北风网
IBEIFENG.COM

人工智能之机器学习

贝叶斯算法

上海育创网络科技股份有限公司

主讲人：刘老师(GerryLiu)

课程要求

- 课上课下“九字”真言
  - 认真听，**善摘录，勤思考**
  - **多温故，乐实践**，再发散
- 四不原则
  - 不懒散惰性，不迟到早退
  - 不请假旷课，不拖延作业
- 一点注意事项
  - 违反“四不原则”，不推荐就业

课程内容

- 朴素贝叶斯
- 贝叶斯网络

贝叶斯定理相关公式

- 先验概率P(A)：在不考虑任何情况下，A事件发生的概率
- 条件概率P(B|A)：A事件发生的情况下，B事件发生的概率

$$P(\mathrm{B} \mid A) = \frac{P(AB)}{P(\mathrm{A})}$$

- 后验概率P(A|B)：在B事件发生之后，对A事件发生的概率的重新评估
- 全概率：如果A和A'构成样本空间的一个划分，那么事件B的概率为：A和A'的概率分别乘以B对这两个事件的概率之和。

$$P(B) = P(A) * P(B \mid A) + P(A^{'}) * P(B \mid A^{'})$$

$$P(B) = \sum_{i=1}^{n} P(A_i) * P(B \mid A_i)$$

贝叶斯定理公式

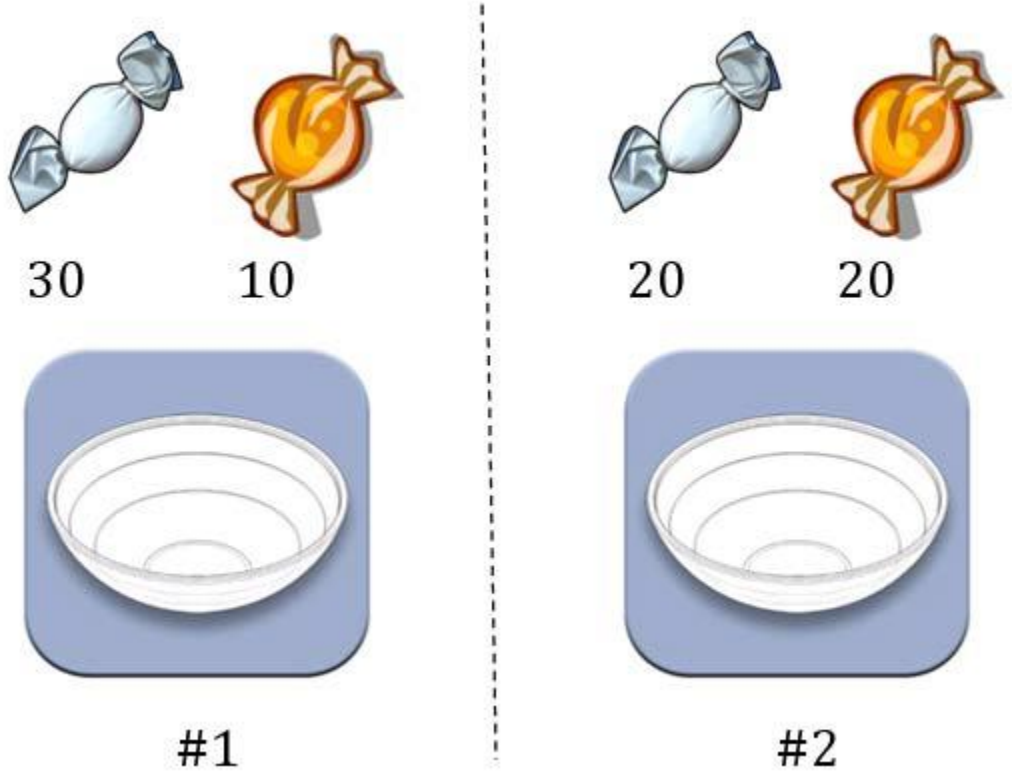
- 基于条件概率的贝叶斯定律数学公式

$$P(\mathrm{A}_j \mid B) = \frac{P(A_j B)}{P(B)} \qquad P(B) = \sum_{i=1}^{n} P(A_i) * P(B \mid A_i)$$

$$P(A_j \mid B) = \frac{P(A_j) * P(\mathrm{B} \mid A_j)}{P(\mathrm{B})} = \frac{P(A_j) * P(\mathrm{B} \mid A_j)}{\sum_{i=1}^{n} P(\mathrm{B} \mid \mathrm{A}_i) * P(\mathrm{A}_i)}$$

贝叶斯公式的应用

- 有两个碗，第一个碗中装有30个水果糖和10个巧克力糖，第二个碗中装有20个水果糖和20个巧克力糖，现在随机选择一个碗，从中取出一颗糖，发现是水果糖，请求出这颗水果糖来自一号碗的概率有多大?
- 现在随机选择一个碗，从中取出一颗糖，发现是水果糖，请问这颗糖属于哪个碗?



朴素贝叶斯算法

- 朴素贝叶斯(Naive Bayes， NB)是基于“**特征之间是独立的**”这一朴素假设，应用贝叶斯定理的监督学习算法
- 对应给定的样本X的特征向量$x_1,x_2,...,x_m$；该样本X的类别y的概率可以由贝叶斯公式得到：

$$P(y \mid x_1, x_2, \ldots x_m) = \frac{P(y)P(x_1, x_2, \ldots x_m \mid y)}{P(x_1, x_2, \ldots x_m)}$$

朴素贝叶斯算法推导

$$P(y \mid x_1, x_2, \ldots x_m) = \frac{P(y)P(x_1, x_2, \ldots x_m \mid y)}{P(x_1, x_2, \ldots x_m)}$$

- 特征属性之间是独立的，所以得到：

$$P(x_i \mid y, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m) = P(x_i \mid y)$$

- 公式优化得到：

$$P(y \mid x_1, x_2, \ldots x_m) = \frac{P(y)P(x_1, x_2, \ldots x_m \mid y)}{P(x_1, x_2, \ldots x_m)} = \frac{P(y)\prod_{i=1}^{m} P(x_i \mid y)}{P(x_1, x_2, \ldots x_m)}$$

- 在给定样本的情况下，$P(x_1, x_2, \ldots, x_m)$是常数，所以得到：

$$P(y \mid x_1, x_2, \ldots x_m) \propto P(y)\prod_{i=1}^{m} P(x_i \mid y)$$

- 从而：

$$\hat{y} = \arg\max_{y} P(y)\prod_{i=1}^{m} P(x_i \mid y)$$

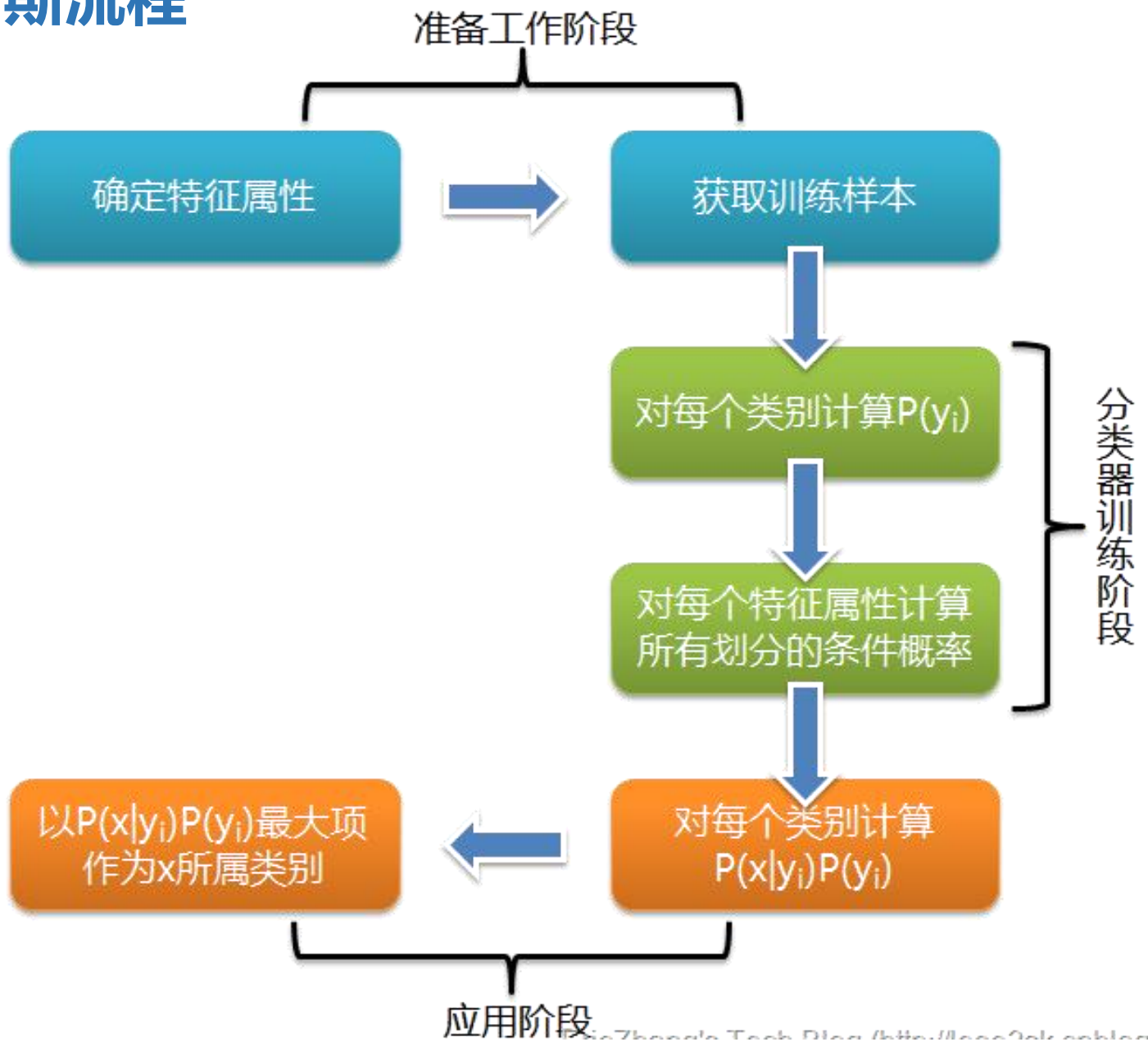
朴素贝叶斯算法流程

- 朴素贝叶斯算法流程/定义如下：
  - 设x={x_1,x_2,...,x_m}为待分类样本，其中x_i为x的一个特征属性
  - 类别集合为C={y_1,y_2,...,y_n}
  - 分别计算P(y_1|x),P(y_2|x),....,P(y_n|x)的值（贝叶斯公式）
  - 如果P(y_k|x)=max{P(y_1|x),P(y_2|x),....,P(y_n|x)},那么认为x为y_k类型

$$P(y \mid x_1, x_2, \ldots x_m) = \frac{P(y)P(x_1, x_2, \ldots x_m \mid y)}{P(x_1, x_2, \ldots x_m)}$$

$$P(y \mid x_1, x_2, \ldots x_m) \propto P(y)\prod_{i=1}^{m} P(x_i \mid y)$$

朴素贝叶斯流程


准备工作阶段
确定特征属性
获取训练样本
对每个类别计算P(yi)
对每个特征属性计算
所有划分的条件概率
分类器训练阶段
对每个类别计算
P(x|yi)P(yi)
以P(x|yi)P(yi)最大项
作为x所属类别
应用阶段

高斯朴素贝叶斯

- Gaussian Naive Bayes是指当特征属性为连续值时，而且分布服从高斯分布，那么在计算P(x|y)的时候可以直接使用高斯分布的概率公式：

$$g(x,\eta,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\eta)^2}{2\sigma^2}}$$

$$P(x_i \mid y_k)=g(x_i,\eta_{i,y_k},\sigma_{i,y_k})$$

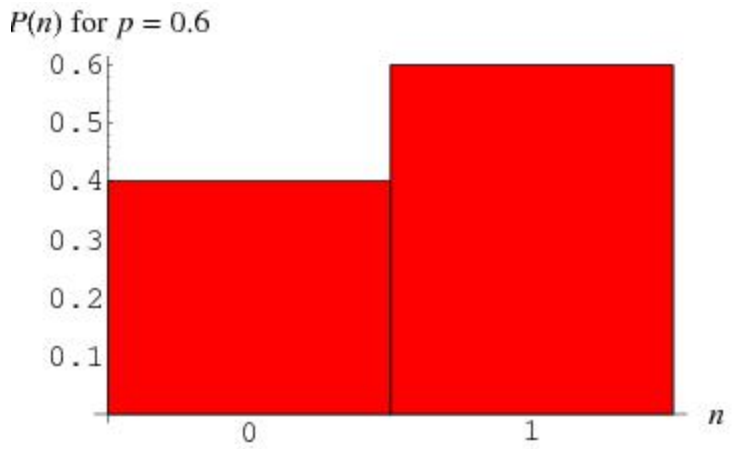
- 因此只需要计算出各个类别中此特征项划分的各个均值和标准差

伯努利朴素贝叶斯

- Bernoulli Naive Bayes是指当特征属性为连续值时，而且分布服从伯努利分布，那么在计算P(x|y)的时候可以直接使用伯努利分布的概率公式：

$$P(x_k \mid y) = P(1 \mid y)x_k + \big(1 - P(1 \mid y)\big)\big(1 - x_k\big)$$

- 伯努利分布是一种离散分布，只有两种可能的结果。1表示成功，出现的概率为p；0表示失败，出现的概率为q=1-p；其中均值为E(x)=p，方差为Var(X)=p(1-p)



多项式朴素贝叶斯

- Multinomial Naive Bayes是指当特征属性服从多项分布(特征是离散的形式的时候)，直接计算类别数目的占比作为先验概率和条件概率。

$$p(y_k)=\frac{N_{y_k}+\alpha}{N+k*\alpha} \qquad p(x_i|y_k)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+n_i*\alpha}$$

- N是总样本个数，k是总的类别个数，N_{yk}是类别为yk的样本个数，α为平滑值。
- N_{yk}是类别为yk的样本个数，n_i为特征属性xi的不同取值数目，$N_{yk,xi}$为类别yk中第i维特征的值为xi的样本个数，α为平滑值。
- 当α=1时，称为Laplace平滑，当0<α<1时，称为Lidstone平滑，α=0时不做平滑；平滑的主要作用是可以克服条件概率为0的问题。

多项式朴素贝叶斯案例理解

- 对于下列训练数据，使用多项式朴素贝叶斯方式对测试样本(2,M,L)做一个预测判断。

	1	2	3	4	5	6	7	8	9	10
x1	1	1	1	2	2	2	2	3	3	4
x2	S	M	S	L	S	S	L	L	L	S
x3	L	H	L	H	L	M	H	M	H	M
y	-1	1	1	-1	-1	-1	1	1	1	1

$$N = 10$$

$$k = 2 \; n_1 = 4$$

$$n_2 = 3 \; n_3 = 3$$

	x1=1	x1=2	x1=3	x1=4	
y=1	2	1	2	1	6
y=-1	1	3	0	0	4
	3	4	2	1	10

	x2=S	x2=M	x2=L	
y=1	2	1	3	6
y=-1	3	0	1	4
	5	1	4	10

	x3=L	x3=M	x3=H	
y=1	1	2	3	6
y=-1	2	1	1	4
	3	3	4	10

多项式朴素贝叶斯案例理解

$\alpha = 0$

- 先验概率：

$$p(y=1) = 6/10 = 0.6 \quad p(y=-1) = 4/10 = 0.4$$

- 条件概率：

$$p(x_1=1|y=1) = \frac{2}{6} \quad p(x_1=1|y=-1) = \frac{1}{4} \quad p(x_2=S|y=1) = \frac{2}{6} \quad p(x_2=S|y=-1) = \frac{3}{4}$$

$$p(x_1=2|y=1) = \frac{1}{6} \quad p(x_1=2|y=-1) = \frac{3}{4} \quad p(x_2=M|y=1) = \frac{1}{6} \quad p(x_2=M|y=-1) = 0$$

$$p(x_1=3|y=1) = \frac{2}{6} \quad p(x_1=3|y=-1) = 0 \quad p(x_2=L|y=1) = \frac{3}{6} \quad p(x_2=L|y=-1) = \frac{1}{4}$$

$$p(x_1=4|y=1) = \frac{1}{6} \quad p(x_1=4|y=-1) = 0$$

多项式朴素贝叶斯案例理解

- 条件概率：

$$p(x_3 = L|y = 1) = \frac{1}{6} \qquad p(x_3 = L|y = -1) = \frac{2}{4}$$

$$p(x_3 = M|y = 1) = \frac{2}{6} \qquad p(x_3 = M|y = -1) = \frac{1}{4}$$

$$p(x_3 = H|y = 1) = \frac{3}{6} \qquad p(x_3 = H|y = -1) = \frac{1}{4}$$

$\alpha = 0$

- 样本(2,M,L)的预测概率：

$$p(y = 1|x) \propto p(y = 1)p(x_1 = 2|y = 1)p(x_2 = M|y = 1)p(x_3 = L|y = 1) = \frac{6}{10} * \frac{1}{6} * \frac{1}{6} * \frac{1}{6} = \frac{1}{360}$$

$$p(y = -1|x) \propto p(y = -1)p(x_1 = 2|y = -1)p(x_2 = M|y = -1)p(x_3 = L|y = -1) = \frac{4}{10} * \frac{3}{4} * 0 * \frac{2}{4} = 0$$

$$\hat{y} = \arg\max_y \{p(y = 1|x), p(y = -1|x)\} = 1$$

多项式朴素贝叶斯案例理解

$\alpha = 1$

- 先验概率:

$$p(y=1)=(6+1)/(10+2*1)=7/12 \quad p(y=-1)=5/12$$

- 条件概率:

$$p(x_1=1|y=1)=\frac{3}{10} \quad p(x_1=1|y=-1)=\frac{2}{8} \quad p(x_2=S|y=1)=\frac{3}{9} \quad p(x_2=S|y=-1)=\frac{4}{7}$$

$$p(x_1=2|y=1)=\frac{2}{10} \quad p(x_1=2|y=-1)=\frac{4}{8} \quad p(x_2=M|y=1)=\frac{2}{9} \quad p(x_2=M|y=-1)=\frac{1}{7}$$

$$p(x_1=3|y=1)=\frac{3}{10} \quad p(x_1=3|y=-1)=\frac{1}{8} \quad p(x_2=L|y=1)=\frac{4}{9} \quad p(x_2=L|y=-1)=\frac{2}{7}$$

$$p(x_1=4|y=1)=\frac{2}{10} \quad p(x_1=4|y=-1)=\frac{1}{8}$$

多项式朴素贝叶斯案例理解

- 条件概率：

$$p(x_3 = L|y = 1) = \frac{2}{9} \qquad p(x_3 = L|y = -1) = \frac{3}{7}$$

$$p(x_3 = M|y = 1) = \frac{3}{9} \qquad p(x_3 = M|y = -1) = \frac{2}{7}$$

$$p(x_3 = H|y = 1) = \frac{4}{9} \qquad p(x_3 = H|y = -1) = \frac{2}{7}$$

$\alpha = 1$

- 样本(2,M,L)的预测概率概率：

$$p(y = 1|x) \propto p(y = 1)p(x_1 = 2|y = 1)p(x_2 = M|y = 1)p(x_3 = L|y = 1) = \frac{7}{12} * \frac{2}{10} * \frac{2}{9} * \frac{2}{9} = \frac{7}{1215}$$

$$p(y = -1|x) \propto p(y = -1)p(x_1 = 2|y = -1)p(x_2 = M|y = -1)p(x_3 = L|y = -1) = \frac{5}{12} * \frac{4}{8} * \frac{1}{7} * \frac{3}{7} = \frac{5}{392}$$

$$\hat{y} = \arg\max_y \{p(y = 1|x), p(y = -1|x)\} = -1$$

案例一：鸢尾花数据分类

- 使用高斯朴素贝叶斯API对鸢尾花数据进行分类操作

class sklearn.naive_bayes. **GaussianNB** ¶ [source]

Attributes:

class_prior_ : array, shape (n_classes,)

probability of each class.

各个类别的概率

class_count_ : array, shape (n_classes,)

number of training samples observed in each class.

各个类别的样本数量

theta_ : array, shape (n_classes, n_features)

mean of each feature per class

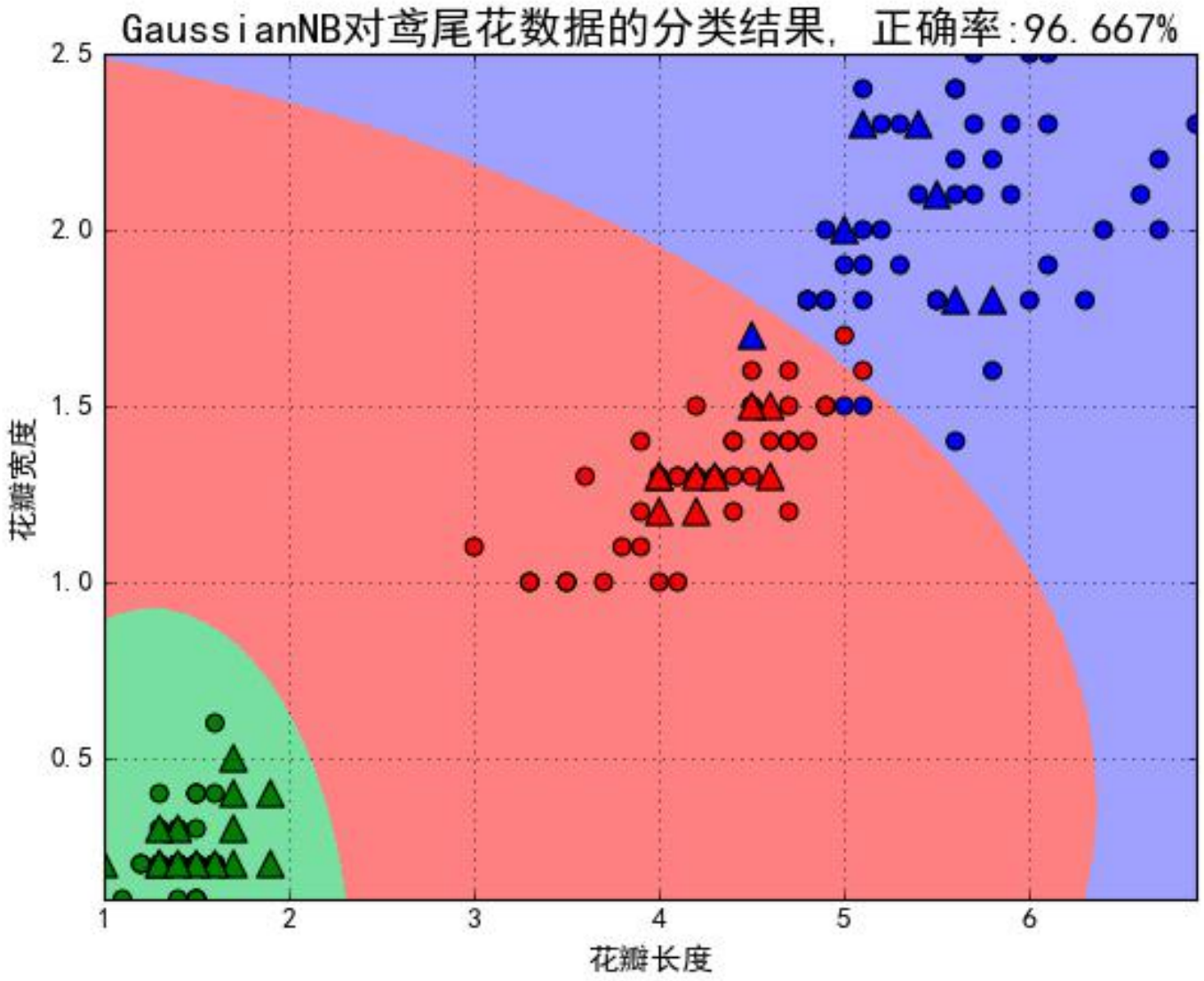
各个类别中各个特征属性的均值

sigma_ : array, shape (n_classes, n_features)

variance of each feature per class

各个类别中各个特征属性的方差

案例一：鸢尾花数据分类


GaussianNB对鸢尾花数据的分类结果，正确率:96.667%
花瓣宽度
花瓣长度
2.5
2.0
1.5
1.0
0.5
1
2
3
4
5
6

案例二：文本数据分类

- 分别使用朴素贝叶斯API和其它分类API对scikit中自带的新闻文本数据进行分类操作；比较各种不同分类算法的效果以及执行消耗时间

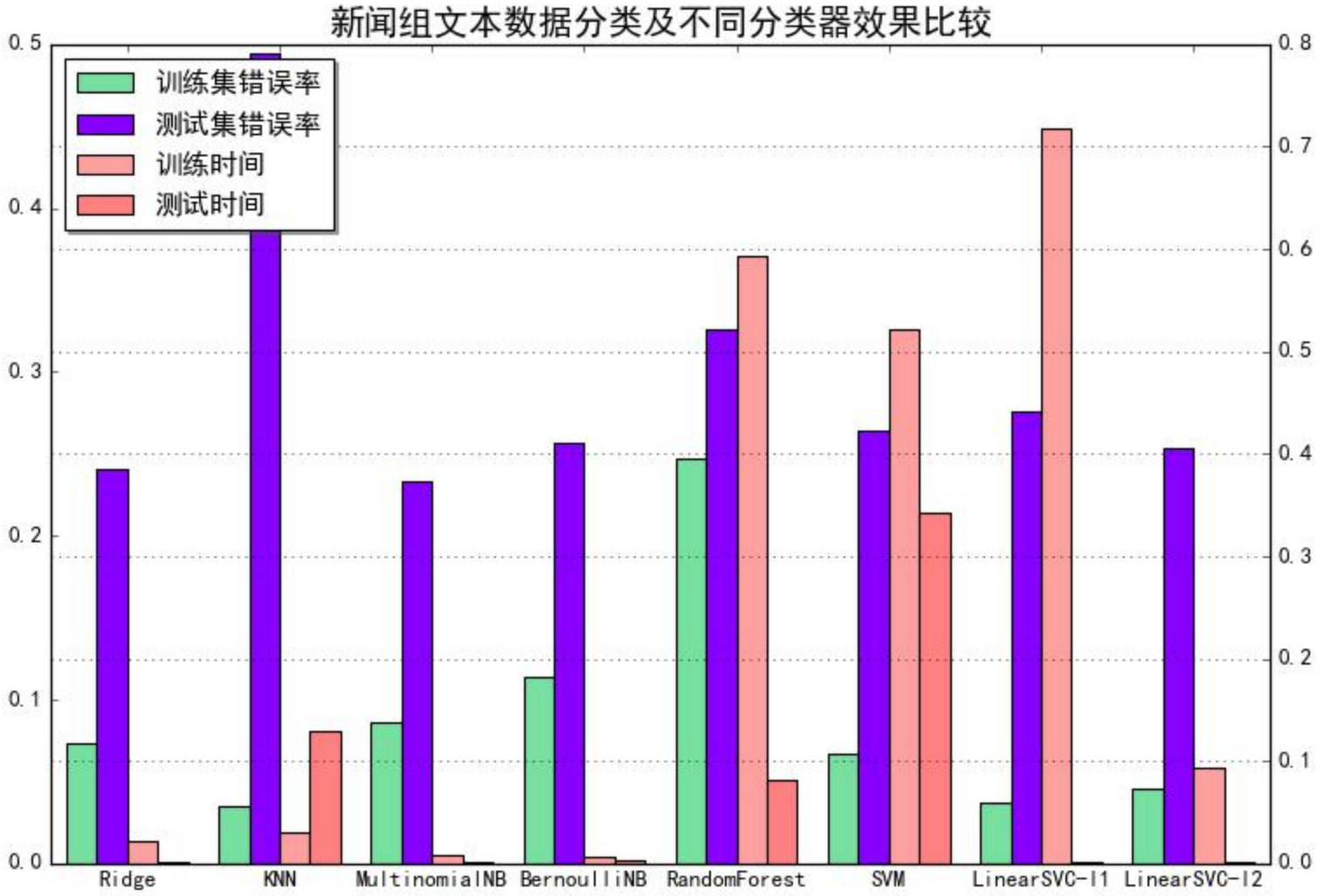
sklearn.datasets. **fetch_20newsgroups** (*data_home=None, subset='train', categories=None, shuffle=True, random_state=42, remove=(), download_if_missing=True*) ¶ [source]

class sklearn.feature_extraction.text. **TfidfVectorizer** (*input='content', encoding='utf-8', decode_error='strict', strip_accents=None, lowercase=True, preprocessor=None, tokenizer=None, analyzer='word', stop_words=None, token_pattern='(?u)\b\w\w+\b', ngram_range=(1, 1), max_df=1.0, min_df=1, max_features=None, vocabulary=None, binary=False, dtype=<class 'numpy.int64'>, norm='l2', use_idf=True, smooth_idf=True, sublinear_tf=False*) ¶ [source]

案例二：文本数据分类

- 文本类数据处理的最重要的是需要将文本数据转换为数值型数据，一般情况是将文本转换为一个向量；
  - 先计算出在文档A中各个单词出现的频率TF(term/token)
  - 在计算出文档A中的各个单词在所有文档中出现的频率DF(出现的文档/总的文档)
  - 这样的话我们可到如果一个词在当前文档中出现的频率越高，在所有文档中出现的频率越低，那么这个单词就越重要，所以可以似乎用TF/DF的值来作为单词的权重；考虑有DF的计算过程中，分母有可能为空，经常使用IDF(逆文件频率)来替代DF，此时的权重等于TF * IDF；有时候防止IDF过大以及IDF为0的情况，可以将公式换为：TF * log(IDF + 1)
- 通过使用文本数据形成的向量之间的距离来量化两个文档的相似性，一般使用夹角余弦公式

案例二：文本数据分类


新闻组文本数据分类及不同分类器效果比较
训练集错误率
测试集错误率
训练时间
测试时间
0.5
0.4
0.3
0.2
0.1
0.0
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Ridge
KNN
MultinomialNB
BernoulliNB
RandomForest
SVM
LinearSVC-l1
LinearSVC-l2

贝叶斯网络

- 把某个研究系统中涉及到的**随机变量**，根据是否条件独立绘制在一个**有向图**中，就形成了贝叶斯网络。
- 贝叶斯网络(Bayesian Network)，又称**有向无环图模型**(directed acyclic graphical model, DAG)，是一种概率图模型，根据概率图的拓扑结构，考察一组随机变量{X_1,X_2,...,X_n}及其N组**条件概率分布**(Conditional Probabililty Distributions, CPD)的性质

贝叶斯网络

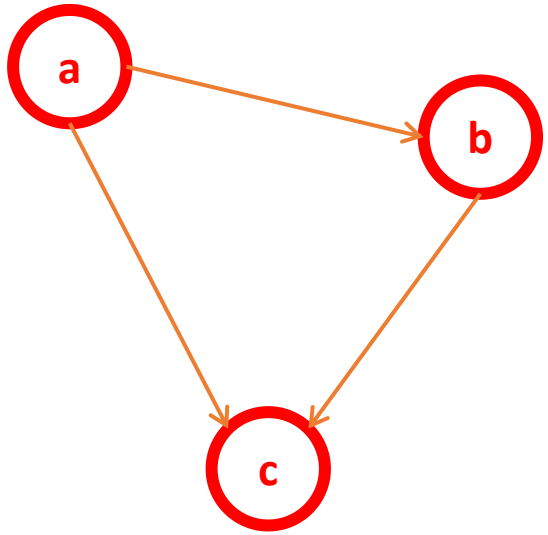
- 当多个特征属性之间存在着某种相关关系的时候，使用朴素贝叶斯算法就没法解决这类问题，那么贝叶斯网络就是解决这类应用场景的一个非常好的算法。
- 一般而言，贝叶斯网络的有向无环图中的节点表示随机变量，可以是可观察到的变量，或隐变量，未知参数等等。连接两个节点之间的箭头代表两个随机变量之间的因果关系(也就是这两个随机变量之间非条件独立)，如果两个节点间以一个单箭头连接在一起，表示其中一个节点是“因”，另外一个是“果”，从而两节点之间就会产生一个条件概率值。

贝叶斯网络

- 贝叶斯网络的关键方法是图模型，构建一个图模型我们需要把具有因果联系的各个变量用箭头连在一起。贝叶斯网络的有向无环图中的节点表示随机变量。连接两个节点的箭头代表此两个随机变量是具有因果关系的。
- 贝叶斯网络是模拟人的认知思维推理模式的，用一组条件概率以及有向无环图对不确定性因果推理关系建模

最简单的一个贝叶斯网络

$$P(a,b,c) = P(c \mid a,b)P(b \mid a)P(a)$$

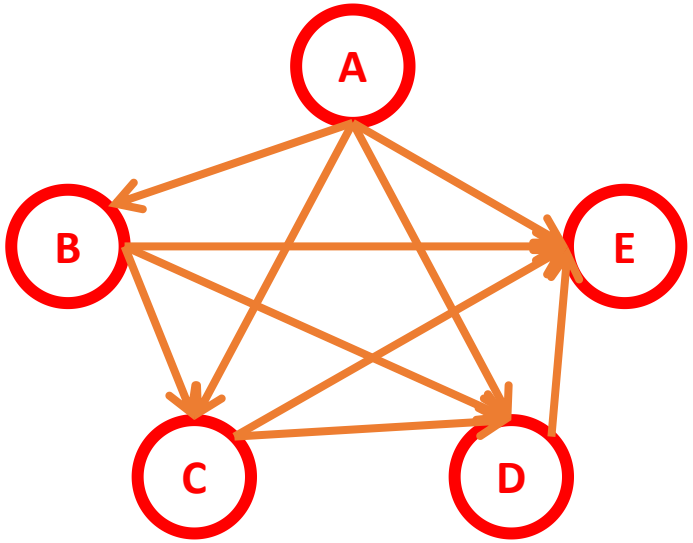


全连接贝叶斯网络

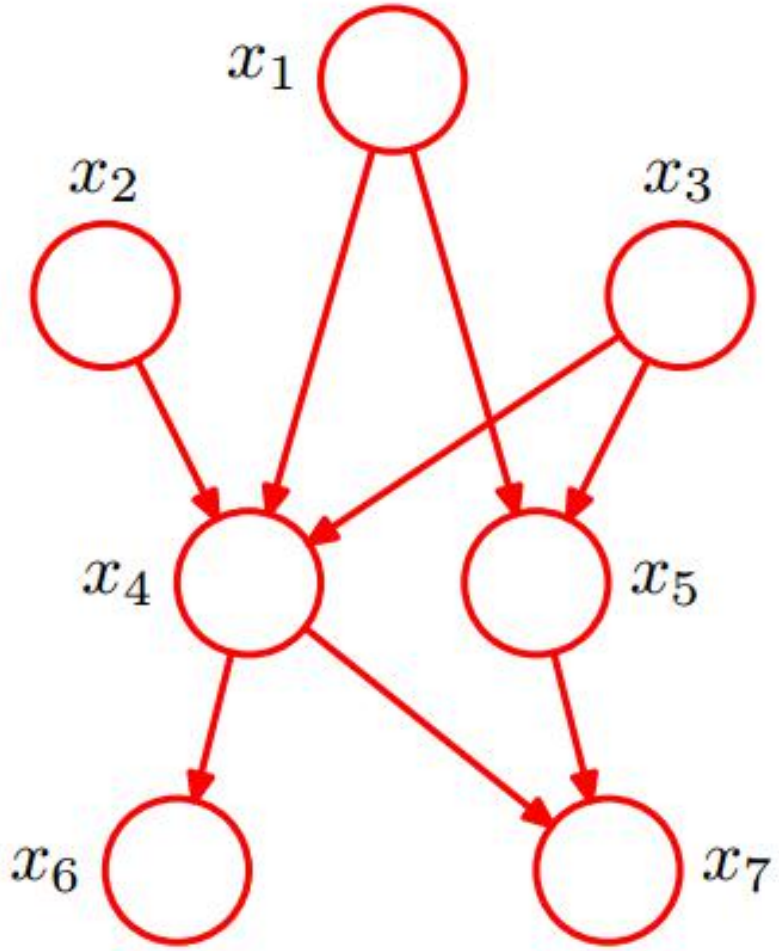
- 每一对节点之间都有边连接

$$P(x_1, x_2, \ldots, x_n) = P(x_n \mid x_1, x_2, \ldots, x_{n-1}) \ldots P(x_2 \mid x_1) P(x_1)$$

$$P(x_1, x_2, \ldots, x_n) = \prod_{i=2}^{n} P(x_i \mid x_1, x_2, \ldots, x_{i-1}) * P(x_1)$$



“正常”贝叶斯网络

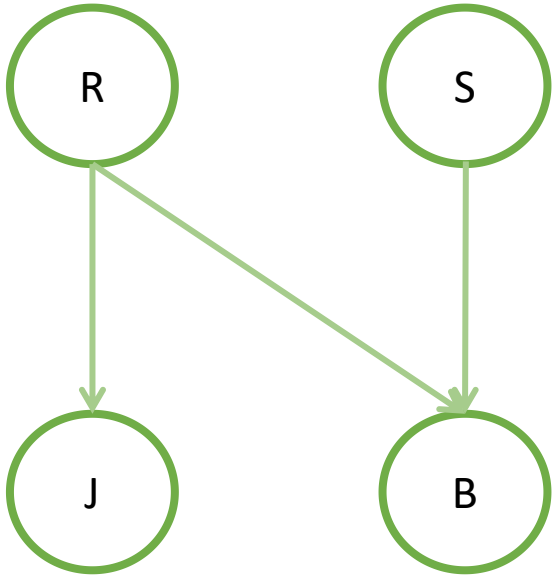


- x1,x2,x3独立
- x6和x7在给定条件下独立
- x1,x2,x3...x7的联合分布为

$$p(x_1,x_2,x_3,x_4,x_5,x_6,x_7)=p(x_1)p(x_2)p(x_3)p(x_4\mid x_1,x_2,x_3)p(x_5\mid x_1,x_3)p(x_6\mid x_4)p(x_7\mid x_4,x_5)$$

实际贝叶斯网络：判断是否下雨

- 有一天早晨，Bruce离开他的房子的时候发现他家花园中的草地是湿的，有两种可能，第一：昨天晚上下雨了，第二：他昨天晚上忘记关掉花园中的喷水器，接下来，他观察他的邻居Joe，发现他家花园中的草地也是湿的，因此，他推断，他家的草地湿了是因为昨天晚上下雨的缘故



R	0	1
P(R)	0.8	0.2

S	0	1
P(S)	0.9	0.1

R	0	1
P(J=0)	0.8	0
P(J=1)	0.2	1

R	S	p(B=0)	p(B=1)
0	0	1	0
0	1	0.1	0.9
1	0	0	1
1	1	0	1

实际贝叶斯网络：判断是否下雨

$$p(S=1|B=1,J=1)=\frac{p(S=1,B=1,J=1)}{p(B=1,J=1)}$$

$$=\frac{\sum_R p(B=1,J=1,R,S=1)}{\sum_{R,S} p(B=1,J=1,R,S)}$$

$$=\frac{\sum_R p(J=1|R)p(B=1|R,S=1)p(R)p(S=1)}{\sum_{R,S} p(J=1|R)p(B=1|R,S)p(R)p(S)}$$

$$=\frac{0.0344}{0.2144}=0.1604$$

$$p(R=1|B=1,J=1)=\frac{p(R=1,B=1,J=1)}{p(B=1,J=1)}$$

$$=\frac{\sum_S p(B=1,J=1,R=1,S)}{\sum_{R,S} p(B=1,J=1,R,S)}$$

$$=\frac{\sum_S p(J=1|R)p(B=1|R=1,S)p(R=1)p(S)}{\sum_{R,S} p(J=1|R)p(B=1|R,S)p(R)p(S)}$$

$$=\frac{0.2}{0.2144}=0.9328$$

所以判断出应该是下雨导致草地变湿

贝叶斯网络判定条件独立-01

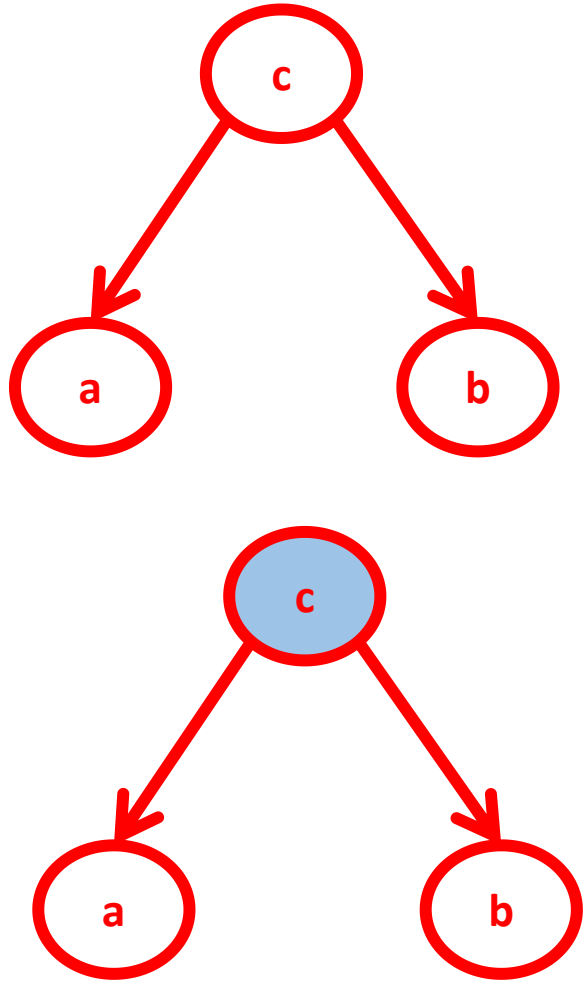
- 在C给定的条件下，a和b被阻断(blocked)是独立的
  - 条件独立：tail - to -tail

$$P(a,b,c) = P(c)P(b|c)P(a|c)$$

$$\Rightarrow P(a,b,c)/P(c) = P(b|c)P(a|c)$$

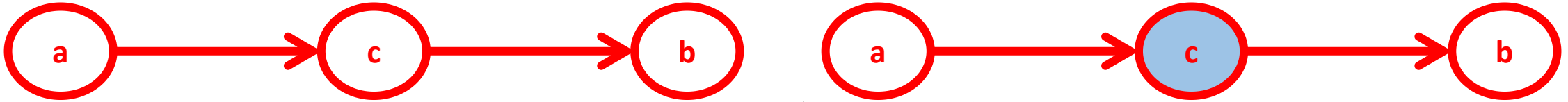
$$\because p(a,b|c) = p(a,b,c)/p(c)$$

$$\therefore p(a,b|c) = p(a|c)p(b|c)$$



贝叶斯网络判定条件独立-02

- 在C给定的条件下，a和b被阻断(blocked)是独立的
  - 条件独立：head- to -tail

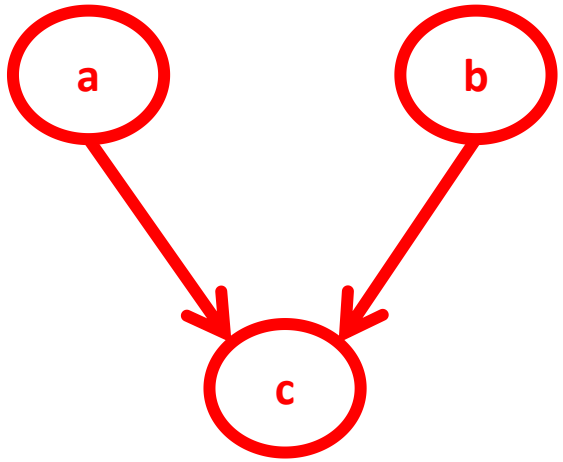


$$P(a,b,c) = P(a)P(c|a)P(b|c)$$

$$\begin{aligned} p(a,b|c) &= p(a,b,c)/\ p(c) \\ &= p(a) * p(c|a) * p(b|c) / p(c) \\ &= p(a,c) * p(b|c) / p(c) \\ &= p(a|c) * p(b|c) \end{aligned}$$

贝叶斯网络判定条件独立-03

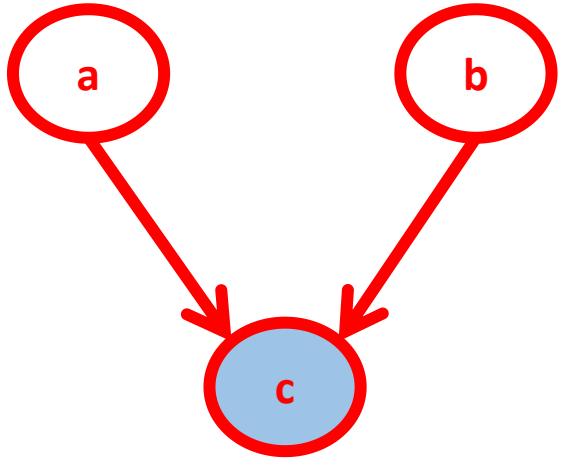
- 在C未知的情况下，a和b被阻断(blocked)，是独立的
  - 条件独立：head - to - head



$$P(a,b,c) = P(\mathrm{a})P(b)P(c \mid a,b)$$

$$\sum_{c} P(a,b,c) = \sum_{c} P(a) * P(b) * P(c \mid a,b)$$

$$\Rightarrow P(a,b) = P(a) * P(b)$$




北风网
IBEIFENG.COM