(n).
4 准确度高，对数据没有假设，对 outlier 不敏感。

**缺点：**

1 计算量大
2 样本不平衡（即有些类别的样本数量很多，而其他样本的数量很少）
3 需要大量的内存