Chapter 7.

1. 类条件概率 (likelihood)，只考虑1个 attribute.

设前3个属性为 x_1, x_2, x_3，c 为 class.（是否好瓜）

$$\begin{cases} P(x_1|c) \approx \dfrac{|D_{x_1,c}|}{|D_c|} \\ \Rightarrow\ p(x_1=\text{乌黑, 浅白, 青绿}|c=\text{是}) = \frac{4}{8}, \frac{1}{8}, \frac{3}{8} \\ p(x_1= \text{———} \quad |\ c=\text{否}) = \frac{2}{9}, \frac{4}{9}, \frac{3}{9}. \end{cases}$$

↑ 这个是 naive Bayesian Classifier 思想.

Maximal Likelihood： assume that $p_1 = P(x_1=\text{乌黑}|c=\text{是})$，$p_2$ —，p_3 —

(likelihood)

$$L(\theta_c) = P(D_c|\theta_c) = \prod_{x\in D_c} P(x|\theta_c) = p_1^4 \cdot p_2 \cdot p_3^3$$

$$LL(\theta_c=\text{是}) = 4\log p_1 + \log p_2 + 3\log p_3 = 4\log p_1 + \log p_2 + 3\log(1-p_1-p_2)$$

$$\frac{\partial LL(\theta_c=\text{是})}{\partial p_1} = \frac{\partial LL(\theta_c=\text{是})}{\partial p_2} = 0$$

$$\Rightarrow \hat{p}_2 = \frac{1}{8},\ \hat{p}_1 = \frac{4}{8},\ \hat{p}_3 = \frac{3}{8}.$$

同理可得 $\theta_c=$否的情况 $\hat{q}_1 = \frac{2}{9}$，$\hat{q}_2 = \frac{4}{9}$，$\hat{q}_3 = \frac{3}{9}$.

2、Bayesian optimal classfier： $h_1(x) = \arg\max_c P(c|x)$.

naive Bayesian classifier： $h_2(x) = \arg\max_c p(x|c)\cdot p(c)$

$$= \arg\max_c \left(\prod_{i=1}^{d} p(x_i|c)\right) p(c) \quad \text{(attribute independent assumption)}$$

如果不满足 $x_1, \cdots x_d$ under condition d 独立.

假设 $d=2$，$c = A$ or A^c. $x_1 = K$，$x_2 = T$，$d=2$.

不妨设 $h_1(x) = A$，$\Leftrightarrow P(c=A|x) > P(c=A^c|x)$.

若 $h_2(x) = A$，$\Leftrightarrow p(x_1=K, x_2=T|c=A)\cdot P(c=A) > p(x_1=K, x_2=T|c=A^c)\cdot P(c=A^c)$

and $p(x_1=K|c)\cdot p(x_2=T|c) \neq p(x_1=K, x_2=T|c)$

假设 $x_1 = x_2$, $K=T$, 则 $p(x_1=K|c) \triangleq p_1$,

有 $p_1^2 \neq p_1$, 即 $p_1 \neq 0$ or 1 时, 不满足 conditional independence.

假设 $p(c=A) = p(c=A^c) = 1/2$、

即 $h_2(x)=A \iff p(x_1=k|c=A) > p(x_1=k|c=A^c) \iff p(c=A|x_1=k) > p(c=A^c|x_1=K)$.

$h_1(x) = A \iff p(c=A|x_1=k) > 1-p(c=A|x_1=k)$

此时需要 $\dfrac{p(c=A|x_1=k)\cdot p(x_1=k)}{p(c=A)} \neq 0$, or 1,

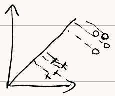
即 $p(c=A|x_1=k)\cdot p(x_1=k) \neq 0$ or $\frac{1}{2}$.

此时 $h_1(x) = h_2(x)$, optimal classifier = naive Bayesian classifier.

4. $h_{nb}(x) = \underset{c\in Y}{\arg\max}\ p(c)\cdot \prod_{i=1}^{d} p(x_i|c)$ (naive Bayesian classifier)

$= \underset{c\in Y}{\arg\max}\ \log p(c) + \sum_{i=1}^{d} \log p(x_i|c)$

5、线性判别分析 (LDA) 见 P60



$\vec{x}_0 \sim Normal(\mu_0, \Sigma_0)$,

$\vec{x}_1 \sim Normal(\mu_1, \Sigma_1)$. 2类数据.

$w = S_w^{-1}(\mu_0-\mu_1) = (\Sigma_0+\Sigma_1)^{-1}(\mu_0-\mu_1)$

划分超平面 (separating superplane) 为

$f(x) = w^T x - \frac{1}{2} w^T(\mu_0+\mu_1) = x^T\Sigma^{-1}(\mu_0-\mu_1) - \frac{1}{2}(\mu_0^T+\mu_1^T)\Sigma^{-1}(\mu_0-\mu_1)$

if $f(x)<0$, $x\in 1$, if $f(x)>0$, $x\in 0$

看最优贝叶斯 (optimal Bayesian classifier)

$h(x) = \underset{c\in\{0,1\}}{\arg\max}\ p(c|x)$

$= \underset{c\in\{0,1\}}{\arg\max}\ p(c)\cdot p(x|c)$

$x_0|c \sim Normal(\mu_0, \Sigma_0)$, $x_1|c \sim Normal(\mu_1, \Sigma_1)$

$$h(x) = \arg\max_{c\in\{0,1\}} \; p(c)\cdot \frac{1}{2\pi\cdot|\Sigma|} \exp\{-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$$

我们写成上述 $f(x)$ 形式. 用 $g(x)>0$ or $g(x)<0$ 进行判断

$$g(x) \triangleq \log p(0)/p(1) + \log\frac{|\Sigma_1|}{|\Sigma_0|} - \tfrac{1}{2}\left((x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1)\right)$$

$$= \log\frac{p(0)}{p(1)} - \tfrac{1}{2}\left(x^T\Sigma^{-1}(\mu_1-\mu_0) + \mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1\right)$$

(since $\Sigma_0 = \Sigma_1$)

$$= \log\frac{p(0)}{p(1)} - \tfrac{1}{2}\left(x^T\Sigma^{-1}(\mu_1-\mu_0) + (\mu_0^T+\mu_1^T)\Sigma^{-1}(\mu_0+\mu_1)\right)$$

assume that $p(0)=p(1)$,

we have $f(x) = -g(x)$, 交换 0,1 类即相同. □

7.10. $LL(\Theta|X,Z) = \log P(X,Z|\Theta)$

E: $Q(\Theta|\Theta^{(t)}) = E_{Z|X,\Theta^t}(LL(\Theta|X,Z))$

M: $\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(t)})$

求 $\arg\max_{\Theta} \ln\sum_Z P(X,Z|\Theta)$

考虑如下图情况,

traing data 为 $(x_1^{(i)}, x_2^{(i)})_{i=1}^m$

$$LL(\Theta) = \ln\sum_Z P(X,Z|\Theta)$$

$$= \ln\sum_Z P(Z|x_1,\Theta)\cdot P(x_2|Z,\Theta)$$

$$= \ln\sum_{i=1}^m\prod_{k=1}^d p(z_k|x_1^{(i)})\cdot p(x_2^{(i)}|z_k,\Theta)$$ we denote $p(k,i) = p(z_k|x_1^{(i)})$

求 $\arg\max_{p,q} \ln\sum_{i=1}^m\prod_{k=1}^d p(k,i)\cdot q(k,i)$ $\quad q(k,i) = p(x_2^{(i)}|z_k,\Theta)$

Laplacian function: $\ln\sum_{i=1}^m\prod_{k=1}^d p(k,i)\cdot q(k,i) + (\sum_{k=1}^d p(k,i) - 1)\lambda$

因为上述方程中对 k 的取值是无关的,

即可令 $\prod_{k=1}^d p(k,i)\cdot q(k,i) = h(i)$, 对结果无影响,

即 $\arg\max_h \ln\sum_{i=1}^m h(i)$, 与 Z 无关, 只与 $x_1^{(i)}, x_2^{(i)}$ 有关

同理可推广到任意 belief network 上

7.7. prior probability 有 $p(c, x_i)$

要求 $|D_{x_i}| \geq 30$, D_{x_i} 指第 i 个属性为 x_i 的集合

最好情况 30个，全为一种值，即 $x_1 = x_2 = \cdots = x_{30}$ （或者60个，所有 $c, x_1, \ldots, x_d$ 均可以有2种取值）

最坏情况，我们先考虑 $c=0$ 的情况，

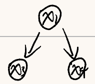
当 $p(x_1=0)$，$p(x_1=1)$ 各有30个样本时，

假设均为 $x_2=0$，则还需要30个 $p(x_2=1)$.

以此类推，需要 $30(d+1)$ 个样本，

考虑 $c=1$ 情况，共需要 $60(d+1)$ 个样本.

7.8. 同父结构：

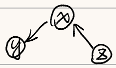


若 x_1 未知，

$$p(x_3, x_4) = \sum_{x_1} p(x_3, x_4, x_1)$$
$$= \sum_{x_1} p(x_1) \cdot p(x_3|x_1) \cdot p(x_4|x_1)$$
$$\neq \sum_{x_1} p(x_1) \cdot p(x_3|x_1) \cdot p(x_4)$$
$$= p(x_3) \cdot p(x_4).$$

故 $x_3 \perp\!\!\!\perp x_4$ 不成立.

顺序结构:



$y \perp z | x$:

$$p(y, z|x) = \frac{p(x, y, z)}{p(x)}$$
$$= \frac{p(y|x) \cdot p(x|z) \cdot p(z)}{p(x)}$$
$$= p(y|x) \cdot p(z|x)$$

$y \perp\!\!\!\perp z$ 不成立：

$$p(y, z) = \sum_x p(y, z, x)$$
$$= \sum_x p(z) \cdot p(x|z) \cdot p(y|x)$$
$$= p(z) \sum_x p(x|z) \cdot p(y|x) \neq p(z) \cdot \sum_x p(x|z) \cdot p(y) = p(z) \cdot p(y)$$

□

7.9. BIC准则： $BIC(B|D) = \frac{\log m}{2}|B| - LL(B|D)$

书上所说，使用贪心策略，每次改变一条边，

找到一个局部最优解，具体算法就不学习了，这涉及到树结构。