独立二值变量贝叶斯决策：

考虑 d 维特征，2分类情况，且

$p_i = p(x_i=1|w_1)$， $q_i = p(x_i=1|w_2)$.

则 $p(x|w_1) = \prod_{i=1}^{d} p(x_i|w_1) = \prod_{i=1}^{d} p_i^{x_i}(1-p_i)^{1-x_i}$

$p(x|w_2) = \prod_{i=1}^{d} p(x_i|w_2) = \prod_{i=1}^{d} q_i^{x_i}(1-q_i)^{1-x_i}$

似然比：$\frac{p(x|w_1)}{p(x|w_2)} = \prod_{i=1}^{d}\left(\frac{p_i}{q_i}\right)^{x_i}\left(\frac{1-p_i}{1-q_i}\right)^{1-x_i}$

判别函数：$g(x) = \log\frac{p(x|w_1)p(w_1)}{p(x|w_2)p(w_2)}$

$= \sum_{i=1}^{d}\left[x_i \ln\frac{p_i}{q_i} + (1-x_i)\ln\frac{1-p_i}{1-q_i}\right] + \ln\frac{p(w_1)}{p(w_2)}$

$\triangleq \sum_{i=1}^{d} w_i x_i + w_0$ 线性判别.

其中 $w_i = \ln\frac{p_i(1-q_i)}{q_i(1-p_i)}$ $w_0 = \sum_{i=1}^{d}\ln\frac{1-p_i}{1-q_i} + \ln\frac{p(w_1)}{p(w_2)}$

参数估计方法：

1. MLE：假设参数为确定值，目标：使似然值最大.
2. Bayesian learning：假设参数为随机变量，估计其分布.

MLE：

基本原理：假设概率密度：$p(x|w_i,\theta_i)$ θ_i 待估.

似然函数：$p(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$.

希望似然最大：$\arg\max_{\theta} p(D|\theta) \Leftrightarrow \nabla_\theta\, p(D|\theta) = 0$

一般计算对数似然：$l(\theta) = \ln p(D|\theta) = \sum_{k=1}^{n}\ln p(x_k|\theta)$

MAP (Maximum a posterior) Estimator：

$\max_{\theta}\; l(\theta)\,p(\theta)$.

若 $p(\theta)$ 是均匀分布，MAPE $\Leftrightarrow$ MLE.

Gaussian 的 MLE 估计：

$p(x|\mu,\Sigma) \sim N(\mu,\Sigma) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}$.

Case Ⅰ. μ未知. Σ已知.

$$\ln(p(x_k|\mu)) = -\frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(x_k-\mu)'\Sigma^{-1}(x_k-\mu)$$

$$\Rightarrow l(\theta) = \sum_{k=1}^{n}\ln(p_k|\mu)$$

$$\nabla_\theta l(\theta) = 0 \iff \sum_{k=1}^{n}\Sigma^{-1}(x_k-\hat{\mu}) = 0 \Rightarrow \hat{\mu}_{MLE} = \frac{1}{n}\sum_{k=1}^{n}x_k$$

Case Ⅱ. μ未知. Σ未知.

似然函数: $p(D|\theta) = (2\pi)^{-\frac{nd}{2}}|\Sigma|^{-\frac{n}{2}}\exp\{-\frac{1}{2}tr(\Sigma^{-1}(V+n(\bar{x}-\mu)'(\bar{x}-\mu)))\}$

其中 $\bar{x} = \frac{1}{n}\sum_{k=1}^{n}x_k$. $V = \sum_{k=1}^{n}(x_k-\bar{x})(x_k-\bar{x})'$

则要始似然函数 $p(D|\theta)$ 最大. $\hat{\mu}_{MLE} = \bar{x} = \frac{1}{n}\sum_{k=1}^{n}x_k$

把 $\hat{\mu}_{MLE} = \bar{x}$ 代入后, 得到:

$$p(D|\theta) \propto |\Sigma|^{-\frac{n}{2}}\exp\{-\frac{1}{2}tr(\Sigma^{-1}V)\}.$$

对于 $\Sigma^{-\frac{1}{2}}V\Sigma^{-\frac{1}{2}}$, 存在正交分解. 即

$$\Sigma^{-\frac{1}{2}}V\Sigma^{-\frac{1}{2}} = U^T\Lambda U, \quad U^TU = I_n. \quad \Lambda = diag\{\lambda_1,\cdots,\lambda_d\}.$$

则 $|\Sigma|^{-1} = |V|^{-1}\prod_{i=1}^{d}\lambda_i$

$tr(\Sigma^{-1}V) = tr(\Lambda) = \sum_{i=1}^{d}\lambda_i$

故 $p(D|\theta) = |V|^{-\frac{n}{2}}\prod_{i=1}^{d}\lambda_i^{\frac{n}{2}}\exp\{-\frac{1}{2}\lambda_i\}$.

要使似然函数最大, 则 $\hat{\lambda}_i = n$. 进而 $\hat{\Sigma} = \frac{V}{n} = \frac{1}{n}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})'$

注意到 $\hat{\Sigma}$ 是有偏估计. 而 $\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\hat{\mu})(x_k-\hat{\mu})'$ 是无偏的.

贝叶斯参数估计:

$$p(w_i|x,D) = \frac{P(x|w_i,D_i)P(w_i)}{\sum_{j=1}^{c}P(x|w_j,D_j)P(w_j)}$$

可忽略.

, 假设先验概率 $p(w_j)$ 已知. 且每个类别样本 D_i 已包含了该类别概率密度所有信息.

则现需求 $p(x|w_i,D_i)$:

* 假设密度函数参数形式 $p(x|\theta)$ 已知

*假设关于参数的先验分布 $p(\theta)$ 已知.

则: $p(x|D)=\int p(x,\theta|D)\,d\theta=\int p(x|\theta)p(\theta|D)\,d\theta$

若 $p(\theta|D)$ 在 $\hat{\theta}$ 处有显著尖峰. $p(x|D)\approx p(x|\hat{\theta})$

Gauss 密度贝叶斯估计:

Case I: 1维、σ^2 已知, μ 未知.

$p(x|\mu)\sim N(\mu,\sigma^2)$

$p(\mu)\sim N(\mu_0,\sigma_0^2)$. μ_0 可以看成当前对 μ 的最好估计. σ_0^2 为对当前估计不确定性.

则 $p(\mu|D)=\dfrac{p(D|\mu)p(\mu)}{\int p(D|\mu)p(\mu)\,d\mu}$

$$=\alpha\prod_{k=1}^{n}p(x_k|\mu)p(\mu).\quad \alpha \text{ 为归一化因子.}$$

$$=\alpha\left\{\prod_{k=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x_k-\mu}{\sigma}\right)^2\right]\right\}\frac{1}{\sqrt{2\pi}\sigma_0}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right]$$

$$=\alpha'\exp\left[-\frac{1}{2}\left(\sum_{k=1}^{n}\left(\frac{x_k-\mu}{\sigma}\right)^2+\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$=\alpha''\exp\left\{-\frac{1}{2}\left[\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\sum_{k=1}^{n}\frac{x_k}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}\right)\mu\right]\right\}.$$

注意到 exp 中为 μ 的二次型, $p(\mu|D)$ 仍为一个正态.

则 $$\mu_n=\frac{\sum_{k=1}^{n}\frac{x_k}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}=\frac{n\hat{\mu}\sigma_0^2+\mu_0\sigma^2}{n\sigma_0^2+\sigma^2}=\frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2}\hat{\mu}+\frac{\sigma^2}{n\sigma_0^2+\sigma^2}\mu_0$$

$$\sigma_n^2=\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}=\frac{\sigma_0^2\sigma^2}{n\sigma_0^2+\sigma^2}$$

当 $n\to\infty$ $\mu_n\longrightarrow\hat{\mu}$, $\sigma_n^2\longrightarrow 0$.

$$p(x|D)=\int p(x|\mu)p(\mu|D)\,d\mu$$

$$=\int\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\frac{1}{\sqrt{2\pi}\sigma_n}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_n}{\sigma_n}\right)^2\right]d\mu$$

$$=\frac{1}{2\pi\sigma\sigma_n}\exp\left[-\frac{1}{2}\frac{(x-\mu_n)^2}{\sigma^2+\sigma_n^2}\right]f(\sigma,\sigma_n)$$

其中 $f(\sigma,\sigma_n)=\int\exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_n^2}{\sigma^2\sigma_n^2}\left(\mu-\frac{\sigma_n^2x+\sigma^2\mu_n}{\sigma^2+\sigma_n^2}\right)^2\right]d\mu$. 为一个常数.

故 $f(x|D)$ 仍为正态. $f(x|D)\sim N(\mu_n,\sigma^2+\sigma_n^2)$

注意到. σ^2 变为 $\sigma^2+\sigma_n^2$, 由 μ_n 估计不确定性导致

Case II 多维, μ 未知, Σ 已知.

$p(x|\mu)\sim N(\mu,\Sigma)$. $p(\mu)\sim N(\mu_0,\Sigma_0)$

$$p(\mu|D)=\alpha\prod_{k=1}^{n}p(x_k|\mu)\cdot p(\mu)$$

$$=\alpha'\exp\{-\frac{1}{2}(\sum_{k=1}^{n}(x_k-\mu)'\Sigma^{-1}(x_k-\mu)+(\mu-\mu_0)'\Sigma_0^{-1}(\mu-\mu_0))\}$$

$$=\alpha''\exp\{-\frac{1}{2}[\mu'(n\Sigma^{-1}+\Sigma_0^{-1})\mu-2\mu^t(\Sigma^{-1}\sum_{k=1}^{n}x_k+\Sigma_0^{-1}\mu_0)]\}.$$

$\Rightarrow p(\mu|D)\sim N(\mu_n,\Sigma_n)$,

其中：$\Sigma_n^{-1}=n\Sigma^{-1}+\Sigma_0^{-1}\Rightarrow\Sigma_n=\Sigma_0(\Sigma_0+\frac{1}{n}\Sigma)^{-1}\frac{1}{n}\Sigma=\frac{1}{n}\Sigma(\Sigma_0+\frac{1}{n}\Sigma)^{-1}\Sigma_0$

$$\mu_n=\Sigma_n(n\Sigma^{-1}\hat{\mu}+\Sigma_0^{-1}\mu_0)=\Sigma_0(\Sigma_0+\frac{1}{n}\Sigma)^{-1}\hat{\mu}+\frac{1}{n}\Sigma(\Sigma_0+\frac{1}{n}\Sigma)^{-1}\mu_0$$

$$p(x|D)=\int p(x|\mu)p(\mu|D)d\mu\sim N(\mu_n,\Sigma+\Sigma_n)$$

一般情况下的贝叶斯估计.

基本条件： ① 知道 $p(x|\theta)$ 的参数形式

② 关于参数 θ 的先验分布 $p(\theta)$ 已知.

③ n 个样本的数据集独立，且一个类只依赖其类样本.

步骤：

① 求 $p(\theta|D)=\dfrac{p(D|\theta)p(\theta)}{\int_\theta p(D|\theta)p(\theta)d\theta}$

② 求 $p(x|D)=\int_\theta p(x|\theta)p(\theta|D)d\theta$，一般取尖峰 $\hat{\theta}$ 来近似 $p(x|D)=p(x|\hat{\theta})$

③ 模型使用.

缺点：复杂模型的参数后验 $p(\theta|D)$ 和数据后验 $p(x|D)$ 很难算

可采用 MCMC + LA 近似.

递归 Bayesian Learning：

$D^n=\{x_1,\cdots,x_n\}$，样本有顺序的到达.

$$p(D^n|\theta)=p(x^n|\theta)\cdot p(D^{n-1}|\theta)$$

$$p(D^n,\theta)=p(x^n|\theta)p(D^{n-1},\theta)=p(x^n|\theta)p(\theta|D^{n-1})\cdot p(D^{n-1}).$$

$$\Rightarrow p(\theta|D) = \frac{p(D|\theta)\,p(\theta)}{\int p(D|\theta)\,p(\theta)\,d\theta} = \frac{p(D,\theta)}{\int_\theta p(D,\theta)\,d\theta} = \frac{p(x_n|\theta)\,p(\theta|D^{n-1})}{\int p(x_n|\theta)\,p(\theta|D^{n-1})\,d\theta}$$

Example.

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} \frac{1}{\theta} & 0<x<\theta \\ 0 & \text{otherwise.} \end{cases}$$

$$p(\theta|D^0) = p(\theta) = U(0,10)$$

$$D = \{4, 7, 2, 8\}.$$

則

$$p(\theta|D^1) \propto p(x|\theta)\cdot p(\theta|D^0) = \begin{cases} \frac{1}{\theta} & 4<\theta\le 10 \\ 0 & \text{otherwise.} \end{cases}$$

$$p(\theta|D^2) \propto p(x|\theta)\cdot p(\theta|D^1) = \begin{cases} \frac{1}{\theta^2} & 7<\theta\le 10 \\ 0 & \text{otherwise.} \end{cases}$$

$$p(\theta|D^3) \propto p(x|\theta)\,p(\theta|D^2) = \begin{cases} \frac{1}{\theta^3} & 7<\theta\le 10 \\ 0 & \text{otherwise.} \end{cases}$$

$$p(\theta|D^4) \propto p(x|\theta)\,p(\theta|D^3) = \begin{cases} \frac{1}{\theta^4} & 8<\theta\le 10 \\ 0 & \text{otherwise.} \end{cases}$$