一叶收获

硬间隔
软间隔
彩图特征组合方法（3通道2个通道组合）

2019

隐写分析深度模型

选择通道

2019-11-20

目录

DIRECTORY

01 自适应隐写算法

02 SRM模型及其变体

03 卷积神经网络—选择通道

04 衍生-区域

自适应隐写算法

以WOW隐写算法为例：

使用三个方向滤波器计算方向残差 $R^{(k)} = K^{(k)} * X \qquad K^{(k)}, k = 1,2,3$

计算每一个像素的嵌入适合性 $\xi_{ij}^{(k)} = |R^{(k)}| * |R^{(k)} - R_{[ij]}^{(k)}|$

综合嵌入适合性计算嵌入代价 $\rho_{ij}^{(p)} = (\sum_{i=1}^{3} |\xi_{ij}^{(k)}|^p)^{-1/p}, p = -1$

使用STC编码最小化损失函数 $D(X,Y) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \rho_{i,j}(X, Y_{i,j}) |X_{i,j} - Y_{i,j}|$

相较于平滑区域，在图像纹理丰富区域嵌入信息

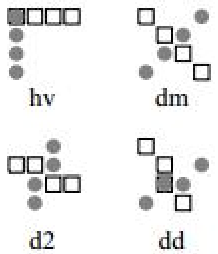
使用嵌入代价计算每个像素嵌入信息的概率——**选择通道**

2014_Adaptive Steganalysis Against WOW Embedding

SRM模型及变体

SRM

使用截断和量化后的残差计算共生矩阵 → 构成富模型输入到分类器



共生矩阵描述纹理特征，以垂直扫描方向为例：

(1,2)　(2,3)　(2,3)　(3,3)　(1,1)
(1,2)　(3,2)　(2,1)　(4,4)　(4,1)

1	2	1	3	4
2	3	1	2	4
3	3	2	1	1

	0	1	2	3	4	5	6	7
0	0	0	0	0	0	0	0	0
1	0	1	2	0	0	0	0	0
2	0	1	0	2	0	0	0	0
3	0	0	1	1	0	0	0	0
4	0	1	0	0	1	0	0	0
5	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0

tSRM

嵌入代价越小，嵌入信息的概率越大，使用嵌入代价选择像素计算共生矩阵

maxSRM

生成共生矩阵时，将+1替换为加上四个值中最大的嵌入变化概率

σmaxSRM

将像素改变转移到对残差失真的统计度量

上页参考文献：
tSRM 2014_Adaptive Steganalysis Against WOW Embedding
maxSRM 2014_Selection-Channel-Aware Rich Model for Steganalysis of Digital Images
σmaxSRM Improving Selection-Channel-Aware Steganalysis Features

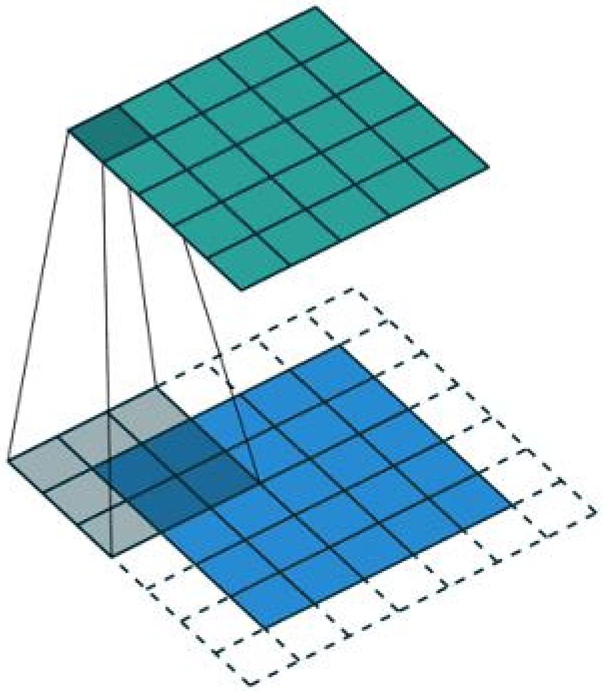
扫描方向 SRM 四维共生矩阵

In other words, instead of adding a 1 to the corresponding co-occurrence bin, we add the maximum of the embedding change probabilities taken across the four residuals. This way, those groups of four of pixels with
small probability of being changed will not affect the co-occurrence values much, while those where at least one pixel is likely to change will.

As pointed out in the introduction, there is a discrepancy in maxSRM in the sense that we accumulate the embedding change probabilities of pixels in the co-occurrence bins of residuals. Thus, we need to move away
from pixel change rates to some measure of the residual distortion. After all, if the features were formed from pixel values rather than residuals, the change rates are proportional to the expected value of the L1 (and L2) distortion.

卷积神经网络—选择通道

《Deep Learning Hierarchical Representations for Image Steganalysis》



像素值±1概率

$n_{ij} \in \{-1, 0, 1\}$ $\{\beta_{ij}, 1-2\beta_{ij}, \beta_{ij}\}$

像素值变化绝对值的期望

$E[|n_{ij}|] = 2\beta_{ij}$

残差失真

$$D = K * Y - K * X = K * (Y - X) = K * N = \sum_{r,c} k^{rc} n_{ij}^{rc} = (d_{ij})$$

残差失真L1范数期望的上界

$$E[|d_{ij}|] = E[|\sum_{r,c} k^{rc} n_{ij}^{rc}|] \le E[\sum_{r,c} |k^{rc}| \cdot |n_{ij}^{rc}|] = 2\sum_{r,c} |k^{rc}| \cdot \beta_{ij}^{rc} = \varphi(\beta_{ij}) \qquad \varphi(P) = P * |K|$$

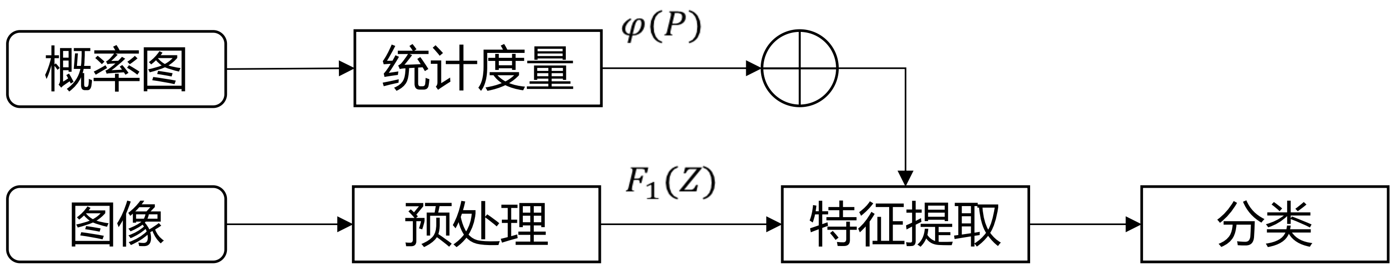
2017_Deep Learning Hierarchical Representations for Image Steganalysis

卷积神经网络—选择通道（续）

按元素相加方式并入网络

$$F_2(Z) = f_2((F_1(Z) + \varphi(P)) * W_2 + B_2)$$

$$F_2(Z) = (F_1(Z) * W_2 + B_2) + \varphi(P) * W_2$$



性能：低嵌入率时，检测准确率提升较大，随着嵌入率的增加，检测准确率提升程度逐渐减小

《Adaptive Steganalysis Based on Selection Region and Combined》

计算每个像素点的嵌入概率，为每个区域计算总的概率选择和最大的区域

2017_Adaptive Steganalysis Based on Selection Region and Combined

《机器学习●周志华》

第六章 支持向量机

目录 /CONTENTS

第六章 支持向量机

1.支持向量机的定义

2.间隔与支持向量

3.对偶问题

4.软间隔SVM

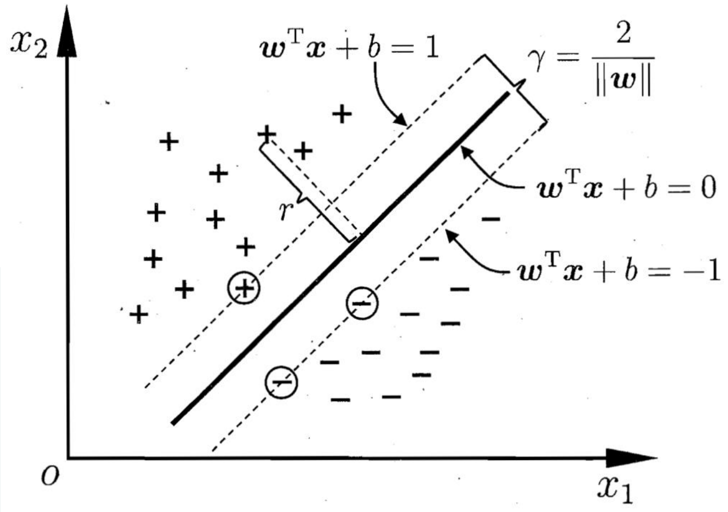
5.核函数

什么是支持向量机

- 支持向量机（Support Vector Machine，常简称为SVM）是一种监督式学习的方法，可广泛地应用于统计分类以及回归分析。支持向量机属于一般化线性分类器，这种分类器的特点是它们能够同时最小化经验误差与最大化几何边缘区，因此支持向量机也被称为最大边缘区分类器。

间隔与支持向量

- 首先考虑线性可分情况下的线性分类器，这是最原始的SVM，它最核心的思想就是找到最大分类间隔。
- 间隔：在样本空间里建立有一个超平面。在分开数据的超平面的两边建有两个互相平行的超平面，分隔超平面到两个平行超平面的距离。
- 支撑平面：是两个和最终求解规划问题得到的划分平面在高维特征空间中平行的平面。
- 支持向量：支持向量就是落在支撑平面上的数据点。



硬间隔SVM与求解

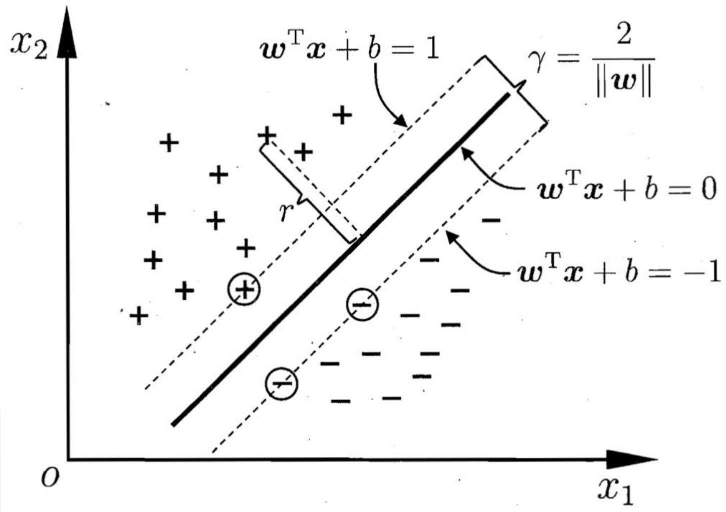
- 硬间隔SVM：当训练数据线性可分时，通过硬间隔最大化可以学习得到一个线性分类器，即硬间隔SVM。
- 最大间隔的求解：欲找到具有“最大间隔”的划分超平面，也就是要找到满足支撑平面的约束参数w和b，使得间距r最大。即

$$\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|}$$
$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m.$$

- 上述式子可以转换为：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geqslant 1, \quad i = 1, 2, \ldots, m.$$

这就是支持向量的基本型。



对偶问题

- 支持向量机基本型：$\min_{\boldsymbol{w},b} \ \frac{1}{2}\|\boldsymbol{w}\|^2$

  $\text{s.t.} \ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\geqslant 1, \quad i=1,2,\ldots,m.$
- 该式就是一个凸二次规划问题，能直接用现成的优化计算包求解，但我们可以有更高效的办法，对该式使用拉格朗日乘子法可得到其“对偶问题” $L(\boldsymbol{w},b,\boldsymbol{\alpha})=\frac{1}{2}\|\boldsymbol{w}\|^2+\sum_{i=1}^{m}\alpha_i\left(1-y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+b)\right)$
- 继而引入拉格朗日函数和对偶变量α，化为对单一因数对偶变量α的求解，将上式中的w和b消去后，得到对偶问题 $\max_{\boldsymbol{\alpha}} \ \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \quad \text{s.t.} \ \sum_{i=1}^{m}\alpha_i y_i=0 \quad \alpha_i\geqslant 0, \quad i=1,2,\ldots,m$
- 解出α后，求出w和b即可得到模型 $f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b \ =\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}+b\,.$
- 该式需要满足（KKT）条件，即要求 $\begin{cases}\alpha_i\geqslant 0\,; \\ y_i f(\boldsymbol{x}_i)-1\geqslant 0\,; \\ \alpha_i\left(y_i f(\boldsymbol{x}_i)-1\right)=0\,.\end{cases}$

对偶问题的求解

- 对偶问题： $\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j$ s.t. $\sum_{i=1}^{m} \alpha_i y_i = 0$ $\alpha_i \geqslant 0$, $i = 1, 2, \ldots, m$
- 这是一个二次规划问题，可使用通用的二次规划算法来求解；然后，该问题的规模正比于训练样本数，这会在实际任务中造成很大的开销，为了避开这个障碍，有一个著名的高效算法SMO。
- 使用SMO算法，对偶问题可以重写成：$\alpha_i y_i + \alpha_j y_j = c$ ，$\alpha_i \geqslant 0$ ，$\alpha_j \geqslant 0$ ，将该式代入对偶问题中，消去αj，则得到一个关于αi的单变量二次规划问题，仅有的约束是 $\alpha_i \geqslant 0$ 。不难发现，这样的二次规划问题具有闭式解，于是不必调用数值优化算法即可高校地计算出更新后的αi和αj。

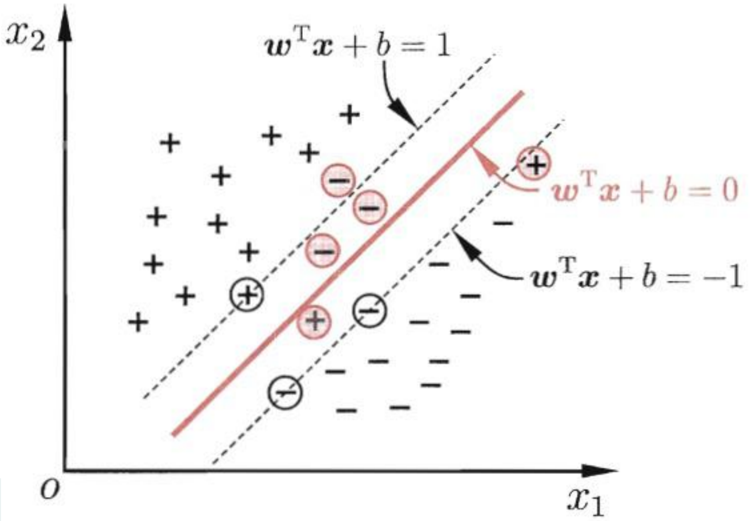
软间隔SVM与最大间隔

- 软间隔SVM：当训练数据不能线性可分但是可以近似线性可分时，通过软间隔(soft margin)最大化也可以学习到一个线性分类器，即软间隔SVM。
- 最大化软间隔要求：SVM对于训练集中的每个样本都引入一个松弛因子ξ，使得函数距离加上松弛因子后的值是大于等于1。
  $y^{(i)}\left(w^T x^{(i)}+b\right) \geq 1-\xi_i, i=1,2,\ldots,m, \xi_i \geq 0$
- 分析上述公式：松弛因子(ξ)越大，表示样本点离超平面越近。如果松弛因子大于1，那么表示允许该样本点分错。
- 最终目标函数转换成：

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m} \xi_i$$

$$s.t: y^{(i)}\left(w^T x^{(i)}+b\right) \geq 1-\xi_i, i=1,2,\ldots, m$$

$$\xi_i \geq 0, i=1,2,\ldots, m$$



软间隔模型优化

- 同线性可分SVM，构造软间隔最大化的约束问题对应的拉格朗日函数：
$$L(w,b,\xi,\beta,\mu)=\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{m}\xi_i+\sum_{i=1}^{m}\beta_i\left[1-\xi_i-y^{(i)}\left(w^T x^{(i)}+b\right)\right]-\sum_{i=1}^{m}\mu_i\xi_i,\beta_i\geq 0,\mu_i\geq 0$$
- 从而将我们的优化目标函数转换为： $\min\limits_{w,b,\xi}\max\limits_{\beta,\mu}L(w,b,\xi,\beta,\mu)$
- 优化目标同样满足KKT条件，所以使用拉格朗日对偶将优化问题转换为等价的对偶问题： $\max\limits_{\beta,\mu}\min\limits_{w,b,\xi}L(w,b,\xi,\beta,\mu)$
- 先求优化函数对于w、b、ξ的极小值，这个可以通过分别对优化函数L求w、b、ξ的偏导数得，从而可以得到w、b、ξ关于β和μ之间的关系：
$$\frac{\partial L}{\partial w}=0\Rightarrow w-\sum_{i=1}^{m}\beta_i y^{(i)}x^{(i)}=0\Rightarrow w=\sum_{i=1}^{m}\beta_i y^{(i)}x^{(i)}$$
$$\frac{\partial L}{\partial b}=0\Rightarrow -\sum_{i=1}^{m}\beta_i y^{(i)}=0\Rightarrow \sum_{i=1}^{m}\beta_i y^{(i)}=0$$
$$\frac{\partial L}{\partial \xi_i}=0\Rightarrow C-\beta_i-\mu_i=0$$
- 将w、b、ξ的值带入L函数中，就可以消去优化函数中的w、b、ξ，定义优化之后的函数为：
$$\ell(\beta)=\sum_{i=1}^{m}\beta_i-\frac{1}{2}\sum_{i=1,j=1}^{m}\beta_i\beta_j y^{(i)}y^{(j)}x^{(i)^T}x^{(j)}$$
$$s.t:\sum_{i=1}^{m}\beta_i y^{(i)}=0$$
$$C-\beta_i-\mu_i=0$$
$$\beta_i\geq 0,i=1,2,\ldots,m$$
$$\mu_i\geq 0,i=1,2,\ldots,m$$
- 最终优化后的目标函数和线性可分SVM模型基本一样，除了约束条件不同而已， 也就是说也可以使用SMO算法来求解。得到最终优化目标函数：
$$\min\frac{1}{2}\sum_{i=1,j=1}^{m}\beta_i\beta_j y^{(i)}y^{(j)}x^{(i)^T}x^{(j)}-\sum_{i=1}^{m}\beta_i$$
$$s.t:\sum_{i=1}^{m}\beta_i y^{(i)}=0$$
$$0\leq\beta_i\leq C,i=1,2,\ldots,m$$

核函数

- 当训练数据线性不可分时，这就需要使用核函数，将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。
- 核函数定义：令X为输入空间，$\kappa(\cdot,\cdot)$是定义在$\mathcal{X}\times\mathcal{X}$上的对称函数，则k是核函数当且仅当对于任意数据D = {x1,x2,...,xm}，“核矩阵”K总是半正定的：

$$\mathbf{K}=\begin{bmatrix}\kappa(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_j) & \cdots & \kappa(\boldsymbol{x}_1,\boldsymbol{x}_m)\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ \kappa(\boldsymbol{x}_i,\boldsymbol{x}_1) & \cdots & \kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) & \cdots & \kappa(\boldsymbol{x}_i,\boldsymbol{x}_m)\\ \vdots & \ddots & \vdots & \ddots & \vdots\\ \kappa(\boldsymbol{x}_m,\boldsymbol{x}_1) & \cdots & \kappa(\boldsymbol{x}_m,\boldsymbol{x}_j) & \cdots & \kappa(\boldsymbol{x}_m,\boldsymbol{x}_m)\end{bmatrix}$$

- 该定理表明，只要一个对称函数所对应的核矩阵半正定，它就能作为核函数使用。事实上，对于一个半正定核矩阵，总能找到一个与之对应的映射Φ。

常用核函数

- 核函数的选择至关重要，下表列出了几种常用的核函数

名称	表达式	参数
线性核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$	
多项式核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j)^d$	$d\geqslant 1$ 为多项式的次数
高斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$	$\sigma>0$ 为高斯核的带宽(width)
拉普拉斯核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}{\sigma}\right)$	$\sigma>0$
Sigmoid 核	$\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\tanh(\beta\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j+\theta)$	tanh 为双曲正切函数, $\beta>0$, $\theta<0$

其它核函数

- 核函数也可以通过函数的组合得到，如：
- 若k1和k2为核函数，则对于任意正数 γ_1、γ_2，其线性组合 $\gamma_1\kappa_1+\gamma_2\kappa_2$ 也是核函数。
- 若k1和k2为核函数，则核函数的直积 $\kappa_1\otimes\kappa_2(\boldsymbol{x},\boldsymbol{z})=\kappa_1(\boldsymbol{x},\boldsymbol{z})\kappa_2(\boldsymbol{x},\boldsymbol{z})$ 也是核函数。
- 若k1为核函数，则对于任意函数g(x)，$\kappa(\boldsymbol{x},\boldsymbol{z})=g(\boldsymbol{x})\kappa_1(\boldsymbol{x},\boldsymbol{z})g(\boldsymbol{z})$ 也是核函数。

机器学习 周志华 chap7 知识点总结

目录 /CONTENTS

Chap 7

贝叶斯分类

- 贝叶斯决策论
- 极大似然估计
- 朴素贝叶斯分类器
- 半朴素贝叶斯分类器
- 贝叶斯网
- EM算法

,

Part 1 贝叶斯决策论

- **贝叶斯决策论**

贝叶斯决策论是概率框架下实施决策的基本方法。基于后验概率P$(c_i|x)$可获得将一个真实标记为c_j的样本x误分类为c_i所产生的期望损失，即在样本x上的“条件风险”：

$$P(c_i|x) = \sum_{j=1}^{N} \lambda_{ij} P(c_j|x).$$

- **贝叶斯判定准则**：$h: X \mapsto Y$，最小化总体风险，只需在每个样本上选择使条件风险最小的类别标记。此时的分类器称为“贝叶斯最优分类器” 与之对应的总体风险$R(h^*)$ 称为“贝叶斯风险”.

$$h^*(\boldsymbol{x}) = \underset{c \in \mathcal{Y}}{\arg\min}\, R(c \mid \boldsymbol{x})$$

  - $1 - R(h^*)$反映了分类器所能达到的最好性能，即通过机器学习所能产生的模型精度的理论上限。
  - 欲使用贝叶斯判定准则来最小化决策风险，首先要获得后验概率 $P(c|x)$，往往难以直接获得。机器学习要实现的是基于有限训练样本集尽可能准确地估计出后验概率。

- **主要策略**
  - 判别式模型：给定x，可通过直接建模$P(c|x)$来预测c；
  - 生成式模型：先对联合概率分布$P(x,c)$建模，然后再由此获得$P(c|x)$。

Part 2 极大似然估计

- 令D_c表示训练集D中第c类样本组成的集合，假设这些样本是独立同分布的，则参数θ_c对于数据集D_c的似然是：

$$\mathrm{P}(D_c|\theta_c)=\prod_{x\in D_c}P(x|\theta_c).$$

- **极大似然估计**

极大似然估计是根据数据采样估计概率分布参数的经典方法。它试图在参数θ_c所有可能的取值中，找到一个能使数据出现在数据集中的“可能性”最大的值。

➢ 为了避免下溢，通常使用对数似然：

$$LL(\theta_c)=\log\mathrm{P}(D_c|\theta_c)$$

- 需要注意的是，这种参数化的方法虽然能使类条件概率估计变得相对简单，但估计结果的准确性严重依赖于所假设的概率分布形式是否符合潜在的真实数据分布。

Part 3 朴素贝叶斯分类器

- **解决的问题**

基于贝叶斯公式来估计后验概率$P(c|x)$的主要困难在于类条件概率$P(x|c)$是所有属性的联合概率，难以从有限的训练样本直接估计而得。

- **假设**

“朴素贝叶斯分类器” 采用了“属性条件独立性假设” ，对已知类别，假设所有属性相对独立。

- **训练过程**

基于训练集D来估计类先验概率$P(c)$，并为每个属性估计条件概率$P(x_i|c)$。

- **拉普拉斯修正**

拉普拉斯修正避免了因训练样本不充分而导致概率估计值为零的问题，而且在训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，使得估值渐趋向于实际概率值。

- **实际应用**
  - 对预测速度要求较高：将涉及的所有概率估值事先计算好存储起来，预测时“查表”即可判别；
  - 任务数据更替频繁：采用“懒惰学习” 方法，先不进行任何训练，待接收到预测请求时再根据当前数据集进行概率估值；
  - 数据不断增加：在现有估值基础上，仅对新增样本的属性值所涉及的概率估值进行计数修正即可实现“增量学习”。

Part 4 半朴素贝叶斯分类器

- **解决的问题**

为降低贝叶斯公式中估计后验概率$P(c|x)$的困难，朴素贝叶斯分类器采用了属性条件独立性假设，但现实任务中此假设也很难成立，于是人们尝试对属性条件独立性假设进行一定程度的放松，由此产生了一类为“半朴素贝叶斯分类器” 的学习方法。

- **方法策略**

采用“独依赖估计” 策略：假设每个属性在类别之外最多仅依赖于一个其他属性，问题转化为如何确定每个属性的父属性（即被依赖的属性）。

- **确定父属性的方法**
  - **SPODE方法**：假设所有属性都依赖于同一个属性，称为“超父” ，然后通过交叉验证等模型选择的方法来确定超父属性。
  - **TAN**：在最大带权生成树算法的基础上，将属性间依赖关系约简为树形结构。TAN实际仅保留了强相关属性之间的依赖性。
  - **AODE**：AODE是一种基于集成学习机制、更为强大的独依赖分类器，AODE尝试将每个属性作为超父来构建SPODE，然后将那些具有足够训练数据支撑的SPODE集成起来作为最终结果。
    - ✓ **优点**：AODE无需模型选择，既能通过预计算节省预测时间，也能采取懒惰学习方式在预测时再进行计数，并且易于实现增量学习。
  - **AkDE**：AODE的扩展：通过来考虑属性间的高阶依赖进一步提升泛化性能。但是只有在训练数据非常充分的情况下泛化性能有可能提升。

Part 5 贝叶斯网

- **描述**

贝叶斯网又称“信念网”，它借助有向无环图来刻画属性之间的依赖关系，并使用条件概率表来描述属性的联合概率分布。

- **结构**

贝叶斯网结构有效地表达了属性间的条件独立性。给定父节点集，贝叶斯网假设每个属性与它的非后裔属性独立。

➢ **贝叶斯网中三个变量之间的典型依赖关系：**同父结构、V型结构、顺序结构。

- 在同父结构中，给定父结点x_1的取值，则x_3与x_4条件独立；
- V型结构亦称“冲撞结构”，给定子结点x_4的取值，x_1与x_2不必独立；
- 在顺序结构中，给定x的值，则y与z条件独立。

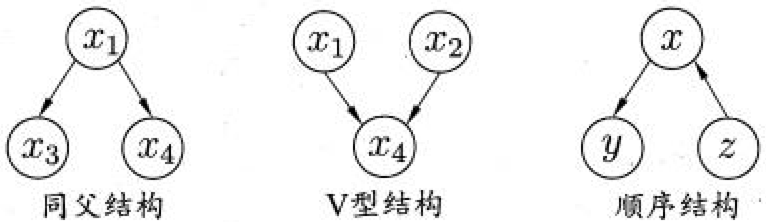
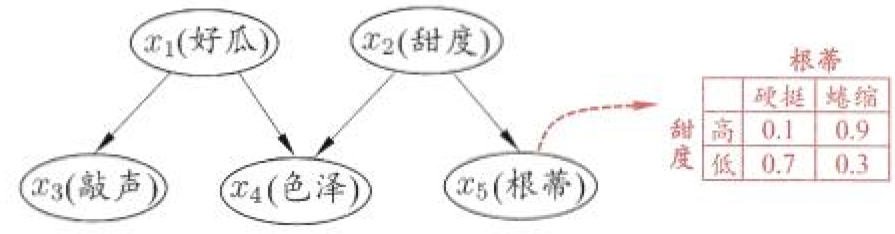


图 7.3 贝叶斯网中三个变量之间的典型依赖关系

➢ **“边际独立性”**：若x_4的取值完全未知，则V型结构下x_1与x_2是相互独立的。

➢ **“道德图”**：为了分析有向图中变量间的条件独立性，使用有向分离将有向图转变为一个无向图，也叫道德图。



		硬挺	蜷缩
甜度	高	0.1	0.9
	低	0.7	0.3

图 7.2 西瓜问题的一种贝叶斯网结构以及属性“根蒂”的条件概率表

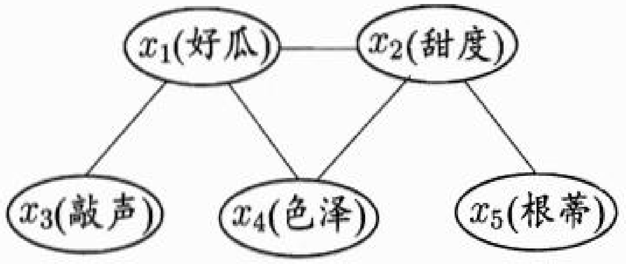


图 7.4 图 7.2 对应的道德图

- **推断**

通过已知变量观测值来推测待查询变量的过程称为“推断”，已知变量观测值称为“证据”.

➢ **吉布斯采样**

现实任务中，贝叶斯网的**近似推断（求后验概率）**常使用吉布斯采样来完成，这是一种随机采样方法。

- 吉布斯采样是在贝叶斯网所有变量的联合状态空间与证据一致的子空间进行“随机漫步”，每一步仅依赖于前一步的状态，这是一个“马尔可夫链”。

如图 7.5 所示，吉布斯采样算法先随机产生一个与证据 $\mathbf{E}=\mathbf{e}$ 一致的样本 $\mathbf{q}^0$ 作为初始点，然后每步从当前样本出发产生下一个样本. 具体来说，在第 t 次采样中，算法先假设 $\mathbf{q}^t=\mathbf{q}^{t-1}$，然后对非证据变量逐个进行采样改变其取值，采样概率根据贝叶斯网 B 和其他变量的当前取值(即 $\mathbf{Z}=\mathbf{z}$)计算获得. 假定经过 T 次采样得到的与 $\mathbf{q}$ 一致的样本共有 n_q 个，则可近似估算出后验概率

$$P(\mathbf{Q}=\mathbf{q} \mid \mathbf{E}=\mathbf{e}) \simeq \frac{n_q}{T} . \tag{7.33}$$

- 在T很大时，吉布斯采样相当于根据$P(Q|E=e)$采样，从而保证了后验概率收敛于$P(Q=q|E=e)$.
- 缺点：收敛速度较慢；若贝叶斯网中存在极端概率“0”或“1”，则不能保证马马尔可夫链存在平稳分布，此时吉布斯采样会给出错误的估计结果。

Part 6 EM算法

- **隐变量**

  属性中存在“未观测”的变量，称为“隐变量”。

- **EM算法**

  EM算法常用于估计参数隐变量，是一种迭代式的方法：若模型参数已知，则可根据训练数据推断出最优隐变量的值；反之，若隐变量的值已知，则可方便地对模型参数做极大似然估计。

- **算法步骤**
  - 期望（E）步：利用当前估计的参数值来计算对数似然的期望值；
  - 最大化（M）步，寻找能使E步产生的似然期望最大化的参数值。然后得到的参数值重新被用于E步，直至收敛到局部最优解。

- **隐变量估计问题的其他解决办法：**梯度下降等优化算法
  - 缺点：求和项数将随着隐变量的数目以指数级上升，计算复杂。