模式识别课程

贝叶斯分类器的训练

张俊超

中南大学
航空航天学院

贝叶斯分类器-训练

- 最小错误率贝叶斯分类决策

$$\text{若}P(\omega_i\mid\mathbf{x})=\max_{j=1,2,\ldots,c}P(\mathbf{x}|\omega_j)P(\omega_j)\text{，则}\mathbf{x}\in\omega_i$$

- 最小风险贝叶斯分类决策

$$R(\alpha_i\mid\mathbf{x})=\sum_{j=1}^{c}\lambda(\alpha_i,\omega_j)P(\omega_j\mid\mathbf{x}),i=1,2,\ldots,k$$

$$\text{若}R(\alpha_i\mid\mathbf{x})=\min_{j=1,2,\ldots,k}R(\alpha_j\mid\mathbf{x})\text{，则}\alpha=\alpha_i$$

类条件概率密度和先验概率需要事先估算。

$P(\mathbf{x}|\omega_j)$ $P(\omega_j)$

贝叶斯分类器-训练

假设类条件密度为高斯分布

$$g_i(\mathbf{x}) = P(\mathbf{x}|\omega_i)P(\omega_i)$$

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

取对数，仍用g表示

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

贝叶斯分类器-训练

- 先验概率的估计

先验概率 $P(\omega_j)$

$\hat{P}(\omega_j)=\frac{n_i}{N}$ n_i 为类别 ω_i 的样本数，N 为总样本数

$\hat{P}(\omega_j)=\frac{1}{c}$ c 为类别数

某个概率分布

模型可根据研究情况适当的复杂化

贝叶斯分类器-训练

- 类条件概率密度的估计

类条件概率密度$P(x|\omega_j)$

参数估计(最大似然估计、贝叶斯估计)

非参数估计(直方图、近邻估计等)

不实用

近邻分类器

贝叶斯分类器-最大似然估计

最大似然估计的基本假设：

- 待估计的参数记作θ(单变量)，$\boldsymbol{\theta}$(多变量)
- 类条件概率密度具有某种确定的函数形式，只是其中的参数$\theta/\boldsymbol{\theta}$未知。
- 每类样本集记作$\aleph_i, i = 1, 2, ..., c$，同类里的样本满足独立同分布条件。
- 不同类别的参数是独立的。(对每一类单独处理)

独立意味着不相关(正态分布)
其他情况，独立一定不相关，但不相关未必独立

为了强调待估计的参数，$p(\omega_i \mid \mathbf{x})$记作$p(\omega_i \mid \mathbf{x}, \theta_i)$或$p(\mathbf{x} | \theta_i)$

贝叶斯分类器-最大似然估计

似然函数：

类别ω_i的样本集为：$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$，在概率密度为$p(\mathbf{x}|\theta)$获得该样本集的概率为：

$$l(\theta) = p(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N | \theta) = \prod_{i=1}^{N} p(\mathbf{x}_i | \theta)$$

反映的是在不同参数取值下取得当前样本集的可能性

参数θ相对于样本集$\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$的似然函数。

对数似然函数：

$$\boxed{H(\theta) = \ln l(\theta) = \ln\left(\prod_{i=1}^{N} p(\mathbf{x}_i | \theta)\right) = \sum_{i=1}^{N} \ln\left(p(\mathbf{x}_i | \theta)\right)}$$

最大似然估计是假设在估计值θ等于真值θ时，似然函数取得最大值，即抽取到样本集的概率最大。

贝叶斯分类器-最大似然估计

对似然函数或对数似然函数求梯度并置零：

$$\nabla_{\theta} l(\theta) = 0$$

$$\nabla_{\theta} H(\theta) = \sum_{i=1}^{N} \nabla_{\theta} \ln\left(p\left(\mathbf{x}_i \mid \theta\right)\right) = 0$$

当$\boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_s\right)$是向量时，对$\boldsymbol{\theta}$的每一维分别求偏导

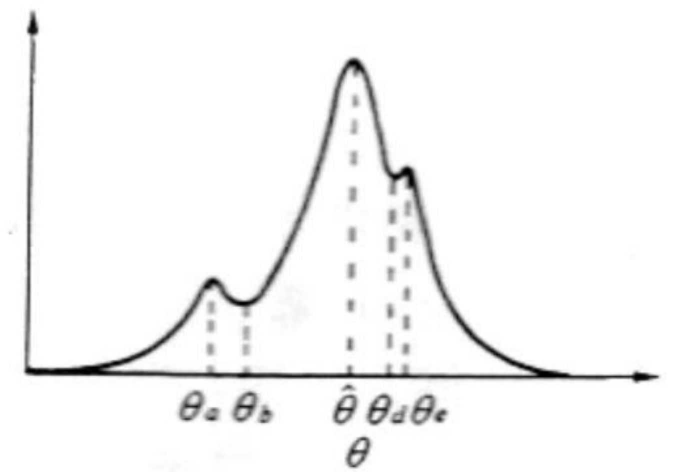
$$\nabla_{\boldsymbol{\theta}} = \left(\frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \cdots, \frac{\partial}{\partial \theta_s}\right)$$

贝叶斯分类器-最大似然估计

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_s)$$

$$\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) = \sum_{i=1}^{N} \nabla_{\boldsymbol{\theta}} \ln\left(p(\mathbf{x}_i \mid \boldsymbol{\theta})\right) = \mathbf{0}$$

$$\begin{cases} \sum_{i=1}^{N} \frac{\partial}{\partial \theta_1} \ln\left(p(\mathbf{x}_i \mid \boldsymbol{\theta})\right) = 0 \\ \vdots \\ \sum_{i=1}^{N} \frac{\partial}{\partial \theta_s} \ln\left(p(\mathbf{x}_i \mid \boldsymbol{\theta})\right) = 0 \end{cases}$$



求得的满足方程的参数估计值有可能有多个，有的是局部最优解，需要寻找到全局最优解！

贝叶斯分类器-最大似然估计

正态分布下的最大似然估计

当$d=1$，即一维情况时，只有两个未知参数：$\boldsymbol{\theta}=\left(\mu,\sigma^2\right)^T$

$$p\left(x\,|\,\boldsymbol{\theta}\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

$$H\left(\boldsymbol{\theta}\right)=\ln\left(p\left(x\,|\,\boldsymbol{\theta}\right)\right)=-\frac{1}{2}\ln\left(2\pi\sigma^2\right)-\frac{1}{2\sigma^2}\left(x-\mu\right)^2$$

$$\nabla_{\boldsymbol{\theta}}H\left(\boldsymbol{\theta}\right)=\sum_{i=1}^{N}\nabla_{\boldsymbol{\theta}}\ln\left(p\left(x_i\,|\,\boldsymbol{\theta}\right)\right)=0$$

$$\begin{cases}\dfrac{\partial H\left(\boldsymbol{\theta}\right)}{\partial\mu}=\sum_{i=1}^{N}\dfrac{1}{\sigma^2}\left(x_i-\mu\right)=0\\[2ex]\dfrac{\partial H\left(\boldsymbol{\theta}\right)}{\partial\sigma^2}=\sum_{i=1}^{N}\left(-\dfrac{1}{2\sigma^2}+\dfrac{\left(x-\mu\right)^2}{2\sigma^4}\right)=0\end{cases}\Rightarrow\begin{cases}\mu=\dfrac{1}{N}\sum_{i=1}^{N}x_i\\[2ex]\sigma^2=\dfrac{1}{N}\sum_{i=1}^{N}\left(x_i-\mu\right)^2\end{cases}$$

$\boxed{\text{有偏估计}}$

$$\mathrm{H}(\theta)=\ln\left(\prod_{i=1}^{N}p(\mathbf{x}_i|\boldsymbol{\theta})\right)=\sum_{i=1}^{N}\ln\left(\frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right\}\right)$$

$$=-\frac{dN}{2}\ln(2\pi)-\frac{N}{2}\ln(|\boldsymbol{\Sigma}|)-\frac{1}{2}\sum_{i=1}^{N}\left((\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right)$$

$$\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{A}\mathbf{x})=(\mathbf{A}+\mathbf{A}^T)\mathbf{x}$$

$$\nabla_{\mathbf{x}}(\mathbf{y}^T\mathbf{x})=\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y})=\mathbf{y}$$

$$\frac{\partial|\mathbf{A}|}{\partial\mathbf{A}}=|\mathbf{A}|(\mathbf{A}^{-1})^T=|\mathbf{A}|(\mathbf{A}^T)^{-1}$$

$$\boxed{\frac{\partial(\mathbf{x}^T\mathbf{A}^{-1}\mathbf{x})}{\partial\mathbf{A}}=\text{-}\mathbf{x}\mathbf{x}^T\mathbf{A}^{-1}\mathbf{A}^{-1}}$$

$$\begin{cases}\boldsymbol{\mu}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\mathbf{x}_i\\[2ex]\boldsymbol{\Sigma}=\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}\left(\mathbf{x}_i-\boldsymbol{\mu}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}\right)^T\end{cases}$$

$$\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} = \mathbf{I}$$

$$\frac{\partial \boldsymbol{\Sigma}}{\partial t}\left(\boldsymbol{\Sigma}^{-1}\right) + \boldsymbol{\Sigma}\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial t} = 0$$

$$\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial t} = -\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial t}\left(\boldsymbol{\Sigma}^{-1}\right)$$

$$\Downarrow$$

$$\frac{\partial \boldsymbol{\Sigma}^{-1}}{\partial \boldsymbol{\Sigma}} = -\boldsymbol{\Sigma}^{-1}\frac{\partial \boldsymbol{\Sigma}}{\partial \boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}^{-1}\right) = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}$$

- 推导过程

将 $(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})$ 记为 $\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$

基于矩阵迹的矩阵求导公式

$$tr(a)=a$$

$$tr(A)=tr(A^T)$$

$$tr(AB)=tr(BA)$$

$$tr(ABC)=tr(CAB)=tr(BCA)$$

$$\boxed{\frac{\partial tr(AB)}{\partial A}=B^T}$$

$$\frac{\partial tr(ABA^TC)}{\partial A}=CAB+C^TAB^T$$

$$\begin{aligned}\frac{\partial(\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y})}{\partial\boldsymbol{\Sigma}}&=\frac{\partial(\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y})}{\partial\boldsymbol{\Sigma}^{-1}}\cdot\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\boldsymbol{\Sigma}}\\&=\frac{\partial\left(tr(\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\mathbf{y})\right)}{\partial\boldsymbol{\Sigma}^{-1}}\cdot\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\boldsymbol{\Sigma}}\\&=\boxed{\frac{\partial\left(tr(\boldsymbol{\Sigma}^{-1}\mathbf{y}\mathbf{y}^T)\right)}{\partial\boldsymbol{\Sigma}^{-1}}}\cdot\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\boldsymbol{\Sigma}}\\&=\boxed{(\mathbf{y}\mathbf{y}^T)^T}\cdot\frac{\partial\boldsymbol{\Sigma}^{-1}}{\partial\boldsymbol{\Sigma}}\\&=\mathbf{y}\mathbf{y}^T(-\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1})\\&=-\mathbf{y}\mathbf{y}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\end{aligned}$$

贝叶斯分类器-贝叶斯估计

把待估计的参数θ看作具有先验分布密度$p(\theta)$的随机变量

\theta 的先验信息

在样本集$\aleph=\{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$下的条件风险为：

$$R\left(\hat{\theta}\,|\,\aleph\right)=\int_{\Theta}\lambda\left(\hat{\theta},\theta\right)p(\theta\,|\,\aleph)\,d\theta$$

条件风险最小化：$\theta^*=\arg\min_{\hat{\theta}}R\left(\hat{\theta}\,|\,\aleph\right)=\int_{\Theta}\lambda\left(\hat{\theta},\theta\right)p(\theta\,|\,\aleph)\,d\theta$

若采用平方误差损失：$\lambda\left(\hat{\theta},\theta\right)=\left(\hat{\theta}-\theta\right)^2$

则，$\theta^*=E[\theta\,|\,\aleph]=\int_{\Theta}\theta p(\theta\,|\,\aleph)\,d\theta$

贝叶斯分类器-贝叶斯估计

最小平方误差损失函数下，贝叶斯估计的步骤：

1. 确定θ的先验分布密度$p(\theta)$
2. 基于样本是独立同分布的，样本集的联合分布为：$p(\aleph|\theta)=\prod_{i=1}^{N}p(x_i|\theta)$
3. 利用贝叶斯公式，求θ的后验概率分布：$p(\theta|\aleph)=\dfrac{p(\aleph|\theta)p(\theta)}{\int_{\Theta}p(\aleph|\theta)p(\theta)d\theta}$
4. θ的贝叶斯估计量是：$\theta^*=\int_{\Theta}\theta p(\theta|\aleph)d\theta$

贝叶斯分类器-贝叶斯估计

正态分布时的贝叶斯估计：

Example

当$d=1$，即一维情况时，μ为未知参数，方差为σ^2

$$p(x|\mu)=\frac{1}{\sqrt{2\pi\sigma}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

假设μ的先验分布也是正态分布：$p(\mu)=\frac{1}{\sqrt{2\pi\sigma_0}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2\right)$

$$\boxed{p(\mu|\aleph)=\frac{1}{\sqrt{2\pi\sigma_N}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_N}{\sigma_N}\right)^2\right)}$$

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0,\quad m_N=\frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

贝叶斯分类器-贝叶斯估计

$$p(\mu|\aleph)=\frac{1}{\sqrt{2\pi}\sigma_N}\exp\left(-\frac{1}{2}\left(\frac{x-\mu_N}{\sigma_N}\right)^2\right)$$

$$\mu_N=\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}m_N+\frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0,\quad m_N=\frac{1}{N}\sum_{i=1}^{N}x_i$$

$$\sigma_N^2=\frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

μ的贝叶斯估计为：$\mu^*=\mu_N$

- 当样本数趋于无穷大时，第一项系数趋于1，第二项系数趋于0，估计的均值就是样本的算术平均，与最大似然估计一致。
- 当样本数有限，先验分布方差σ_0^2很小，第一项系数就很小，第二项系数接近于1，估计主要由先验知识来决定。

贝叶斯分类器-贝叶斯估计

举例说明：

抛硬币实验，记正面朝上的概率记为θ。且假设抛硬币的模型是二项分布。

我们抛 10 次，得到的数据x是“反正正正正反正正正反”，我们想求正面朝上的概率θ。

则似然函数为：

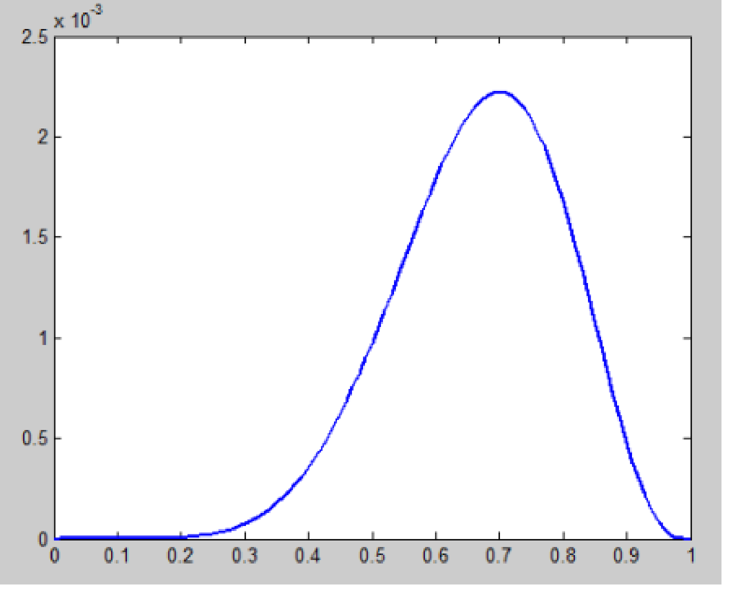
$$\boxed{P(x|\theta)=(1-\theta)\times\theta^4\times(1-\theta)\times\theta^3\times(1-\theta)=\theta^7(1-\theta)^3}$$

贝叶斯分类器-贝叶斯估计

则似然函数为：

$$P(x|\theta)=(1-\theta)\times\theta^4\times(1-\theta)\times\theta^3\times(1-\theta)=\theta^7(1-\theta)^3$$

即 $\theta=0.7$ 时，似然函数取得最大值。



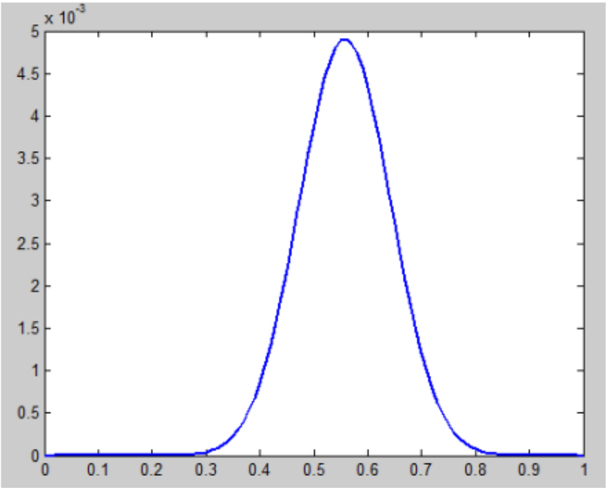
贝叶斯分类器-贝叶斯估计

贝叶斯学派

我们先验地知道$\theta=0.5$的概率很大，取其他值的概率小一些。我们用一个高斯分布来描述这个先验信息：$\mu=0.5,\sigma=0.1$

$$P(\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(\theta-\mu)^2}{2\sigma^2}}$$

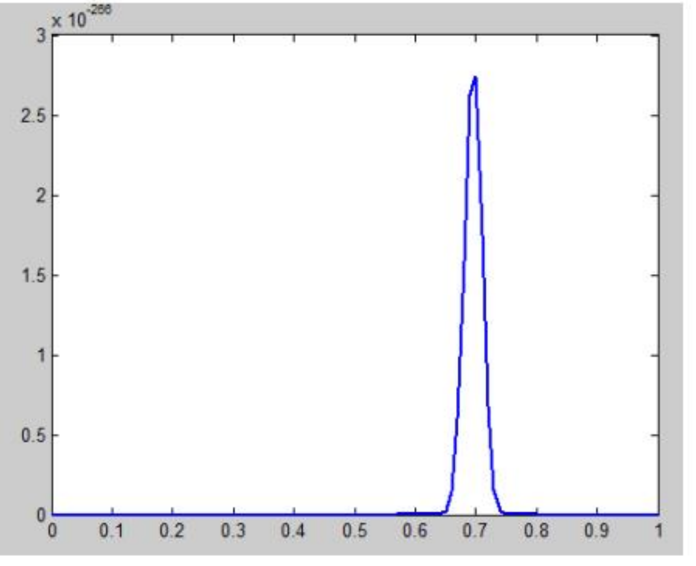
则$P(x|\theta)P(\theta)$的图像为：



在$\theta=0.56$时取得最大值。

贝叶斯分类器-贝叶斯估计

那么，怎么让贝叶斯派相信$\theta=0.7$呢？多做点实验，做 1000 次实验，700 次都是正面朝上，这时$P(x|\theta)P(\theta)$的图像为：



此时，$\theta=0.7$，就算一个考虑了先验概率的贝叶斯派，也不得不承认得把θ估计在 0.7 附近了。

贝叶斯分类器-贝叶斯估计

- 最大似然估计和贝叶斯估计的区别：

➢最大似然估计：把待估计的参数当作未知固定的量，根据观测数据估计这个量；

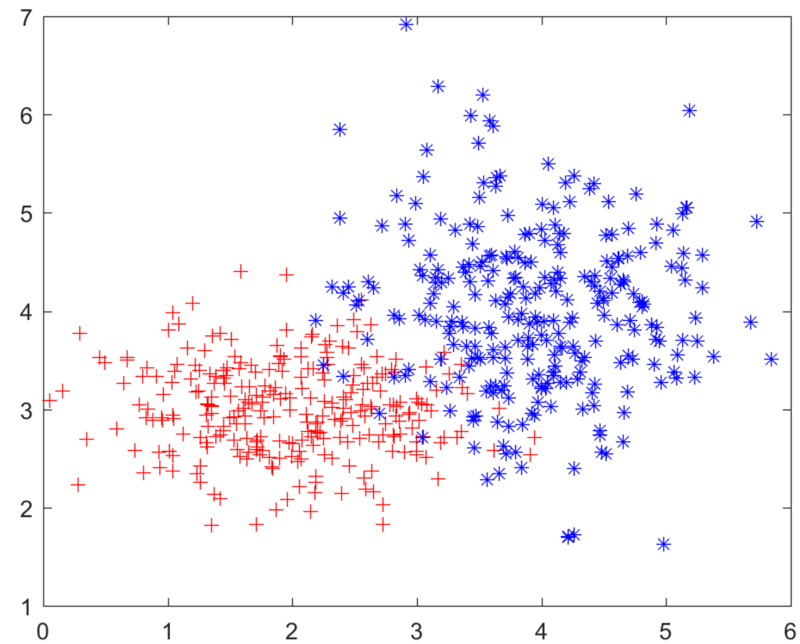
➢贝叶斯估计：把待估计的参数本身也看作是随机变量，根据观测数据对参数的分布进行估计。

模式识别-贝叶斯分类器

【举例】最大似然估计

```
clc;
close all;
clear all;
%% 最大似然估计
% 生成数据
randn('seed',2020);
mu1 = [2 3];
sigma1 = [0.5 0;
          0 0.2];
data1 = mvnrnd(mu1,sigma1,300);

randn('seed',2021);
mu2 = [4 4];
sigma2 = [0.5 0;
          0 0.8];
data2 = mvnrnd(mu2,sigma2,300);
```



模式识别-贝叶斯分类器

```
%%
N1 = size(data1,1);
N2 = size(data2,1);
Test_Num = 20;
Training_Num1 = N1 - Test_Num;
Training_Num2 = N2 - Test_Num;
```

```
%% 训练
mu1_hat = mean(data1(1:Training_Num1,:),1);
mu2_hat = mean(data2(1:Training_Num2,:),1);

Tmp1 = data1(1:Training_Num1,:)-repmat(mu1_hat,[Training_Num1,1]);
Tmp2 = data2(1:Training_Num2,:)-repmat(mu2_hat,[Training_Num2,1]);
sigma1_hat = Tmp1'*Tmp1/Training_Num1;
sigma2_hat =Tmp2'*Tmp2/Training_Num2;
```

模式识别-贝叶斯分类器

```
K>> mu1_hat

mu1_hat =

    1.9919    3.0047

K>> sigma1_hat

sigma1_hat =

    0.5543   -0.0249
   -0.0249    0.2128
```

```
K>> mu2_hat

mu2_hat =

    3.9366    4.0085

K>> sigma2_hat

sigma2_hat =

    0.5074   -0.0423
   -0.0423    0.6971
```

模式识别-贝叶斯分类器

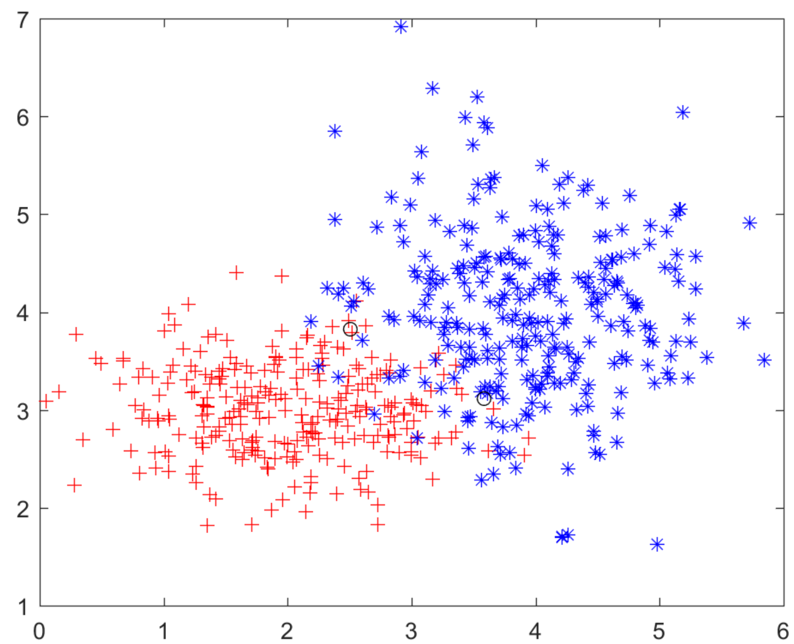
```
%% 预测
Data = data1(N1-Test_Num+1:end,:);
P_1_1 = Predict(Data,sigma1_hat,mu1_hat);
P_1_2 = Predict(Data,sigma2_hat,mu2_hat);
P_1 = [P_1_1,P_1_2];
[~,ind1] = max(P_1,[],2);
C_N1 = sum(ind1==ones(size(ind1)));%% 分类正确的数目


Data = data2(N2-Test_Num+1:end,:);
P_2_1 = Predict(Data,sigma1_hat,mu1_hat);
P_2_2 = Predict(Data,sigma2_hat,mu2_hat);
P_2 = [P_2_1,P_2_2];
[~,ind2] = max(P_2,[],2);
C_N2 = sum(ind2==1+ones(size(ind1)));%% 分类正确的数目

C_N = C_N1 + C_N2;
ratio = C_N/(Test_Num*2);%% 分类正确率
```

模式识别-贝叶斯分类器

```
%% 显示错误分类的样本
figure(2),plot(data1(1:N1-Test_Num,1),data1(1:N1-Test_Num,2),'r+');hold on;
plot(data2(1:N2-Test_Num,1),data2(1:N2-Test_Num,2),'b*');hold on;
D = data1(N1-Test_Num+1:end,:);
plot(D(ind1==1+ones(size(ind1)),1),D(ind1==1+ones(size(ind1)),2),'ko');hold on;
D = data2(N2-Test_Num+1:end,:);
plot(D(ind2==ones(size(ind2)),1),D(ind2==ones(size(ind2)),2),'ko');hold on;
```



贝叶斯分类器-作业

3 二维空间中的两类样本均服从正态分布，其参数分别为：

均值向量：$\boldsymbol{\mu}_1 = (1,0)^T, \boldsymbol{\mu}_2 = (-1,0)^T$

协方差矩阵：$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$

且两类的先验概率相等，试证明其基于最小错误率判决准则的决策分界面方程为一圆，并求其方程。

模式识别-贝叶斯分类器

【作业】

1. 编程实现正态分布下的5种case(2D情况下)，最大似然估计下的分类结果，基于test数据，计算分类准确率
2. 数据可以自己随机生成
3. 给出case 1- case4的决策面方程，并绘制出来
4. 完成实验报告

案例：字写数字的识别

基于Mnist数据集，请用贝叶斯分类器对手写数字进行识别。

模式识别 -贝叶斯分类器

【要求】

1. 编程语言：Matlab (或Python)
2. 不能使用额外的库函数，自己编写实现算法(采用最大似然估计)。
3. 可以用最小错误类贝叶斯决策或最小风险决策(决策表自己设置)
4. 提交实验报告和源代码(**命名规则：贝叶斯_学号_姓名**)
5. 作业迟交n天，本次作业分数乘以0.98^n。


若干素材取自网络，特此致谢！

谢谢聆听！