2.4 朴素贝叶斯分类器

在实际问题中，所得到的特征之间往往存在一些关系，如果能够利用这些关系，可能更容易解决分类问题。特征之间的一个重要关系就是条件独立关系，而朴素贝叶斯分类器就是利用条件独立性的一个分类器。

简单来说，朴素贝叶斯分类器（Naïve Bayes）假定在给定样本类别的情况下，所有特征都是独立的。例如，当给定一个水果是苹果后，则假定它的颜色是红色的概率和形状是圆形的概率之间是独立的。当然，这是对实际复杂问题的一种简化。

我们可以使用贝叶斯网络来表示朴素贝叶斯分类器，如图 2.x 。贝叶斯网络是一种概率图模型，可以很好地表示变量之间条件独立性。图模型中每一个节点表示一个随机变量，变量之间的箭头连线表示因果关系。贝叶斯网络也被称之为因果网络（causal network），或贝叶斯信念网络（Bayesian Belief Network）。例如，图 2.x 表示变量w是变量x_1的因，变量x_1是变量w的果。根据贝叶斯网络的表示方法知道，在已知变量w的情况下，变量$x_1, x_2, \dots, x_d$是条件独立的，即：

$$p(x_1, x_2, \dots, x_d|w) = p(x_1|w)p(x_2|w) \dots p(x_d|w) \quad （2\text{-xxx}）$$

在一个分类问题中，如果特征个数很多，那么使用贝叶斯决策时就需要计算一组随机变量的条件联合分布。而在朴素贝叶斯网络中，这一组随机变量的条件联合分布可以用单个随机变量的条件分布的乘积来代替。

在下面一章的学习中我们会知道，对于多个变量构成的条件联合分布进行估计会有很大的困难。这表现在：一来这需要大量的样本。而实际中我们很难得到足够的样本来估计一个高维的概率分布；另一个是要存储这个分布有时需要大量的存储空间，例如离散变量的概率分布。而当条件独立性满足时，利用公式（2-xxx）可以有效解决上面两个困难。

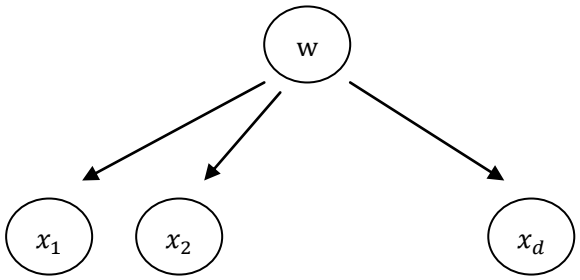

图 2.x

在实际中，上面的条件独立性假设可能是不满足的。但是朴素贝叶斯分类器在实际应用中有时非常有效。公式（2-xxx）意味着，条件联合概率密度可以通过一系列的一维概率分布的估计而完成。这样做有利于缓解由于样本不足带来的问题，而实际问题中往往是特征很多而样本很不够。和所有基于最大后验决策准则的分类器一样，只要正确类别的概率大于其他类别的概率，那么该分类器就能得到正确的分类结果。因此类条件概率分布有时候不需要估计的非常准确。