数据挖掘与机器学习

1

潘斌

panbin@nankai.edu.cn

范孙楼227

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
- 典型分类器：决策树

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

$$g_i(\boldsymbol{x})=-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)-\frac{1}{2}\ln|\Sigma_i|+\ln P(\omega_i)$$

$$=\boldsymbol{x}^T W_i\boldsymbol{x}+\boldsymbol{w}_i^T\boldsymbol{x}+w_{i0}$$

$$W_i=-\frac{1}{2}\Sigma_i^{-1} \qquad (d\times d\text{ 矩阵})$$

$$\boldsymbol{w}_i=\Sigma_i^{-1}\boldsymbol{\mu}_i \qquad (d\text{ 维列向量})$$

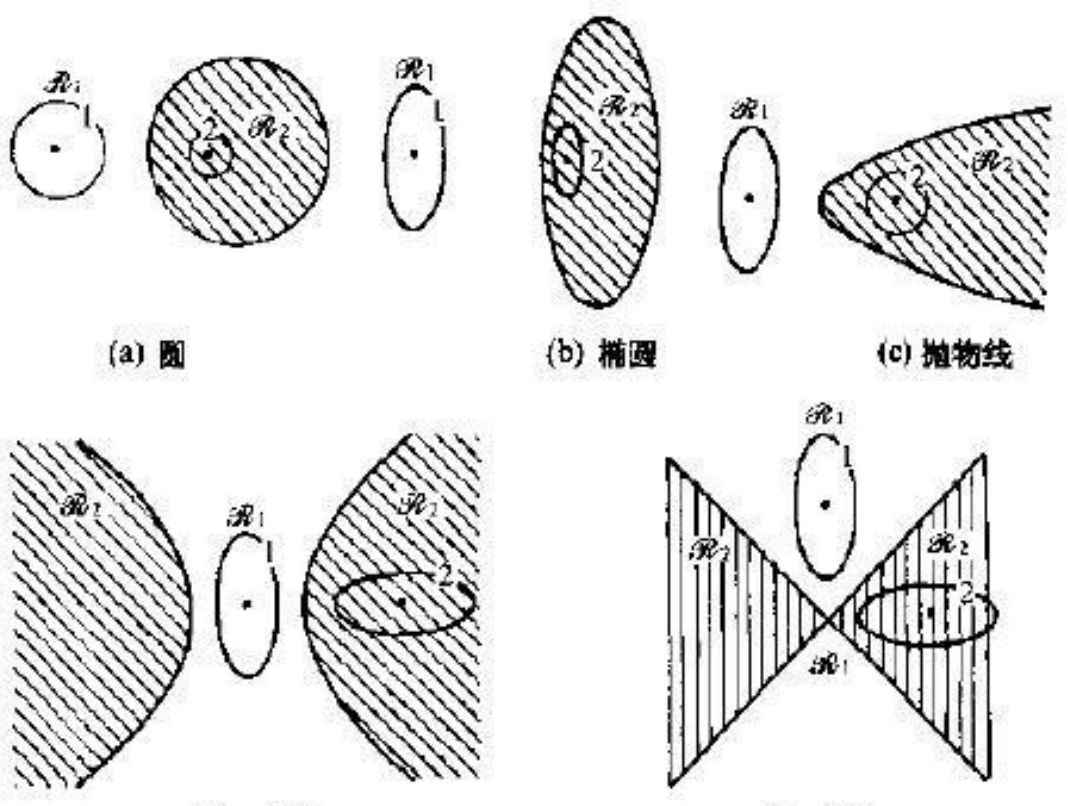
$$w_{i0}=-\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i-\frac{1}{2}\ln|\Sigma_i|+\ln P(\omega_i)$$

  - 决策面方程

相减

4.5.2 正态分布决策面

- 决策面示例



(a) 圆　(b) 椭圆　(c) 抛物线

(d) 双曲线　(e) 直线

4.5.2 正态分布决策面

$$g_i(\boldsymbol{x}) = -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$
$$= \boldsymbol{x}^T W_i \boldsymbol{x} + \boldsymbol{w}_i^T \boldsymbol{x} + w_{i0}$$

- 决策面特例一
  - $\Sigma_i = \sigma^2 I$　　　　　$i = 1, 2, \ldots, C$

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

朴素贝叶斯分类器

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

准垂直平分

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

$$\boldsymbol{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

$$\boldsymbol{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2}\ln\frac{P(\omega_i)}{P(\omega_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

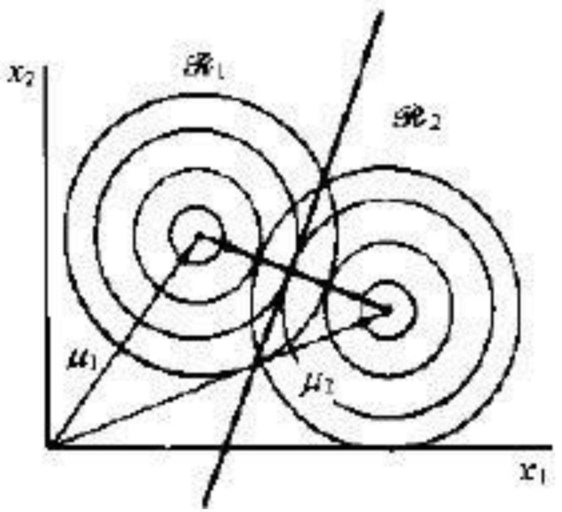
4.5.2 正态分布决策面

- 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
  - $\Sigma_i = \Sigma$　　　　　　　　$i = 1, 2, \ldots, C$

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

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

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln\dfrac{P(\omega_i)}{P(\omega_j)}}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T\Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

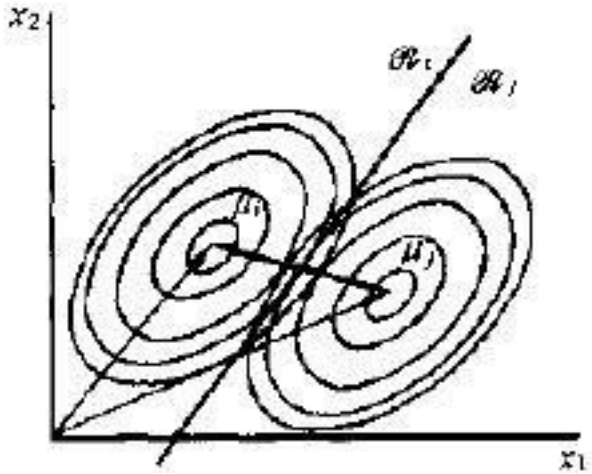
4.5.2 正态分布决策面

- 决策面特例二

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

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

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



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

$$P(e)=\int_{\Re_1} P(\omega_2)\, p(x\mid\omega_2)dx+\int_{\Re_2} P(\omega_1)\, p(x\mid\omega_1)dx$$

$$P(e_{12})=\int_{\Re_1} P(\omega_2)\, p(x\mid\omega_2)dx=P(\omega_2)\int_{\Re_1} p(x\mid\omega_2)dx=P(\omega_2)P_2(e)$$

$$P(e_{21})=\int_{\Re_2} P(\omega_1)\, p(x\mid\omega_1)dx=P(\omega_1)\int_{\Re_2} p(x\mid\omega_1)dx=P(\omega_1)P_1(e)$$

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。

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

$h(x)$是x的函数，x是随机向量，因此$h(x)$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

$$P_1(e)=\int_{\mathscr{R}_2}p(x|\omega_1)\mathrm{d}x=\int_t^{\infty}p(h|\omega_1)\mathrm{d}h \tag{2-114}$$

$$P_2(e)=\int_{\mathscr{R}_1}p(x|\omega_2)\mathrm{d}x=\int_{-\infty}^{t}p(h|\omega_2)\mathrm{d}h \tag{2-115}$$

其中 $t=\ln[P(\omega_1)/P(\omega_2)]$ (2-116)

4.5.3 错误率估计

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

$$\mu=E\{x\}$$

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

- 错误率的计算

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

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

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

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

4.5.3 错误率估计

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

- **错误率的计算**
  - hx是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

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

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

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

X是正态分布

$$\eta_1=-\eta$$

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

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

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

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

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

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

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

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

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

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

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

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

  - Bhattacharyya巴氏系数上界

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

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

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

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

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

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

$\text{Naive\_Bayes\_Learn}(examples)$

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

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

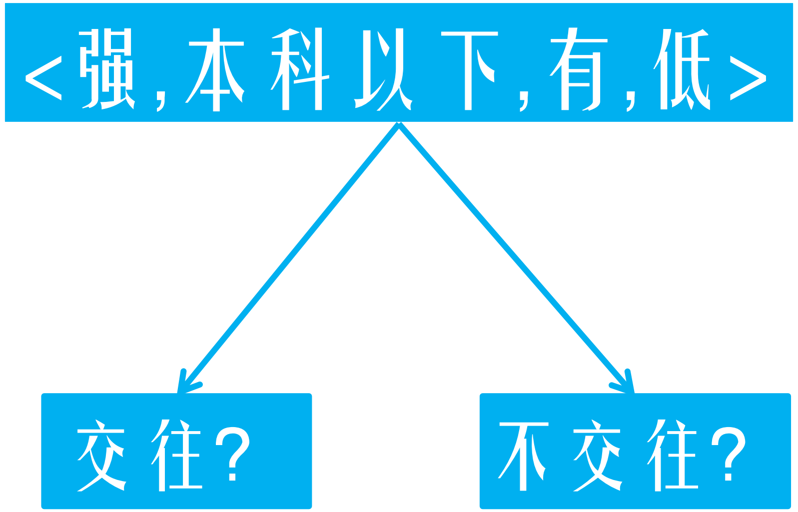
$\text{Classify\_New\_Instance}(x)$

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

对比 $\longrightarrow$ $P(X|\theta) \times P(\theta)$

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



- 例（续）$v_{NB} = \underset{v_j \in \{是,否\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$

$$= \underset{v_j \in \{是,否\}}{\operatorname{argmax}} P(v_j) P(强|v_j) P(本科以下|v_j) P(有|v_j) P(低|v_j)$$

  - 根据数据集，可以计算出上式需要的概率值
    - P(否)=9/14=0.64; P(是)=5/14=0.36;
    - P(强|否)=2/9=0.22; P(强|是)=3/5=0.6;
    - P(弱|否)=4/9=0.44; P(弱|是)=0/5=0;
    - P(一般|否)=3/9=0.33; P(一般|是)=2/5=0.4;
    - P(研究生|否)= 2/9=0.22 ; P(研究生|是)=2/5=0.4;
    - P(本科|否)= 4/9=0.44 ; P(本科|是)=2/5=0.4;
    - P(本科以下|否)=3/9=0.33; P(本科以下|是)=1/5=0.2;
    - P(无|否)=6/9=0.67; P(无|是)=1/5=0.2;
    - P(有|否)=3/9=0.33; P(有|是)=4/5=0.8;
    - P(高|否)=6/9=0.67; P(高|是)=2/5=0.4;
    - P(低|否)=3/9=0.33; P(低|是)=3/5=0.4

$$v_{NB} = \underset{v_j \in \{是,否\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$
$$= \underset{v_j \in \{是,否\}}{\operatorname{argmax}} P(v_j) P(强|v_j) P(本科以下|v_j) P(有|v_j) P(低|v_j)$$

- 求 v_{NB}
  - P(否)P(强 | 否)P(本科以下 | 否)P(有 | 否)P(低 | 否) =0.0053
  - P(是)P(强 | 是)P(本科以下 | 是)P(有 | 是)P(低 | 是) =0.0206
  - v_{NB}=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	ve in Wat	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mamma
salmon	no	no	yes	no	non-mamma
whale	yes	no	yes	no	mammals
frog	no	no	sometime	yes	non-mamma
komodo	no	no	no	yes	non-mamma
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mamma
cat	yes	no	no	yes	mammals
leopard shar	yes	no	yes	no	non-mamma
turtle	no	no	sometime	yes	non-mamma
penguin	no	no	sometime	yes	non-mamma
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mamma
salamander	no	no	sometime	yes	non-mamma
gila monster	no	no	no	yes	non-mamma
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mamma
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mamma

Give Birth	Can Fly	ive in Wate	Have Legs	Class
yes	no	yes	no	?

例

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**

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

估计连续型属性的条件概率

- 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。

Tid	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。

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

Tid	**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|Evade=No) = P(Refund=No| Evade =No)
  P(Married| Evade =No)
  P(Income=No| Evade =No)
  $= 4/7 \times 4/7 \times 4/7 = 0.1866$

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

- Since P(X|No)P(No) > P(X|Yes)P(Yes)

  Therefore P(No|X) > P(Yes|X)
  => Evade = No

用概率分布来估计条件概率

- 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 Ci，属性值x_j属于类Ci的概率为

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

  - μ_{ij}和σ_{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：$(\mu, \sigma^2) = (90, 5^2)$
  - $P(Income = 120|No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

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

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

Since $P(X|No)P(No) > P(X|Yes)P(Yes)$

Therefore $P(No|X) > P(Yes|X)$
=> Class = 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(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
$\boxed{\text{P(Refund=Yes|Yes) = 0}}$
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
$\boxed{\text{P(Marital Status=Married|Yes) = 0}}$

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$\frac{\boxed{n_a}+mp}{\boxed{n_c}+m}$
    - 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(Married | Class=Yes)=0
- 使用m-估计，m=3，p=1/3
- P(Married | Class=Yes)=(0+ 3× 1/3)/(3+3)= 1/6
- P(X | Class=No) = P(Refund=No | Class=No)P(Married | Class=No)P(Income=120K | Class=No) =6/10 × 6/10 ×0.0072= 0.0026
- P(X | Class=Yes) = P(Refund=No | Class=Yes) P(Married | Class=Yes) P(Income=120K | Class=Yes) =4/6 × 1/6 ×1.2 × 10^(-9)= 1.3 × 10^(-10)

- 小结
  - 朴素贝叶斯分类器假定了属性a1,…,an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

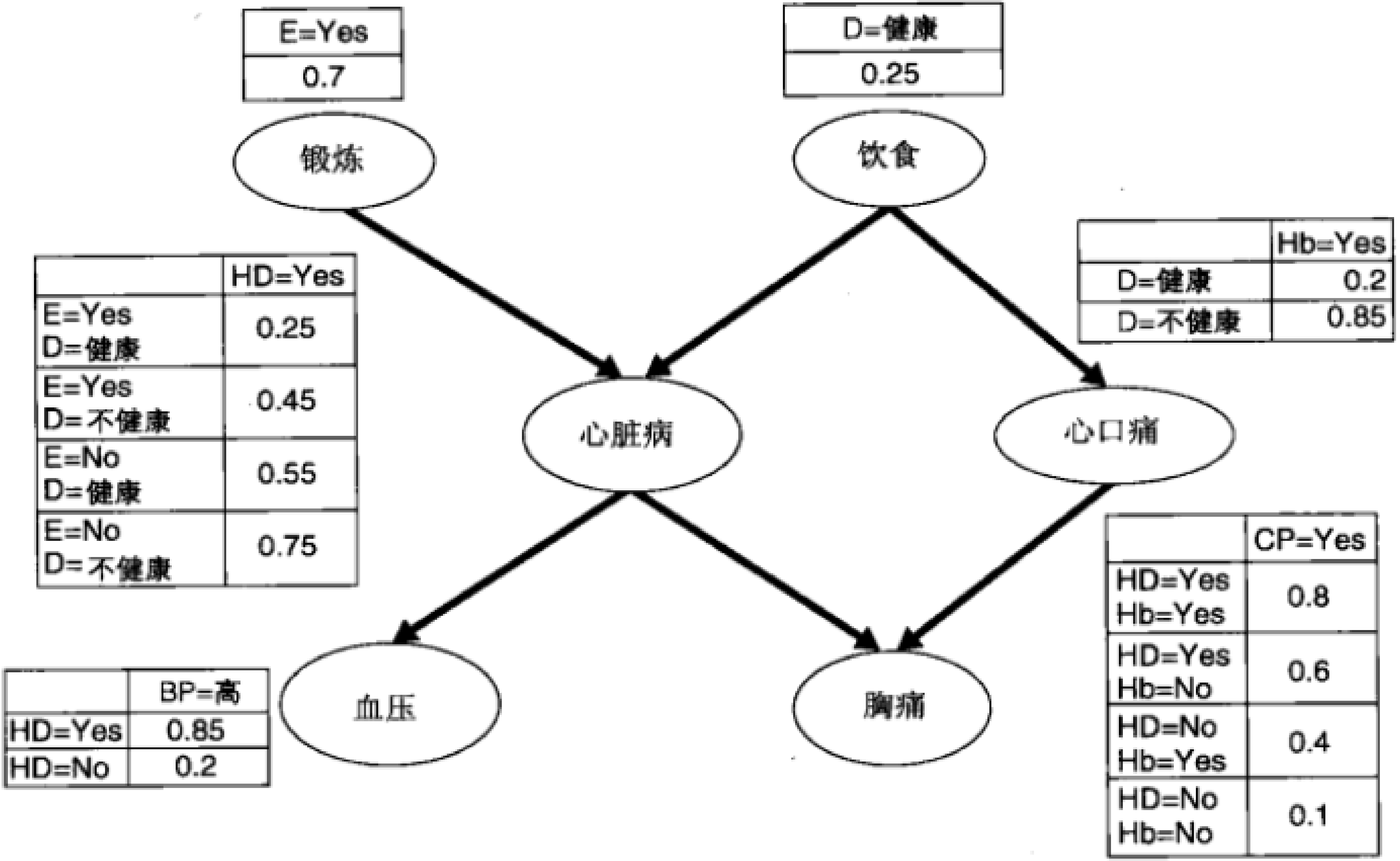
- Python编程实现

贝叶斯信念网络（BAYES BELIEF NETWORKS, BBN）

- 简称贝叶斯网络，<u>表述变量的一个子集上的条件独立性假定</u>。
- 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{Y1,…,YK}，则表中包含条件概率P(X| Y1,…,YK)



E=Yes
0.7

D=健康
0.25

	HD=Yes
E=Yes D=健康	0.25
E=Yes D=不健康	0.45
E=No D=健康	0.55
E=No D=不健康	0.75

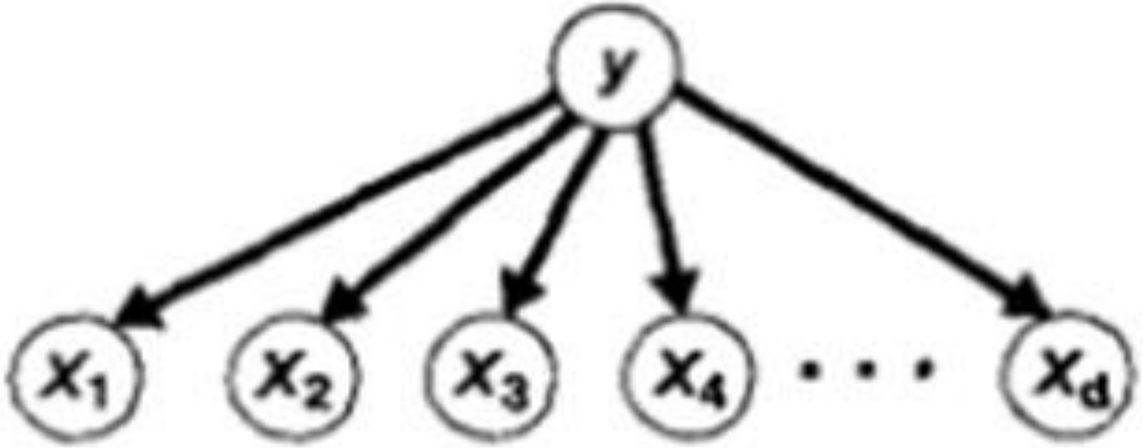
	Hb=Yes
D=健康	0.2
D=不健康	0.85

	BP=高
HD=Yes	0.85
HD=No	0.2

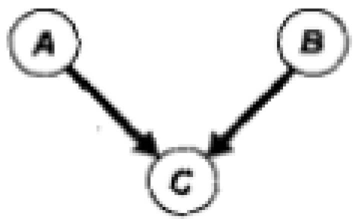
	CP=Yes
HD=Yes Hb=Yes	0.8
HD=Yes Hb=No	0.6
HD=No Hb=Yes	0.4
HD=No Hb=No	0.1

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。

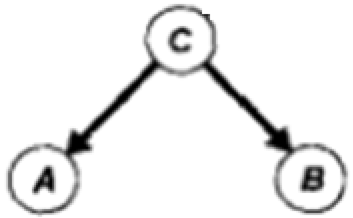


贝叶斯网络中三个变量间的典型依赖关系



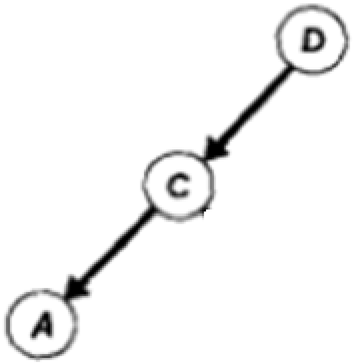
(a)

V型结构（V-Structure）、冲撞结构



(b)

“同父”结构（CommonParent）



(c)

顺序结构

节点

$B = \langle G, \Theta \rangle$

$\boxed{\theta_{x_i|\pi_i} = P_B(x_i \mid \pi_i)}$

给定父结点集，贝叶斯网假设每个属性与它的非后裔属性独立，于是 $B = \langle G, \Theta \rangle$ 将属性 $x_1, x_2, \ldots, x_d$ 的联合概率分布定义为

$$P_B(x_1, x_2, \ldots, x_d) = \prod_{i=1}^{d} P_B(x_i \mid \pi_i) = \prod_{i=1}^{d} \theta_{x_i|\pi_i} .$$

随机变量，XI父节点的集合

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

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

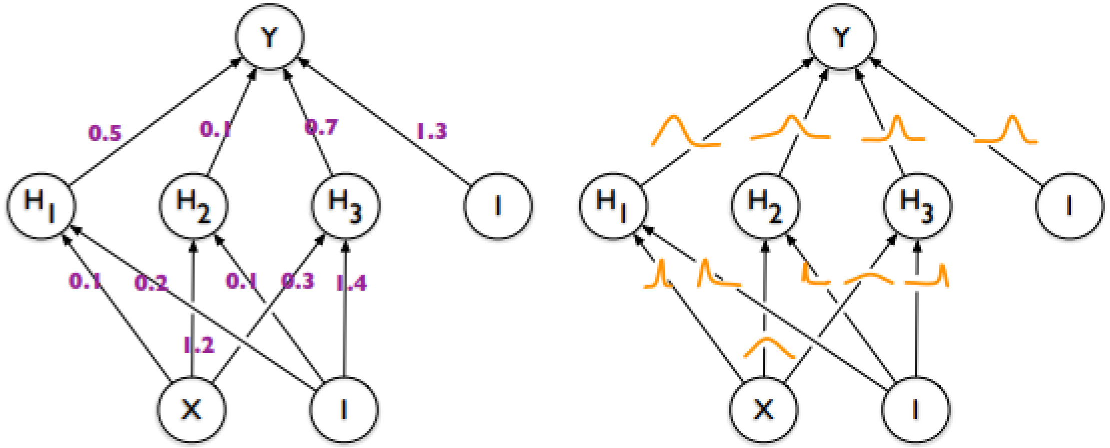


Figure 1. Left: each weight has a fixed value, as provided by classical backpropagation. Right: each weight is assigned a distribution, as provided by Bayes by Backprop.

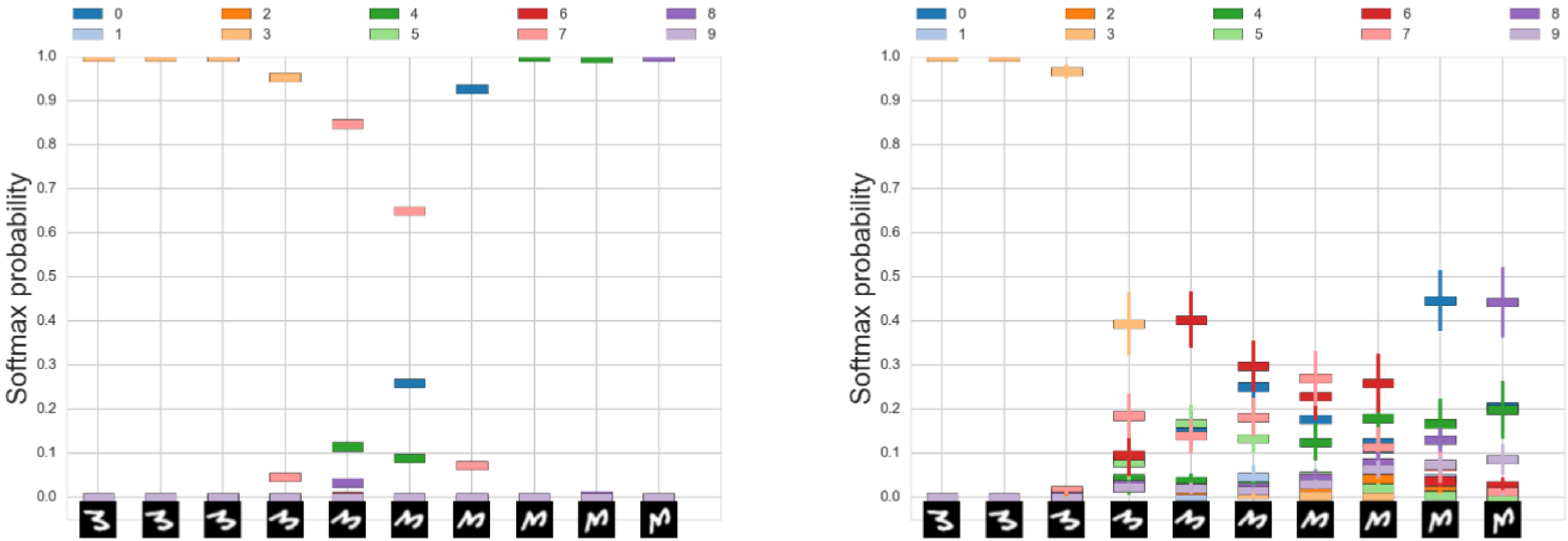
为什么需要贝叶斯神经网络

- 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

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



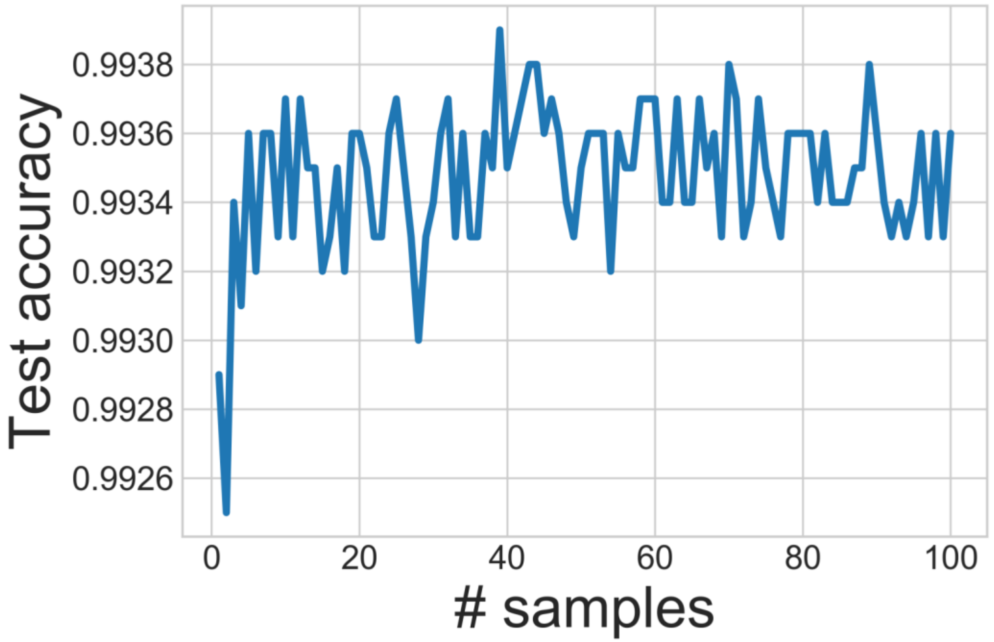
(a) LeNet with weight decay

(b) LeNet with multiplicative formalizing flows

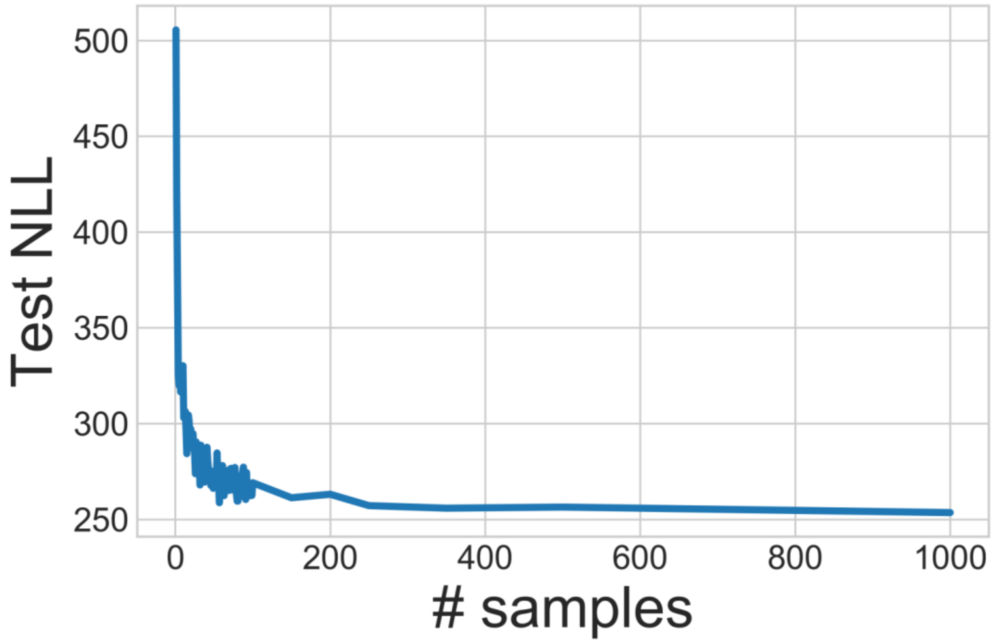
Louizos, Christos, and Max Welling. "Multiplicative normalizing flows for variational bayesian neural networks." ICML 2017

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



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

- θ的求导过程：

$$\theta^* = argmin_\theta D_{KL}[q(w|\theta)||P(w|D)]$$

$$= argmin_\theta \int q(w|\theta) log \frac{q(w|\theta)}{P(w|D)} dw$$

注意到 $P(w|D) = \frac{P(D|w)P(w)}{P(D)}$

$$= argmin_\theta \int q(w|\theta) log \frac{q(w|\theta)P(D)}{P(w)P(D|w)} dw$$

在此处键入公式。

$$= argmin_\theta \int q(w|\theta) log \frac{q(w|\theta)}{P(w)} dw - \int q(w|\theta) log P(D|w) dw + \int q(w|\theta) log P(D) dw$$

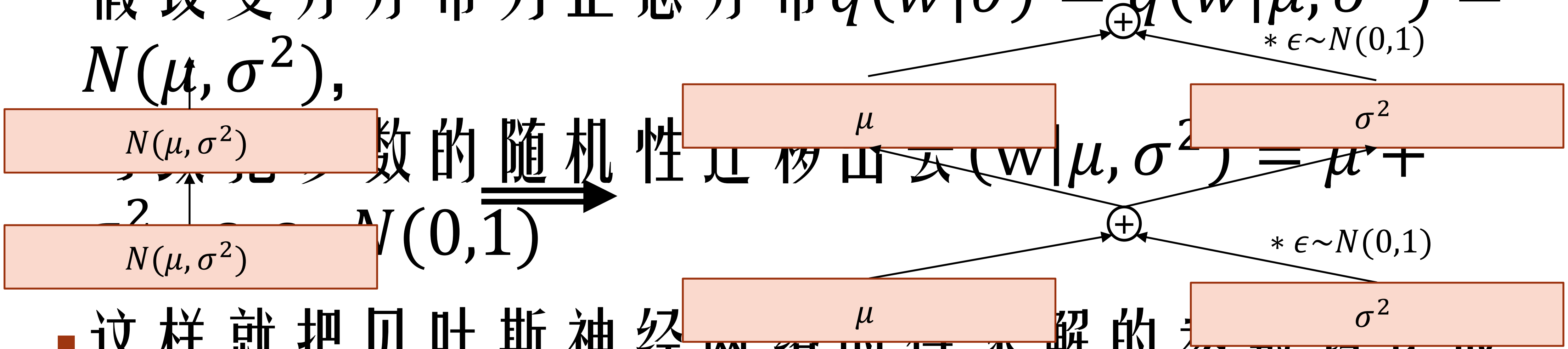
$$= argmin_\theta D_{KL}[q(w|\theta)||P(w)] - E_{q(w|\theta)}[logP(D|w)] + \underbrace{logP(D)}_{\text{常数}}$$

- 最终，网络的优化目标为：

$$F(D,\theta) = \underbrace{-E_{q(w|\theta)}[logP(D|w)]}_{\text{负对数似然，理解为一般损失函数}} + \underbrace{D_{KL}[q(w|\theta)||P(w)]}_{\text{正则化项}}$$

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$q(w|\theta)=q(w|\mu,\sigma^2)=N(\mu,\sigma^2)$，
  …数的随机性迁移出去$(w|\mu,\sigma^2)=\mu+\sigma^2 * \epsilon$，$\epsilon\sim N(0,1)$



- 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

THE END!