**[朴素贝叶斯与贝叶斯信念网络](http://blog.csdn.net/sealyao/article/details/6537670" \o "朴素贝叶斯与贝叶斯信念网络)**

贝叶斯分类是统计学分类方法。它们可以预测类成员关系的可能性，如给定样本属于一个特定类的概率。

贝叶斯定理是就是在给定的数据概率来表示未知的后验概率。比如已知某水果是红色的情况下，判断该水果有多大的概率是苹果，用数学符号表示就是[clip_image002](http://hi.csdn.net/attachment/201106/11/0_1307756315VfPK.gif)(后验概率)，其中X表示“这个水果是红色的”，H表示“这个水果是苹果”。这个概率我们是不知道的，但是如果我们有大量的水果样本，就可以计算水果样本中的统计信息来逼近这个概率。

下面的三个概率都是可以通过样本统计简单计算得到的：“一个水果的颜色是红色”的概率P(X)；“一个水果是苹果” 的概率 P(H)(先验概率)；“如果一个水果是苹果那么它的颜色是红色”的概率[clip_image004](http://hi.csdn.net/attachment/201106/11/0_13077563159W2I.gif)。

我们可以推理得到：

[clip_image002[1]](http://hi.csdn.net/attachment/201106/11/0_1307756315HP0W.gif)=[clip_image006](http://hi.csdn.net/attachment/201106/11/0_1307756316jslV.gif)

=[clip_image008](http://hi.csdn.net/attachment/201106/11/0_13077563162Zd7.gif)/[clip_image010](http://hi.csdn.net/attachment/201106/11/0_13077563167PAi.gif)

=[clip_image012](http://hi.csdn.net/attachment/201106/11/0_13077563163Fcv.gif)

=[clip_image014](http://hi.csdn.net/attachment/201106/11/0_1307756316imP7.gif)

抽象得到的公式[clip_image016](http://hi.csdn.net/attachment/201106/11/0_1307756316v4Vv.gif)就是著名的贝叶斯定理。

贝叶斯分类的基本思路就是把[clip_image018](http://hi.csdn.net/attachment/201106/11/0_13077563178w3T.gif)、[clip_image020](http://hi.csdn.net/attachment/201106/11/0_13077563170MEw.gif)、[clip_image022](http://hi.csdn.net/attachment/201106/11/0_1307756317ed1R.gif)、[clip_image024](http://hi.csdn.net/attachment/201106/11/0_1307756317gM88.gif)……的概率都算出来，当有未知分类的样本时，就认为这个样本是后验概率最大的那个分类的。

**朴素贝叶斯分类**

实际情况要比上面的那个例子复杂一些，因为一个事物的属性是多维的，一个水果的属性可能就要包括：颜色、形状、重量、体积……。

实际上的分类可能是要算这样一个概率：

[clip_image026](http://hi.csdn.net/attachment/201106/11/0_1307756317c2O0.gif)

如果属性相互独立，那么

[clip_image028](http://hi.csdn.net/attachment/201106/11/0_1307756318A1CD.gif)

朴素贝叶斯的朴素之处在于不管属性独不独立，都按独立来算，这样可以使运算大大简化。

归纳一下朴素贝叶斯的运算流程：

1. 每个数据样本用一个n维特征向量[clip_image030](http://hi.csdn.net/attachment/201106/11/0_1307756318rj3s.gif)表示，描述由属性[clip_image032](http://hi.csdn.net/attachment/201106/11/0_1307756318pN34.gif)对样本的n个度量。

2. 假定有m个类[clip_image034](http://hi.csdn.net/attachment/201106/11/0_1307756318nPWe.gif)。给定一个未知的数据样本X（即，没有类标号），分类法将预测X属于具有最高后验概率（条件X下）的类。即，朴素贝叶斯分类将未知的样本分配给类Ci ，当且仅当：

[clip_image036](http://hi.csdn.net/attachment/201106/11/0_1307756318Zp1y.gif)

这样，我们最大化[clip_image038](http://hi.csdn.net/attachment/201106/11/0_1307756319gg6S.gif)。其[clip_image038[1]](http://hi.csdn.net/attachment/201106/11/0_1307756319gNon.gif)最大的类Ci称为最大后验假定。根据贝叶斯定理：

[clip_image041](http://hi.csdn.net/attachment/201106/11/0_130775631933B7.gif)

3．由于P(X) 对于所有类为常数，只需要[clip_image043](http://hi.csdn.net/attachment/201106/11/0_13077563190244.gif)最大即可。如果类的先验概率未知，则通常假定这些类是等概率的；即，[clip_image045](http://hi.csdn.net/attachment/201106/11/0_1307756319zbLX.gif)。并据此对只[clip_image038[2]](http://hi.csdn.net/attachment/201106/11/0_13077563209Bi9.gif)最大化。否则，我们最大化[clip_image043[1]](http://hi.csdn.net/attachment/201106/11/0_1307756320RoOO.gif)。注意，类的先验概率可以用[clip_image049](http://hi.csdn.net/attachment/201106/11/0_13077563201gos.gif)计算；其中，si是类C中的训练样本数，而s是训练样本总数。

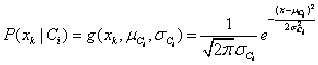
4．给定具有许多属性的数据集，计算[clip_image051](http://hi.csdn.net/attachment/201106/11/0_13077563206J62.gif)的开销可能非常大。为降低计算[clip_image051[1]](http://hi.csdn.net/attachment/201106/11/0_13077563200mZ8.gif)的开销，可以做类条件独立的朴素假定。给定样本的类标号，假定属性值条件地相互独立。即，在属性间，不存在依赖关系。这样，

[clip_image054](http://hi.csdn.net/attachment/201106/11/0_1307756321aO9v.gif)

概率[clip_image056](http://hi.csdn.net/attachment/201106/11/0_1307756321RRbj.gif),[clip_image058](http://hi.csdn.net/attachment/201106/11/0_1307756321ZVTv.gif),...,[clip_image060](http://hi.csdn.net/attachment/201106/11/0_1307756321huuc.gif)可以由训练样本估值，其中，

(a) 如果Ak是分类属性，则[clip_image062](http://hi.csdn.net/attachment/201106/11/0_1307756321KQEM.gif)；其中sik 是在属性Ak 上具有值xk 的类Ci 的训练样本数，而si 是Ci中的训练样本数。

(b) 如果是连续值属性，则通常假定该属性服从高斯分布。因而，

[](http://hi.csdn.net/attachment/201106/11/0_13077563227ag3.gif)

其中，给定类Ci的训练样本属性Ak的值，[clip_image066](http://hi.csdn.net/attachment/201106/11/0_1307756322573V.gif)是属性Ak的高斯密度函数，而[clip_image068](http://hi.csdn.net/attachment/201106/11/0_13077563227oEY.gif)分别为平均值和标准差。

5．为对未知样本X分类，对每个类Ci，计算[clip_image043[2]](http://hi.csdn.net/attachment/201106/11/0_13077563232hV7.gif)。样本X被指派到类Ci，当且仅当：

[clip_image071](http://hi.csdn.net/attachment/201106/11/0_1307756323TJQq.gif)

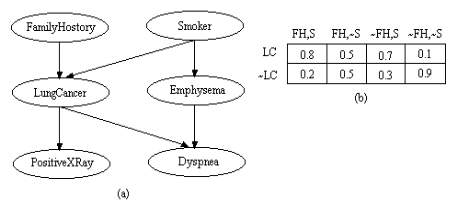
换言之，X被指派到其[clip_image043[3]](http://hi.csdn.net/attachment/201106/11/0_1307756323llBg.gif)最大的类Ci。

**贝叶斯信念网络**

朴素贝叶斯假定属性之间是独立的。贝叶斯信念网络说明联合概率分布，它提供一种因果关系的图形，可以在其上进行学习。

信念网络由两部分定义。第一部分是有向无环图，其每个结点代表一个随机变量，而每条弧代表一个概率依赖。如果一条弧由结点Y到Z，则Y是Z的双亲或直接前驱，而Z是Y的后继。第二部分是每个属性一个条件概率表（CPT）。

下面是一个LungCancer的CPT。

[](http://hi.csdn.net/attachment/201106/11/0_1307756323L8lT.gif)

在贝叶斯信念网络中对应于属性或变量[clip_image075](http://hi.csdn.net/attachment/201106/11/0_130775632394lb.gif)的任意元组[clip_image077](http://hi.csdn.net/attachment/201106/11/0_13077563230r6z.gif)的联合概率由下式计算：

[clip_image079](http://hi.csdn.net/attachment/201106/11/0_1307756324dEca.gif)

如上图，对于FamilyHistory，Smoker，LungCancer这三个属性，用朴素贝叶斯计算，得到的联合概率是[clip_image081](http://hi.csdn.net/attachment/201106/11/0_1307756324Z64F.gif)。

但是如果用贝叶斯信念网络计算得到的联合概率将会是：

[clip_image083](http://hi.csdn.net/attachment/201106/11/0_1307756324DRas.gif)，会更为准确。

贝叶斯信念网络的问题

1、如果贝叶斯信念网络的网络结构和所有数值都是给定的，那么可以直接进行计算。但是，数据是隐藏的，比如上图中的 FamilyHistory/Somker到LungCancer的条件概率是未知的，只是知道存在这样的依存关系，这时就需要进行条件概率的估算。梯度 训练算法和EM算法常被用于处理此问题。

2、贝叶斯网络的数据结构可能是未知的，这时就需要根据已知数据启发式学习贝叶斯网络结构。K2算法可用于解决此问题。

**梯度训练算法**

梯度训练是用于解决信念网络中隐藏数据问题的，就是已知上图（a），但是不知道上图（b）。

设D是d个训练样本[clip_image085](http://hi.csdn.net/attachment/201106/11/0_1307756324L1lc.gif)的集合，[clip_image087](http://hi.csdn.net/attachment/201106/11/0_1307756324q05U.gif)是具有双亲[clip_image089](http://hi.csdn.net/attachment/201106/11/0_1307756325AynF.gif)= [clip_image091](http://hi.csdn.net/attachment/201106/11/0_13077563257jdj.gif)的变量[clip_image093](http://hi.csdn.net/attachment/201106/11/0_1307756325O99G.gif) = [clip_image095](http://hi.csdn.net/attachment/201106/11/0_1307756325sl84.gif)的CPT项。例如，如果[clip_image087[1]](http://hi.csdn.net/attachment/201106/11/0_1307756325pp3p.gif)是上图(b)左上角的CPT项，则[clip_image093[1]](http://hi.csdn.net/attachment/201106/11/0_1307756325Yaot.gif)是LungCancer；[clip_image095[1]](http://hi.csdn.net/attachment/201106/11/0_13077563265p54.gif)是其值“yes”； [clip_image089[1]](http://hi.csdn.net/attachment/201106/11/0_1307756326KeHA.gif)列出[clip_image093[2]](http://hi.csdn.net/attachment/201106/11/0_1307756326JcRk.gif)的双亲结点{FamilyHistory, Smoker}；而[clip_image091[1]](http://hi.csdn.net/attachment/201106/11/0_1307756326RikX.gif)列出双亲结点的值{“yes”, “yes”}。[clip_image087[2]](http://hi.csdn.net/attachment/201106/11/0_1307756326uhHR.gif)可以看作权，类似于神经网络中隐藏单元的权。权的集合记作[clip_image097](http://hi.csdn.net/attachment/201106/11/0_1307756327r11Q.gif)。

梯度训练算法就是求出最为满足[clip_image085[1]](http://hi.csdn.net/attachment/201106/11/0_1307756327809E.gif)训练版本集的权的集合[clip_image097[1]](http://hi.csdn.net/attachment/201106/11/0_1307756327In3i.gif)，用数学公式表示就是[clip_image100](http://hi.csdn.net/attachment/201106/11/0_130775632761Hx.gif)最大（[clip_image102](http://hi.csdn.net/attachment/201106/11/0_13077563279cnN.gif)就表示[clip_image104](http://hi.csdn.net/attachment/201106/11/0_1307756327G58J.gif)）。

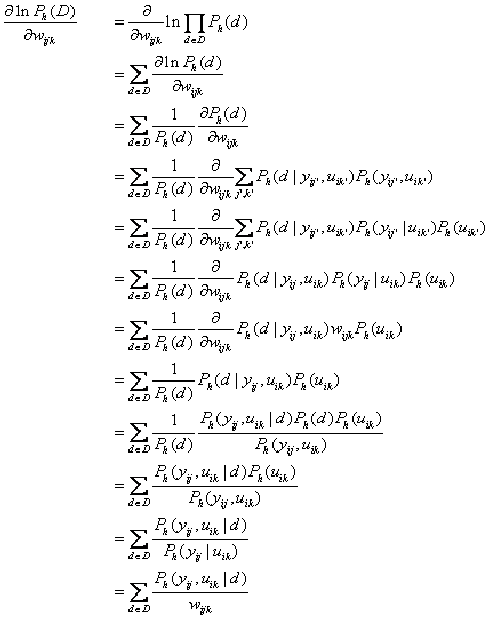
具体的算法：

1、就是对每一个[clip_image087[3]](http://hi.csdn.net/attachment/201106/11/0_1307756328D22Z.gif)取偏导数，[clip_image107](http://hi.csdn.net/attachment/201106/11/0_13077563285U5v.gif)

2、更新权值：沿梯度方向前进一小步。[clip_image109](http://hi.csdn.net/attachment/201106/11/0_1307756328byRk.gif)，其中l是学习率，是一个小常数。

3、由于权值[clip_image087[4]](http://hi.csdn.net/attachment/201106/11/0_1307756328jpJP.gif)是概率值，它们必须在0.0和1.0之间，并且对于所有的i,k，[clip_image112](http://hi.csdn.net/attachment/201106/11/0_13077563286JY8.gif)必须等于1。在权值被式更新后，可以对它们重新规格化来保证这一条件。[clip_image114](http://hi.csdn.net/attachment/201106/11/0_1307756329xsPN.gif)

关于[clip_image087[5]](http://hi.csdn.net/attachment/201106/11/0_1307756329nTPT.gif)的偏导数推导如下：

[](http://hi.csdn.net/attachment/201106/11/0_1307756329EzeX.gif)