

# 数据挖掘与机器学习

潘斌

panbin@nankai.edu.cn

范孙楼227

1

# 上节回顾

- 贝叶斯准则
  - 最小错误率Bayes决策
  - 最小风险Bayes决策
  - 最大最小Bayes决策
- 贝叶斯分类器的设计

# 本节提要

- 贝叶斯分类器
- 典型分类器：朴素贝叶斯
- 贝叶斯网络
- 贝叶斯神经网络

## 4.5.2 正态分布决策面

- 正态分布时分类器的决策面方程

- 二次型判别函数

展开

$$\begin{aligned}g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \\&= \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}\end{aligned}$$

$$\mathbf{W}_i = -\frac{1}{2} \boldsymbol{\Sigma}_i^{-1} \quad (d \times d \text{ 矩阵})$$

$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i \quad (d \text{ 维列向量})$$

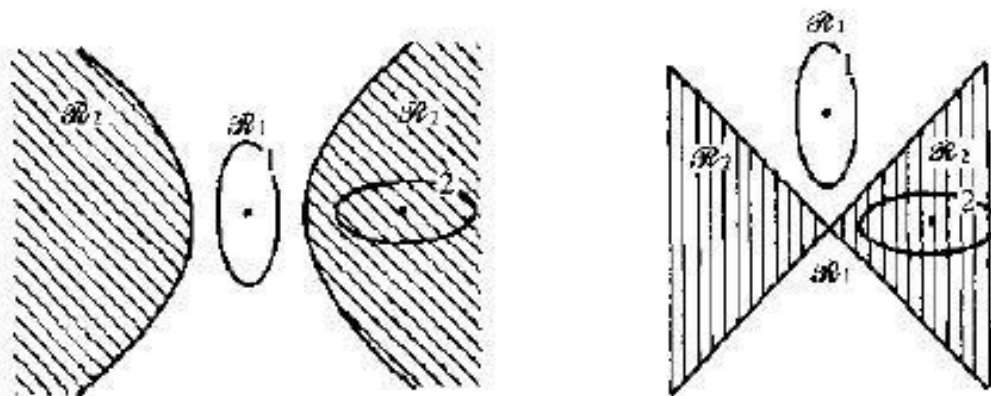
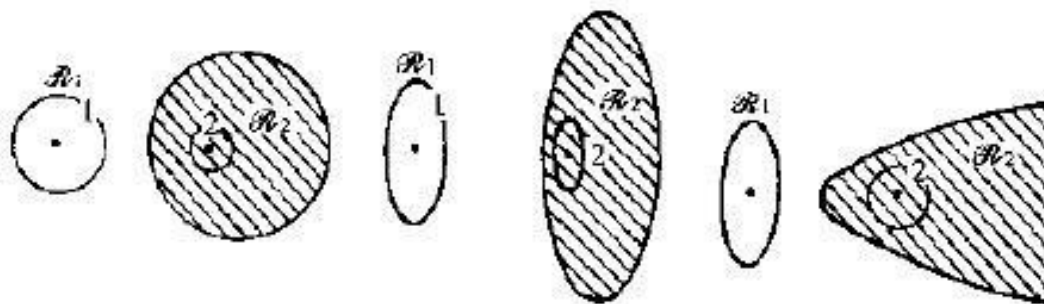
$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i)$$

- 决策面方程

相减

## 4.5.2 正态分布决策面

- 决策面示例



(d) 双曲线

(e) 直线

## 4.5.2 正态分布决策面

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i) \\ &= \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0} \end{aligned}$$

- 决策面特例一

- $\Sigma_i = \sigma^2 \mathbf{I} \quad i = 1, 2, \dots, C$

$$\Sigma_i = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix}$$

朴素贝叶斯  
分类器

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i) + \ln P(\omega_i)$$

准垂直平分

$$\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0) = 0$$

$$\mathbf{w} = \mu_i - \mu_j$$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

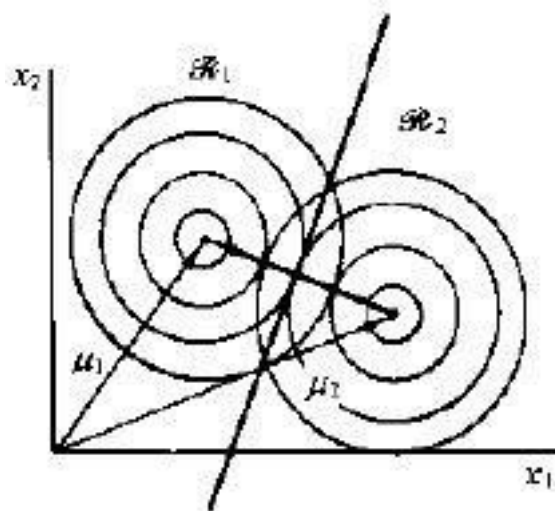
## 4.5.2 正态分布决策面

- 决策面特例一

$$w^T(x - x_0) = 0$$

$$w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$



## 4.5.2 正态分布决策面

- 决策面特例二

- $\Sigma_i = \Sigma \quad i = 1, 2, \dots, C$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma^{-1}(\mathbf{x} - \mu_i) + \ln P(\omega_i)$$

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$$
$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln \frac{P(\omega_i)}{P(\omega_j)}}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$



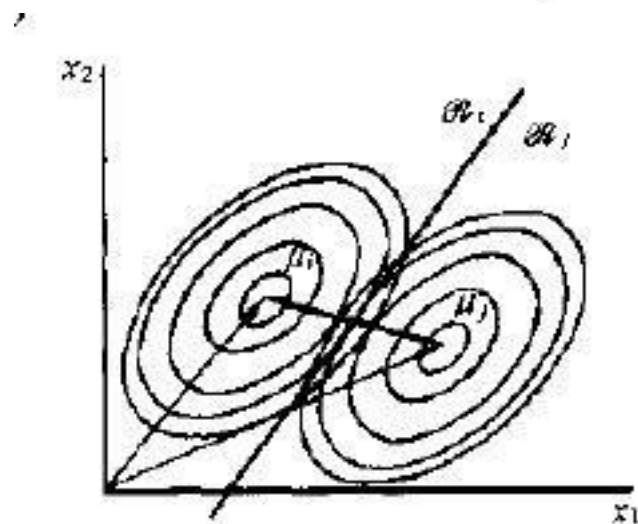
## 4.5.2 正态分布决策面

- 决策面特例二

$$\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0) = 0$$

$$\mathbf{w} = \Sigma^{-1}(\mu_i - \mu_j)$$

$$\mathbf{x}_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\ln \frac{P(\omega_i)}{P(\omega_j)}}{(\mu_i - \mu_j)^T \Sigma^{-1}(\mu_i - \mu_j)}(\mu_i - \mu_j)$$



## 4.5.3 错误率估计

- **Bayes**分类器错误率的估计方法
  - 按理论公式求解计算\*
  - 计算错误率上界
  - 实验估计错误率
- \*注：直接计算十分困难，只能在特例情况下进行。

## 4.5.3 错误率估计

- 错误率的计算
  - 两类正态分布等协方差阵情况下

$$P(e) = \int_{\mathfrak{R}_1} P(\omega_2) p(x | \omega_2) dx + \int_{\mathfrak{R}_2} P(\omega_1) p(x | \omega_1) dx$$

$$P(e_{12}) = \int_{\mathfrak{R}_1} P(\omega_2) p(x | \omega_2) dx = P(\omega_2) \int_{\mathfrak{R}_1} p(x | \omega_2) dx = P(\omega_2) P_2(e)$$

$$P(e_{21}) = \int_{\mathfrak{R}_2} P(\omega_1) p(x | \omega_1) dx = P(\omega_1) \int_{\mathfrak{R}_2} p(x | \omega_1) dx = P(\omega_1) P_1(e)$$

## 4.5.3 错误率估计

- 错误率的计算

回顾最小错误率贝叶斯决策规则的负对数似然比形式。

$$h(\mathbf{x}) = -\ln l(\mathbf{x}) = -\ln p(\mathbf{x}|\omega_1) + \ln p(\mathbf{x}|\omega_2) \leq \ln \left[ \frac{P(\omega_1)}{P(\omega_2)} \right] \rightarrow \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

$h(\mathbf{x})$  是  $\mathbf{x}$  的函数,  $\mathbf{x}$  是随机向量, 因此  $h(\mathbf{x})$  是随机变量。我们记它的分布密度函数为  $p(h|\omega_i)$ 。由于它是一维密度函数, 因此易于积分, 所以用它计算错误率有时较为方便。这样式(2-113)可表示为

$$P_1(e) = \int_{\mathcal{X}_2} p(\mathbf{x}|\omega_1) d\mathbf{x} = \int_t^\infty p(h|\omega_1) dh \quad (2-114)$$

$$P_2(e) = \int_{\mathcal{X}_1} p(\mathbf{x}|\omega_2) d\mathbf{x} = \int_{-\infty}^t p(h|\omega_2) dh \quad (2-115)$$

$$\text{其中 } t = \ln[P(\omega_1) | P(\omega_2)] \quad (2-116)$$

## 4.5.3 错误率估计

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \right\}$$

$$\mu = E(\mathbf{x})$$

$$\Sigma = E\{(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\}$$

- 错误率的计算

$$\begin{aligned} h(\mathbf{x}) &= -\ln l(\mathbf{x}) = -\ln p(\mathbf{x}|\omega_1) + \ln P(\mathbf{x}|\omega_2) \\ &= -\left[ -\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1}(\mathbf{x} - \mu_1) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_1| \right] \\ &\quad + \left[ -\frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma_2^{-1}(\mathbf{x} - \mu_2) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_2| \right] \\ &= \frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma_1^{-1}(\mathbf{x} - \mu_1) - \frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma_2^{-1}(\mathbf{x} - \mu_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} \\ &\leq \ln \frac{P(\omega_1)}{P(\omega_2)} \rightarrow x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases} \end{aligned}$$

## 4.5.3 错误率估计

- 错误率的计算
  - 两类正态分布等协方差阵情况下

$$h(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1}(\mathbf{x} - \mu_1) - \frac{1}{2}(\mathbf{x} - \mu_2)^T \Sigma^{-1}(\mathbf{x} - \mu_2) + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|}$$

$$h(\mathbf{x}) = (\mu_2 - \mu_1)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)$$

$$\lesssim \ln \frac{P(w_1)}{P(w_2)} \rightarrow \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

## 4.5.3 错误率估计

$$h(\mathbf{x}) = (\mu_2 - \mu_1)^T \Sigma^{-1} \mathbf{x} + \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)$$
$$\leq \ln \frac{P(\omega_1)}{P(\omega_2)} \rightarrow \mathbf{x} \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

- 错误率的计算

- $h\mathbf{x}$ 是 $\mathbf{x}$ 的线性组合，线性组合仍然服从正态分布：通过均值和方差来估计错误率

$$\eta_1 = E[h(\mathbf{x}) | \omega_1] = (\mu_2 - \mu_1)^T \Sigma^{-1} \mu_1 + \frac{1}{2} (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)$$

$$= -\frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$$

$$\eta = \frac{1}{2} [(\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)]$$

$$\eta_1 = -\eta$$

$$\sigma_1^2 = E\{[h(\mathbf{x}) - \eta]^2 | \omega_1\} = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) = 2\eta$$

$$\eta_2 = \frac{1}{2} (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) = \eta$$

$$\sigma_2^2 = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2) = 2\eta$$

$\mathbf{x}$ 是正态分布

## 4.5.3 错误率估计

- 错误率的计算
  - 线性组合仍然服从正态分布

$$\begin{aligned}P_1(e) &= \int_t^{\infty} p(h | \omega_1) dh \\&= \int_t^{\infty} \frac{1}{(2\pi)^{\frac{1}{2}} \sigma} \exp\left\{-\frac{1}{2}\left(\frac{h + \eta}{\sigma}\right)^2\right\} dh \\&= \int_t^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{h + \eta}{\sigma}\right)^2\right\} d\left(\frac{h + \eta}{\sigma}\right) \\&= \int_{\left(\frac{t + \eta}{\sigma}\right)}^{\infty} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\xi^2\right\} d\xi\end{aligned}$$

$$t = \ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right], \sigma = \sqrt{2\eta}$$



## 4.5.3 错误率估计

- 错误率的计算
  - 线性组合仍然服从正态分布

$$\begin{aligned}P_2(e) &= \int_{-\infty}^t p(h|\omega_2)dh \\&= \int_{-\infty}^t (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\left(\frac{h-\eta}{\sigma}\right)^2\right\} d\left(\frac{h-\eta}{\sigma}\right) \\&= \int_{-\infty}^{\frac{t-\eta}{\sigma}} (2\pi)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\xi^2\right\} d\xi\end{aligned}$$

$$t = \ln\left[\frac{P(\omega_1)}{P(\omega_2)}\right], \sigma = \sqrt{2\eta}$$

## 4.5.3 错误率上界

- 两类情况下最小错误率的上界

- Chernoff切尔诺夫上界

$$P(e_{12}) \leq P(\omega_2) e^{[-\mu(s) + (1-s)t]}$$

$$P(e_{21}) \leq P(\omega_1) e^{[-\mu(s) - st]}$$

- Bhattacharyya巴氏系数上界

$$J_B = -\ln \left[ \int \sqrt{p(x | \omega_1) p(x | \omega_2)} dx \right]$$

$$P(e) \leq \sqrt{P(\omega_1) P(\omega_2)} \exp \{-J_B\}$$

## 4.5.3 实验估计错误率

- 错误率的实验估计
  - 测试样本集（区别于训练样本集）
  - 测试结果（有限样本情况，混淆矩阵）
  - 错误率评价

## 4.5.4 讨论

- **Bayes**决策的先决条件
  - 类条件概率密度的估计
  - 例：密度函数估计的简化
    - 假设特征向量各分量相互独立

## 4.5.4 讨论

- **Bayes**决策的最小错误概率
  - 设计难度大（类条件概率密度估计）
  - 寻求高“性价比”的分类器
    - 错误概率逼近
    - 决策面逼近

# 朴素贝叶斯分类器

**Naïve Bayes Classifier**

# 朴素贝叶斯分类器 ( NAÏVE BAYES CLASSIFIER )

- 在给定目标值时，属性值之间相互条件独立，即

$$P(a_1, \dots, a_n | v_j) = \prod_i P(a_i | v_j)$$

$$P(x_1, \dots, x_n | w_j) = \prod_i P(x_i | w_j)$$

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

贝叶斯公式

$$P(\theta | X) = \frac{P(X | \theta) \times P(\theta)}{P(X)}$$

- 决策面特例一

- $\Sigma_i = \sigma^2 I \quad i = 1, 2, \dots, C$

$$\Sigma_i = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix}$$

朴素贝叶斯  
分类器

$$g_i(x) = -\frac{1}{2\sigma^2}(x - \mu_i)^T(x - \mu_i) + \ln P(\omega_i)$$

准垂直平分

$$w^T(x - x_0) = 0$$

$$w = \mu_i - \mu_j$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$



Naive\_Bayes\_Learn(*examples*)

For each target value  $v_j$


$\hat{P}(v_j) \leftarrow \text{estimate } P(v_j)$

For each attribute value  $a_i$  of each attribute  $a$

$\hat{P}(a_i|v_j) \leftarrow \text{estimate } P(a_i|v_j)$

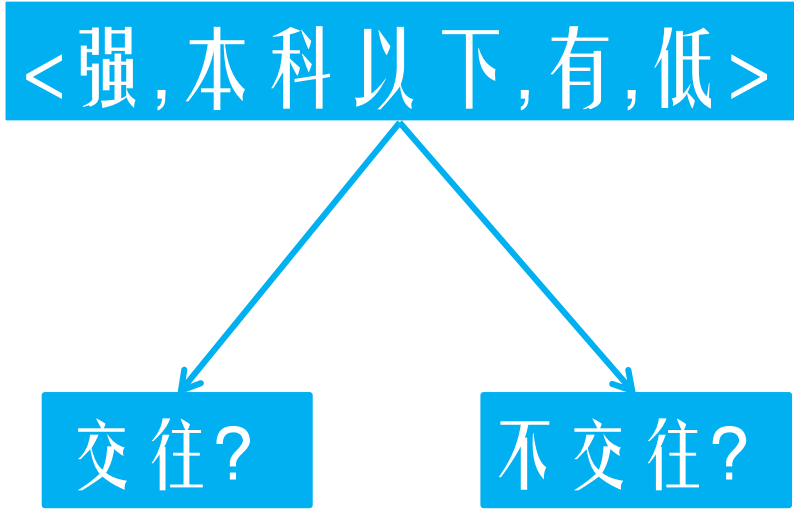
Classify\_New\_Instance( $x$ )

$$v_{NB} = \operatorname{argmax}_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

对比   $P(X|\theta) \times P(\theta)$

例：

上进心	学历	房产	年薪	交往
强	研究生	有	高	是
强	研究生	有	低	是
弱	研究生	有	高	否
一般	本科	有	高	否
一般	本科以下	无	高	否
一般	本科以下	无	低	是
弱	本科以下	无	低	否
强	本科	有	高	是
强	本科以下	无	高	否
一般	本科	无	高	否
强	本科	无	低	否
弱	本科	有	低	否
弱	研究生	无	高	否
一般	本科	有	低	是



$$\begin{aligned}
 \blacksquare \text{ 例 ( 续 ) } v_{NB} &= \operatorname{argmax}_{v_j \in \{\text{是}, \text{否}\}} P(v_j) \prod_i P(a_i | v_j) \\
 &= \operatorname{argmax}_{v_j \in \{\text{是}, \text{否}\}} P(v_j) P(\text{强} | v_j) P(\text{本科以下} | v_j) P(\text{有} | v_j) P(\text{低} | v_j)
 \end{aligned}$$

■ 根据数据集，可以计算出上式需要的概率值

- $P(\text{否})=9/14=0.64$ ； $P(\text{是})=5/14=0.36$ ；
- $P(\text{强} | \text{否})=2/9=0.22$ ； $P(\text{强} | \text{是})=3/5=0.6$ ；
- $P(\text{弱} | \text{否})=4/9=0.44$ ； $P(\text{弱} | \text{是})=0/5=0$ ；
- $P(\text{一般} | \text{否})=3/9=0.33$ ； $P(\text{一般} | \text{是})=2/5=0.4$ ；
- $P(\text{研究生} | \text{否})=2/9=0.22$ ； $P(\text{研究生} | \text{是})=2/5=0.4$ ；
- $P(\text{本科} | \text{否})=4/9=0.44$ ； $P(\text{本科} | \text{是})=2/5=0.4$ ；
- $P(\text{本科以下} | \text{否})=3/9=0.33$ ； $P(\text{本科以下} | \text{是})=1/5=0.2$ ；
- $P(\text{无} | \text{否})=6/9=0.67$ ； $P(\text{无} | \text{是})=1/5=0.2$ ；
- $P(\text{有} | \text{否})=3/9=0.33$ ； $P(\text{有} | \text{是})=4/5=0.8$ ；
- $P(\text{高} | \text{否})=6/9=0.67$ ； $P(\text{高} | \text{是})=2/5=0.4$ ；
- $P(\text{低} | \text{否})=3/9=0.33$ ； $P(\text{低} | \text{是})=3/5=0.4$

$$\begin{aligned}
 v_{NB} &= \operatorname{argmax}_{v_j \in \{\text{是}, \text{否}\}} P(v_j) \prod_i P(a_i | v_j) \\
 &= \operatorname{argmax}_{v_j \in \{\text{是}, \text{否}\}} P(v_j) P(\text{强} | v_j) P(\text{本科以下} | v_j) P(\text{有} | v_j) P(\text{低} | v_j)
 \end{aligned}$$

- 求  $v_{NB}$ 
  - $P(\text{否})P(\text{强} | \text{否})P(\text{本科以下} | \text{否})P(\text{有} | \text{否})P(\text{低} | \text{否}) = 0.0053$
  - $P(\text{是})P(\text{强} | \text{是})P(\text{本科以下} | \text{是})P(\text{有} | \text{是})P(\text{低} | \text{是}) = 0.0206$
  - $v_{NB} = \text{是}$
- 通过将上述的量归一化，可计算给定观察值下目标值为“是”的条件概率为  $0.0206 / (0.0206 + 0.0053) = 0.795$ 。

# ■ 练习：

Name	Give Birth	Can Fly	live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometime	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometime	yes	non-mammals
penguin	no	no	sometime	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometime	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	live in Water	Have Legs	Class
yes	no	yes	no	?

例

<i>Tid</i>	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$  ?

# 估计连续型属性的条件概率-离散化

- 将每一连续属性离散化，然后利用相应的离散区间替换连续属性值
- 把 “Taxable Income” 划分成两个区间，最佳的候选划分点为97K，对应区间为  $(0, 97)$  和  $[97, 10000)$ 。通过计算不同类别中属性 “Taxable Income” 落入对应区间的比例来估计条件概率。

<i>Tid</i>	Refund	Marital Status	Taxable Income < 97K	Evade
1	Yes	Single	No	No
2	No	Married	No	No
3	No	Single	Yes	No
4	Yes	Married	No	No
5	No	Divorced	Yes	Yes
6	No	Married	Yes	No
7	Yes	Divorced	No	No
8	No	Single	Yes	Yes
9	No	Married	Yes	No
10	No	Single	Yes	Yes

- 用Bayes方法估计每个条件概率后，对之前新给出的样本可以进行判别。

$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = \text{No})$



<i>Tid</i>	Refund	Marital Status	Taxable Income< 97K	Evade
1	Yes	Single	No	No
2	No	Married	No	No
3	No	Single	Yes	No
4	Yes	Married	No	No
5	No	Divorced	Yes	Yes
6	No	Married	Yes	No
7	Yes	Divorced	No	No
8	No	Single	Yes	Yes
9	No	Married	Yes	No
10	No	Single	Yes	Yes

- $$P(X | \text{Evade}=\text{No}) = P(\text{Refund}=\text{No} | \text{Evade}=\text{No}) \\ P(\text{Married} | \text{Evade}=\text{No}) \\ P(\text{Income}=\text{No} | \text{Evade}=\text{No}) \\ = 4/7 \times 4/7 \times 4/7 = 0.1866$$

- $$P(X | \text{Evade}=\text{Yes}) = P(\text{Refund}=\text{No} | \text{Evade}=\text{Yes}) \\ P(\text{Married} | \text{Evade}=\text{Yes}) \\ P(\text{Income}=\text{No} | \text{Evade}=\text{Yes}) \\ = 1 \times 0 \times 0 = 0$$

- $$\text{Since } P(X | \text{No})P(\text{No}) > P(X | \text{Yes})P(\text{Yes})$$

Therefore  $P(\text{No} | X) > P(\text{Yes} | X)$   
 $\Rightarrow \text{Evade} = \text{No}$

# 估计连续型属性的条件概率-概率分布

- 假设连续型属性服从某种概率分布（通常假设服从正态分布），然后用训练数据估计出分布的参数，进而计算相应的条件概率。如上例中，假设“Taxable Income”属性为随机变量

$$X_3 \sim N(\mu, \sigma^2)$$

- 对于每个类  $C_i$ ，属性值  $x_j$  属于类  $C_i$  的概率为

$$P(x_j|C_i) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp\left(-\frac{(x_j - \mu_{ij})^2}{2\sigma_{ij}^2}\right)$$

- $\mu_{ij}$  和  $\sigma_{ij}$  分别为类  $C_i$  中随机变量  $x_j$  的期望和方差，可分别用  $C_i$  中  $x_j$  的观察值的样本均值和标准差估计

- 如上表数据中 “Taxable Income” 数据，分别属于两类，设类别  $C1=$  “No”， $C2=$  “Yes”，对应的观察值如下：

Taxable Income	125	100	70	120	95	60	220	85	75	90
Evade	No	No	No	No	Yes	No	No	Yes	No	Yes

- 类别  $C1=$  “No” 的两个参数估计如下：

- $\bar{X} = \frac{1}{7}(125 + 100 + 70 + 120 + 60 + 220 + 75) = 110$

- $S^2 = \frac{1}{6}\{(125 - 110)^2 + (100 - 110)^2 + (70 - 110)^2 + (120 - 110)^2 + (60 - 110)^2 + (220 - 110)^2 + (75 - 110)^2\} = 2975$

- $(\mu, \sigma^2) = (110, 54.54^2)$

- 同理，类别  $C2=$  “Yes” 的两个参数估计为： $(\mu, \sigma^2) = (90, 5^2)$

- $P(\text{Income} = 120 | \text{No}) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$

## naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No:      sample mean=110  
                         sample variance=2975

If class=Yes:      sample mean=90  
                         sample variance=25

$$\begin{aligned} P(X | \text{Class}=\text{No}) &= P(\text{Refund}=\text{No} | \text{Class}=\text{No}) \\ &\times P(\text{Married} | \text{Class}=\text{No}) \\ &\times P(\text{Income}=120\text{K} | \text{Class}=\text{No}) \\ &= 4/7 \times 4/7 \times 0.0072 = 0.0024 \end{aligned}$$

$$\begin{aligned} P(X | \text{Class}=\text{Yes}) &= P(\text{Refund}=\text{No} | \text{Class}=\text{Yes}) \\ &\times P(\text{Married} | \text{Class}=\text{Yes}) \\ &\times P(\text{Income}=120\text{K} | \text{Class}=\text{Yes}) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} = 0 \end{aligned}$$

Since  $P(X | \text{No})P(\text{No}) > P(X | \text{Yes})P(\text{Yes})$

Therefore  $P(\text{No} | X) > P(\text{Yes} | X)$   
 $\Rightarrow \text{Class} = \text{No}$

# 问题

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

naive Bayes Classifier:

$$P(\text{Refund}=\text{Yes}|\text{No}) = 3/7$$

$$P(\text{Refund}=\text{No}|\text{No}) = 4/7$$

$$P(\text{Refund}=\text{Yes}|\text{Yes}) = 0$$

$$P(\text{Refund}=\text{No}|\text{Yes}) = 1$$

$$P(\text{Marital Status}=\text{Single}|\text{No}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{No}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{No}) = 4/7$$

$$P(\text{Marital Status}=\text{Single}|\text{Yes}) = 2/7$$

$$P(\text{Marital Status}=\text{Divorced}|\text{Yes}) = 1/7$$

$$P(\text{Marital Status}=\text{Married}|\text{Yes}) = 0$$

For taxable income:

If class=No:      sample mean=110  
                         sample variance=2975

If class=Yes:     sample mean=90  
                         sample variance=25

## ■ 利用训练样例估计后验概率的潜在问题

- 若一属性的类条件概率为0，在整个类的后验概率为0
- 一个极端的例子，训练样例不能覆盖那么多属性值时，可能无法分类某些测试记录

## ■ 一种解决方案

- 当样本很小时，采用平滑技术，**m-估计**  $\frac{n_a + mp}{n_c + m}$

- $p$ 是将要确定的概率的先验估计，而 $m$ 是一称为等效样本大小的常量，可取属性个数
- 在缺少其他信息时，选择 $p$ 的一种典型的方法是均匀概率，比如某属性有 $k$ 个可能值，那么 $p=1/k$ 。

- 例 ( 续 )
- 前例中,  $P(\text{Married} \mid \text{Class}=\text{Yes})=0$
- 使用m-估计,  $m=3$ ,  $p=1/3$
- $P(\text{Married} \mid \text{Class}=\text{Yes})=(0+ 3 \times 1/3)/(3+3)= 1/6$
- $P(X \mid \text{Class}=\text{No}) = P(\text{Refund}=\text{No} \mid \text{Class}=\text{No})P(\text{Married} \mid \text{Class}=\text{No})P(\text{Income}=120\text{K} \mid \text{Class}=\text{No}) = 6/10 \times 6/10 \times 0.0072 = 0.0026$
- $P(X \mid \text{Class}=\text{Yes}) = P(\text{Refund}=\text{No} \mid \text{Class}=\text{Yes}) P(\text{Married} \mid \text{Class}=\text{Yes}) P(\text{Income}=120\text{K} \mid \text{Class}=\text{Yes}) = 4/6 \times 1/6 \times 1.2 \times 10^{-9} = 1.3 \times 10^{-10}$

## ■ 小结

- 朴素贝叶斯分类器假定了属性 $a_1, \dots, a_n$ 的值在给定目标值 $v$ 下是条件独立的。
- 这一假定显著地减小了目标函数学习的计算复杂度。
- 当此条件成立时，朴素贝叶斯分类器可得到最优贝叶斯分类。
- 但在多数情况下，这一条件独立假定过于严厉了。



# 实验3：基于朴素贝叶斯的犯罪类型预测

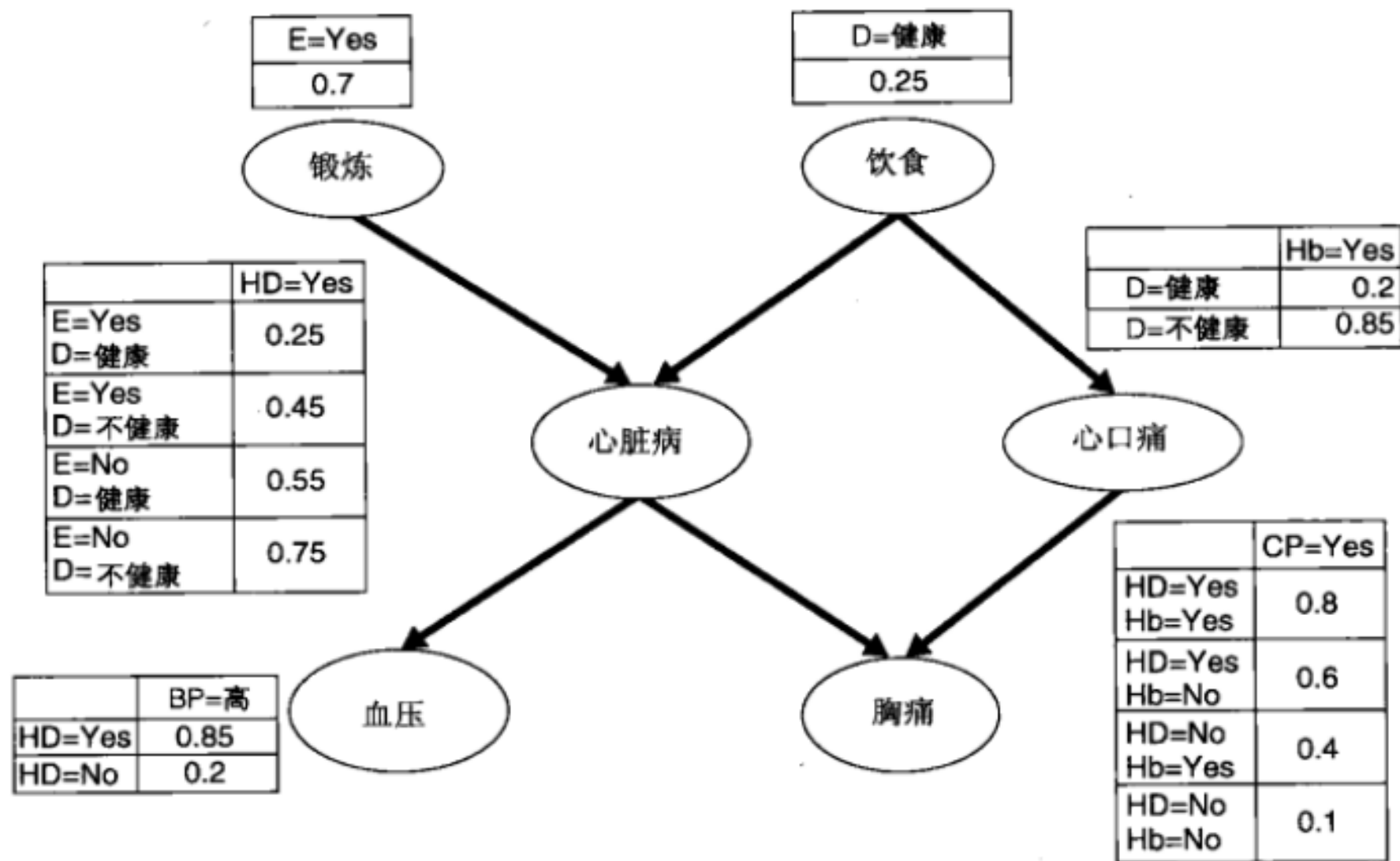
- 给定LA犯罪数据集
- 根据某案件的“案发时间”“所属警区”“案发地点”等属性，预测案件的类型，例如“抢劫”“偷窃”“杀人”等
- 属性均为离散的
- 共44948个训练样本，14983个测试样本
- Python编程实现

# 贝叶斯信念网络 ( BAYES BELIEF NETWORKS, BBN )

- 简称贝叶斯网络，表述变量的一个子集上的条件独立性假定。
- 贝叶斯网络中的一个节点，如果它的父母节点已知，则它条件独立与它的所有非后代节点。

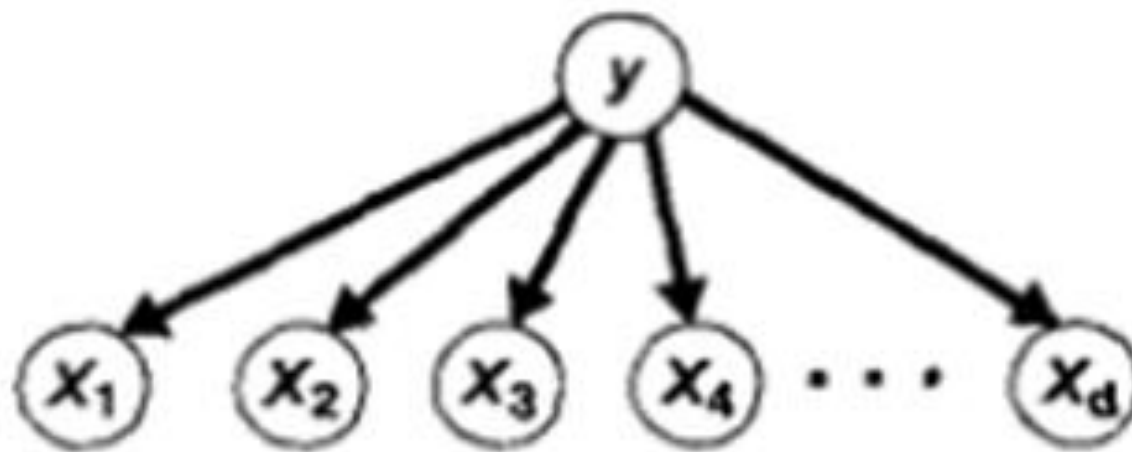
# 贝叶斯信念网络的表示

- 一个有向无环图(directed acyclic graph, or DAG)，指定一组条件独立性假定，表示变量间的依赖关系
  - 每个节点表示一个变量，每条弧表示两个变量间的依赖关系
  - 如从 $X$ 到 $Y$ 有一条有向弧，则 $X$ 是 $Y$ 的父母， $Y$ 是 $X$ 的子女
  - 如网络中存在一条从 $X$ 到 $Z$ 的有向路径，则 $X$ 是 $Z$ 的祖先， $Z$ 是 $X$ 的后代
- 一个概率表，即一组局部条件概率集合，把各节点和它的直接父节点关联起来
  - 如节点 $X$ 没有父母节点，则表中只包含先验概率 $P(X)$
  - 如节点 $X$ 只有一个父母节点 $Y$ ，则表中包含条件概率 $P(X|Y)$
  - 如节点 $X$ 有多个父母节点 $\{Y_1, \dots, Y_K\}$ ，则表中包含条件概率 $P(X|Y_1, \dots, Y_K)$

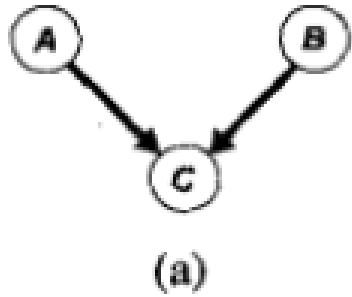


发现心脏病和心口痛病人的贝叶斯网络

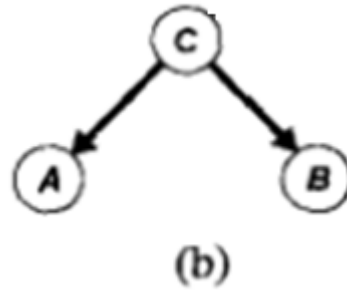
朴素贝叶斯分类器中的条件独立假设也可以用贝叶斯网络表示。



# 贝叶斯网络中三个变量间的典型依赖关系



V型结构 ( V-Structure )、  
冲撞结构



“同父”结构 ( CommonParent )



顺序结构

# 贝叶斯神经网络 ( BAYESIAN NEURAL NETWORK )

- 什么是贝叶斯神经网络
- 为什么需要贝叶斯神经网络
- 贝叶斯神经网络如何实现



# 什么是贝叶斯神经网络

- 传统神经网络通过最小化损失函数求出参数的点估计
- 贝叶斯神经网络希望求 $w$ 的分布而不是 $w$ 的极大似然估计
- 网络的输出表示为

$$P(y|x) = E_{P(w|D)}[P(y|x, w)]$$

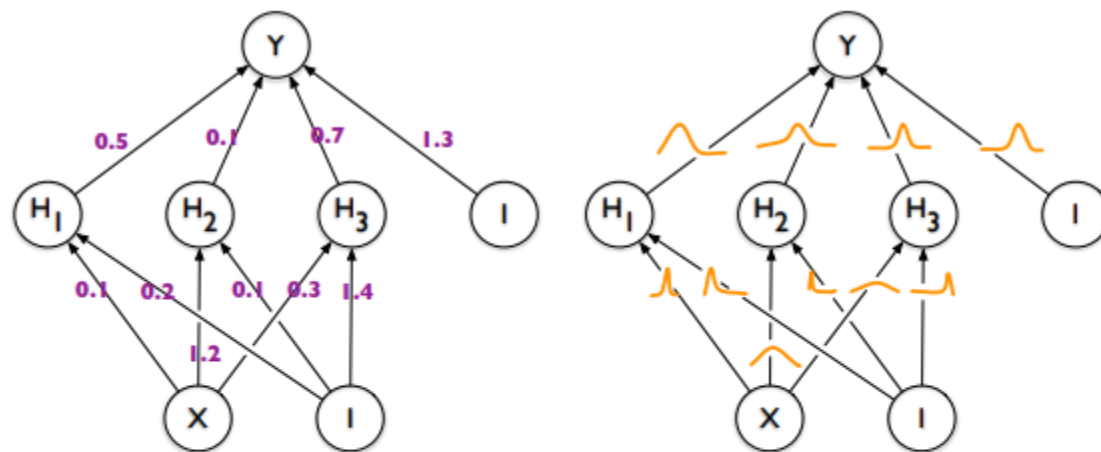


Figure 1. Left: each weight has a fixed value, as provided by classical backpropagation. Right: each weight is assigned a distribution, as provided by Bayes by Backprop.





# 为什么需要贝叶斯神经网络

- 1、贝叶斯神经网络相当于最全面的集成模型
- $w \sim P(w|D)$ , 每个  $w$  的样本都对应一个可能的模型
- $P(y|x) = E_{P(w|D)}[P(y|x, w)]$  即是所有可能模型的预测结果的加权平均

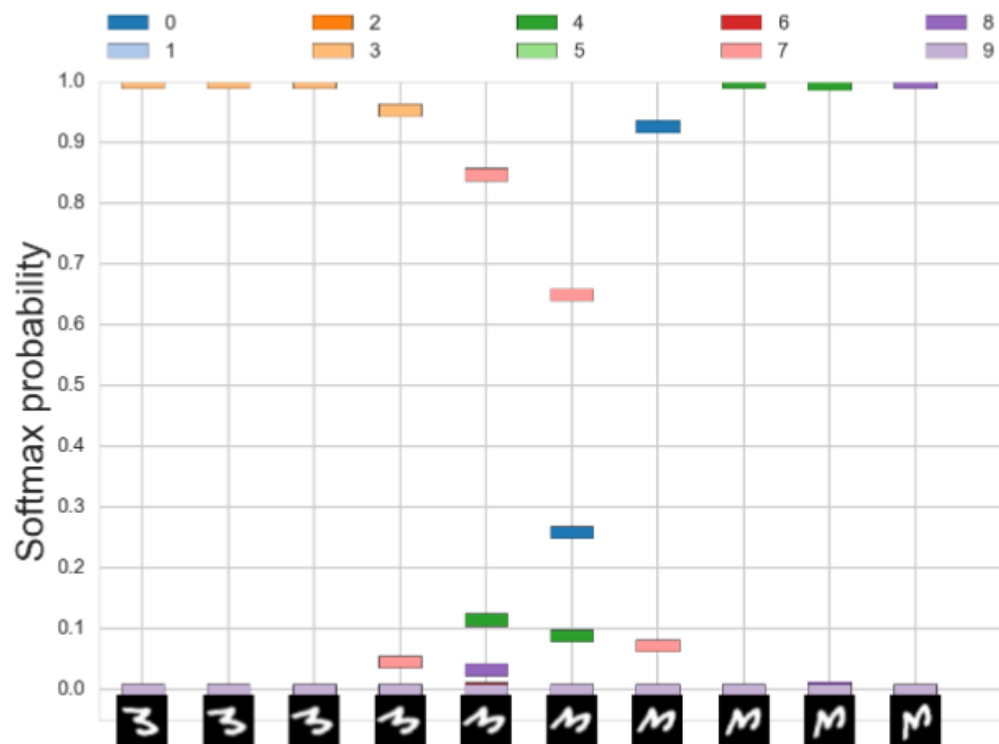


# 为什么需要贝叶斯神经网络

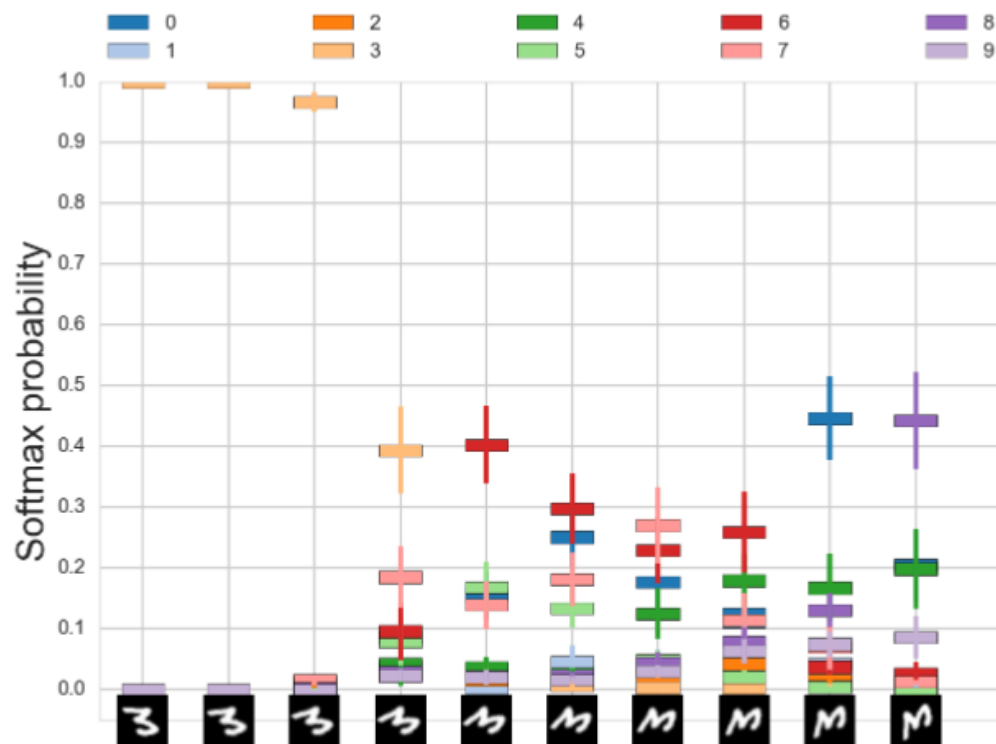
- 2、贝叶斯神经网络可以估计预测结果的 **不确定性**
- 普通神经网络的输出是点估计，通常过于自信
- 贝叶斯神经网络可以推导出输出的分布



## 贝叶斯神经网络可以估计预测结果的不确定性 ( UNCERTAINTY )



(a) LeNet with weight decay



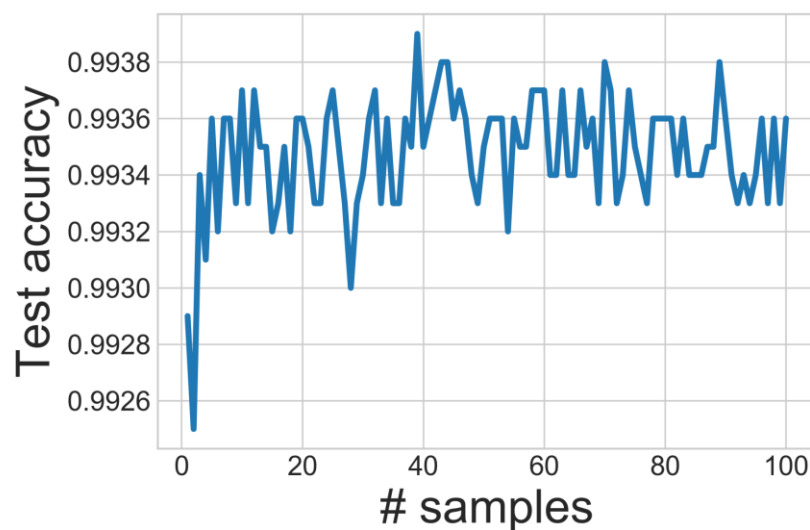
(b) LeNet with multiplicative formalizing flows

Louizos, Christos, and Max Welling. "Multiplicative normalizing flows for variational bayesian neural networks." ICML 2017

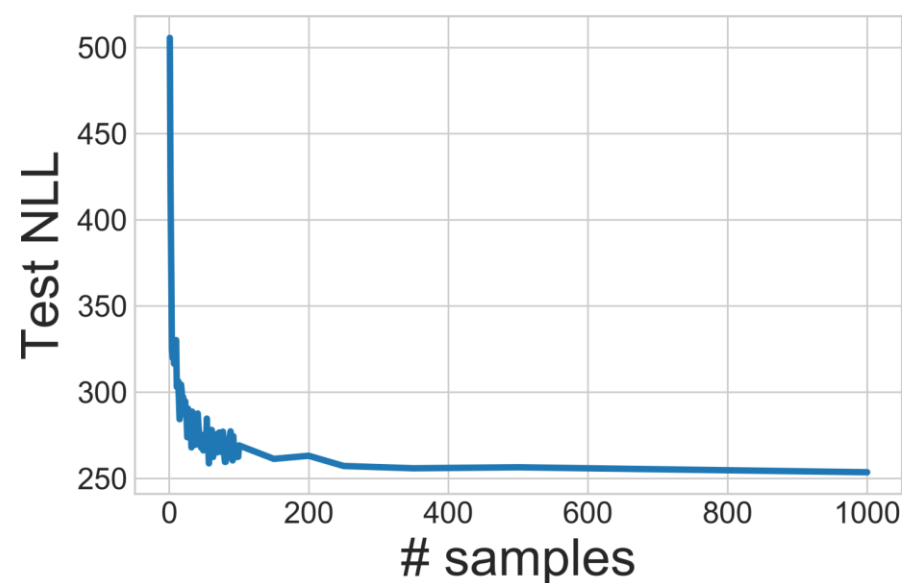


# 为什么需要贝叶斯神经网络

- 贝叶斯神经网络还可以反映网络训练过程中的不确定性



测试准确率快速接近最优值并且开始摆动



但负对数似然持续下降



# 贝叶斯神经网络如何实现

- $P(w|D)$  维度极高，直接推导比较困难，而且不易于使用
- 使用变分推断 (variational inference) 估计  $P(w|D)$ ，即寻找一个简单的分布  $q(w|\theta)$  来逼近  $P(w|D)$
- 逼近过程可以通过最小化两个分布的 KL 散度 (Kullback-Leibler

divergence) 实现： 
$$D_{KL}[q(w)||p(w)] = \int q(w) \log \frac{q(w)}{p(w)} dw$$



## ■ $\theta$ 的求导过程：

$$\begin{aligned}\theta^* &= \operatorname{argmin}_{\theta} D_{KL}[q(w|\theta)||P(w|D)] \\&= \operatorname{argmin}_{\theta} \int q(w|\theta) \log \frac{q(w|\theta)}{P(w|D)} dw && \text{注意到 } P(w|D) = \frac{P(D|w)P(w)}{P(D)} \\&= \operatorname{argmin}_{\theta} \int q(w|\theta) \log \frac{q(w|\theta)P(D)}{P(w)P(D|w)} dw \\&= \operatorname{argmin}_{\theta} \int q(w|\theta) \log \frac{q(w|\theta)}{P(w)} dw - \int q(w|\theta) \log P(D|w) dw + \int q(w|\theta) \log P(D) dw \\&= \operatorname{argmin}_{\theta} D_{KL}[q(w|\theta)||P(w)] - E_{q(w|\theta)}[\log P(D|w)] + \log P(D) \\&\hspace{15em} \text{常数}\end{aligned}$$

## ■ 最终，网络的优化目标为：

$$F(D, \theta) = -E_{q(w|\theta)}[\log P(D|w)] + D_{KL}[q(w|\theta)||P(w)]$$

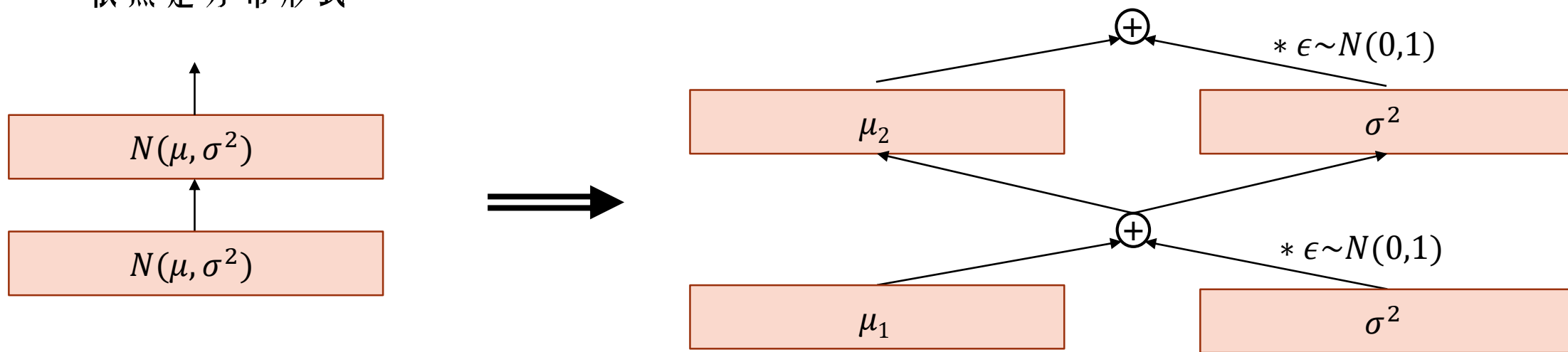
负对数似然，  
理解为一般损失函数

正则化项



# 贝叶斯神经网络如何实现

- 虽然利用变分推断简化了贝叶斯神经网络，但如何利用梯度下降法对分布进行优化？
- 重参数法（Reparameterization）：  
假设变分分布为正态分布  $q(w|\theta) = q(w|\mu, \sigma^2) = N(\mu, \sigma^2)$ ，  
可以把参数的随机性迁移出去  $(w|\mu, \sigma^2) = \mu + \sigma^2 \cdot \epsilon, \epsilon \sim N(0,1)$
- 这样就把贝叶斯神经网络的待求解的参数转化成了普通数值，但最终向前传播时的参数依然是分布形式





**THE END !**