第5章 概率密度函数的估计

5.1 引言

5.2 参数估计的基本概念

5.3 最大似然估计与正态分布的参数估计

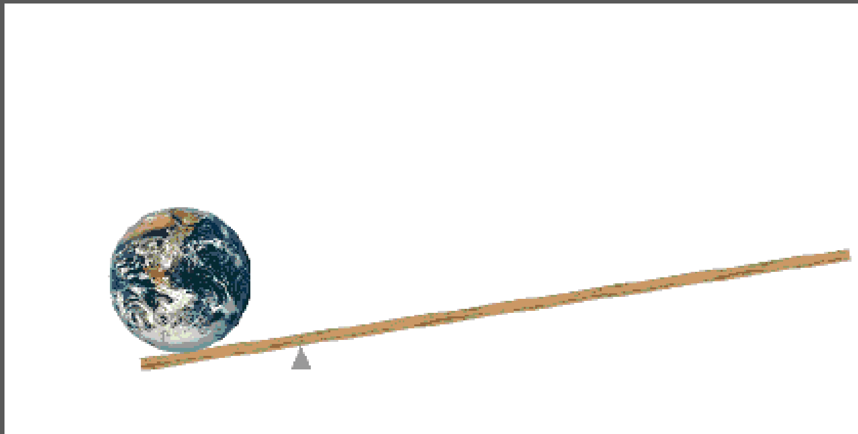
5.4 Bayes估计与正态分布参数的估计

5.1 引言

贝叶斯决策理论

完美（表面上）： ***最低的分类错误率***

标杆



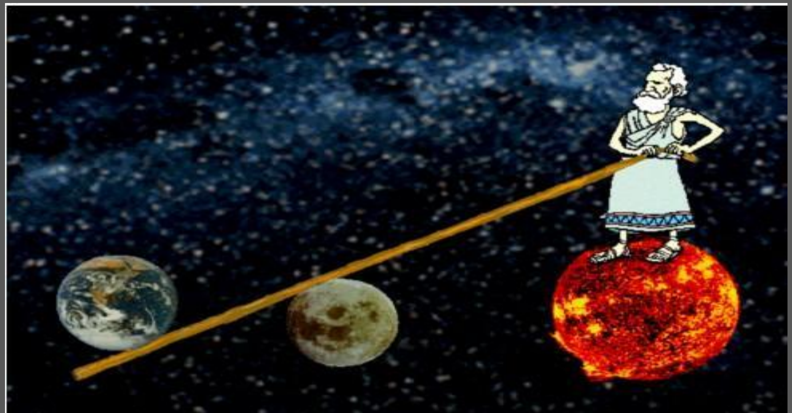
5.1 引言

在贝叶斯决策理论中，基本的已知条件是：

类先验概率 $P(\omega_i)$

类条件概率密度 $p(x|\omega_i)$

疑问： 它们从何而来？

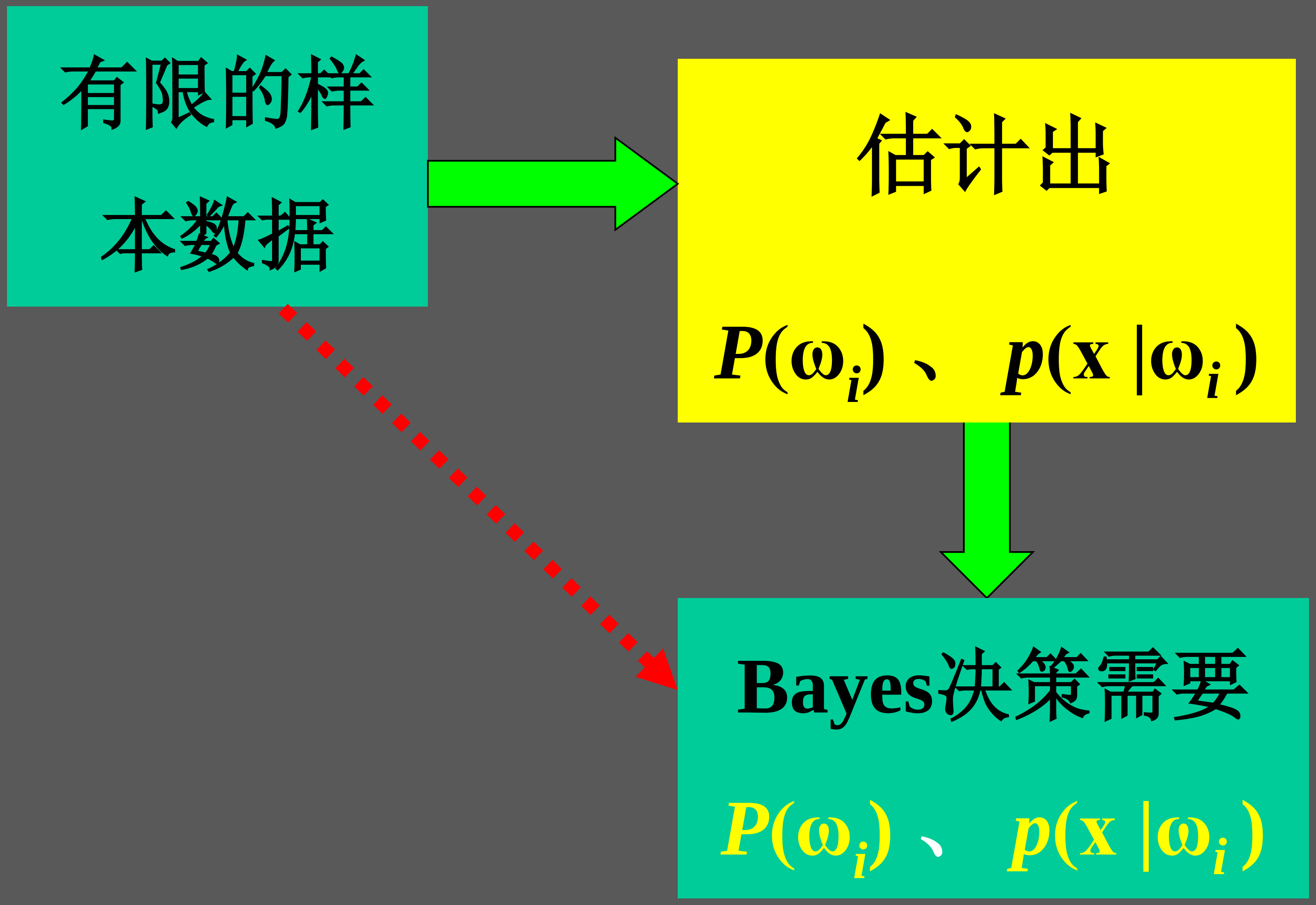


支点在哪儿？

我又该站在哪儿？

面临的实际情况是：

对于一个具体问题，我们只有**有限数目的样本**（所属类别有可能还是未知的）


有限的样本数据
估计出
$P(\omega_i)$、$p(x|\omega_i)$
Bayes决策需要
$P(\omega_i)$、$p(x|\omega_i)$

分类器的设计分成两步来完成：

1 利用样本集估计出$P(\omega_i)$ 、 $p(x\,|\omega_i)$（本章要解决的基本问题）

2 利用Bayes决策理论设计分类器（前一章已经解决的问题）

本章要解决的三个问题

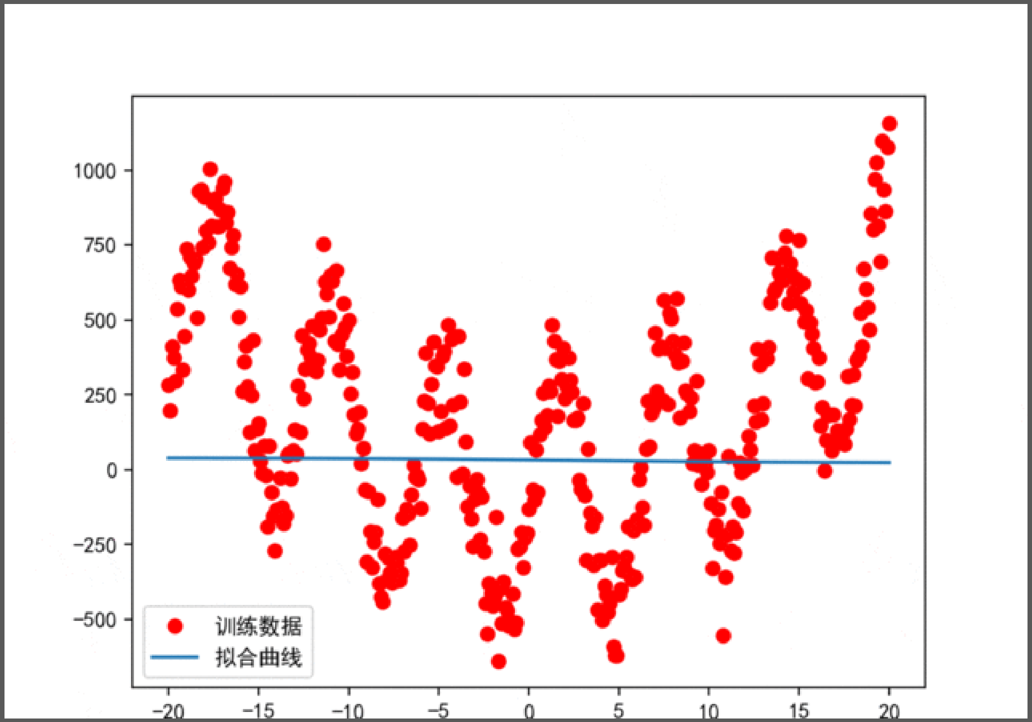
1 如何用样本集估计出$P(\omega_i)$ 、 $p(\mathbf{x}\,|\omega_i)$的估计量 $\hat{P}(\omega_i), \hat{p}(\mathbf{x}\,|\,\omega_i)$

2 评估与分析估计量的性质

从样本集推断总体概率分布的方法

- 估计方法
  - 参数估计
    - 监督参数估计
    - 非监督参数估计
  - 非参数估计

一个例子：曲线拟合



已知： 样本（红色数据点）

需求： 根据数据得出一条曲线

说明：

监督：样本的类别是已知的

非监督：样本的类别是未知的

参数估计：概率密度形式已知（假设其服从某种分布），只需推断出其中的未知参数（确定出分布的参数）

非参数估计：直接推断出概率密度本身

监督参数估计

条件：已知样本所属的**类别**及类条件总体概率密度函数的**形式**，未知概率密度函数的某些**参数**

监督参数估计：从已知类别的样本集，推断（估计）出总体分布（每一类概率密度函数）的某些参数的方法

例如：

（1）假设数据服从正态分布

（2）从样本求正态分布的均值向量与协方差矩阵

非监督参数估计

条件：未知样本所属**类别**，已知总体概率密度函数**形式**，但未知其中的某些**参数**

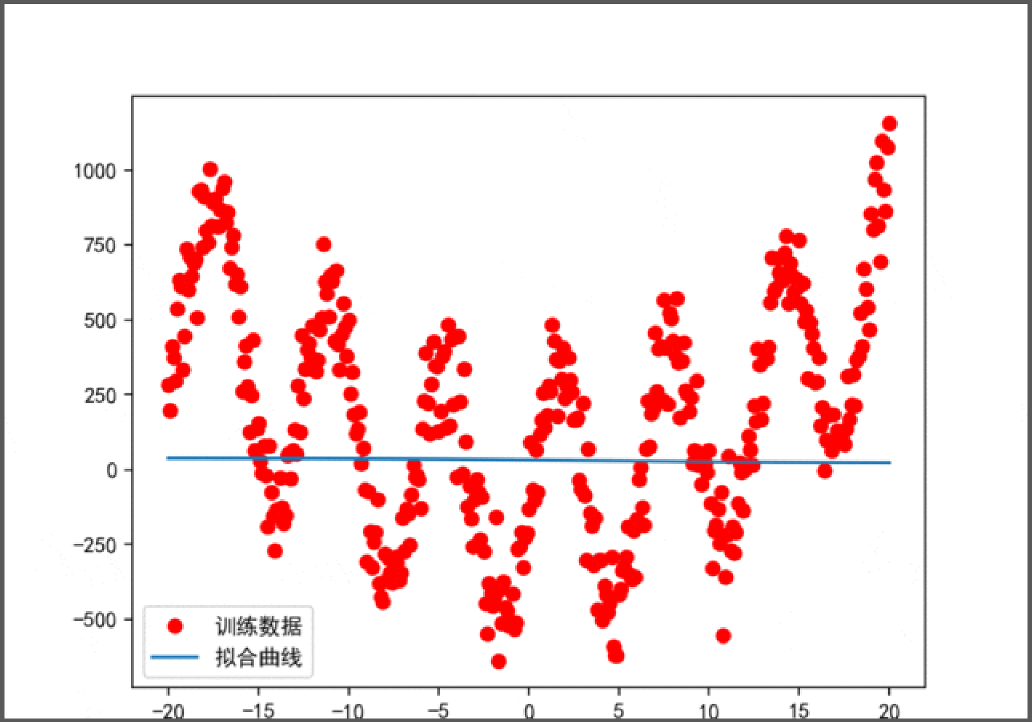
非监督参数估计：推断（估计）出总体概率密度函数中的某些参数的方法

非参数估计

条件：已知样本所属**类别**，但未知总体概率密度函数的**形式**

非参数估计：从已知类别的样本数据中，直接推断出概率密度函数本身

一个例子：曲线拟合

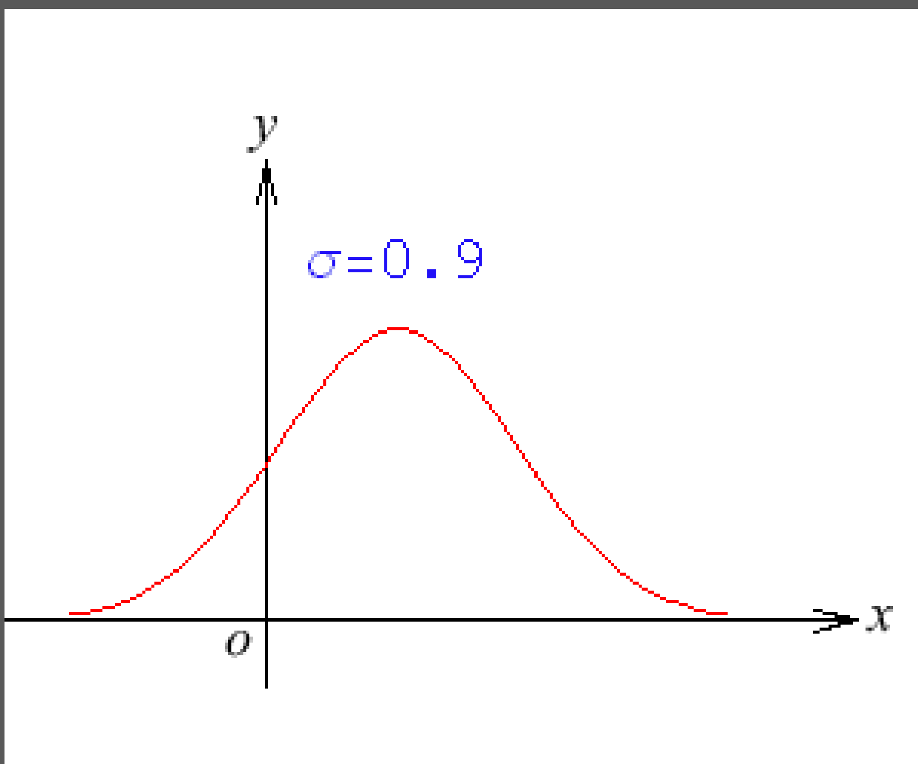


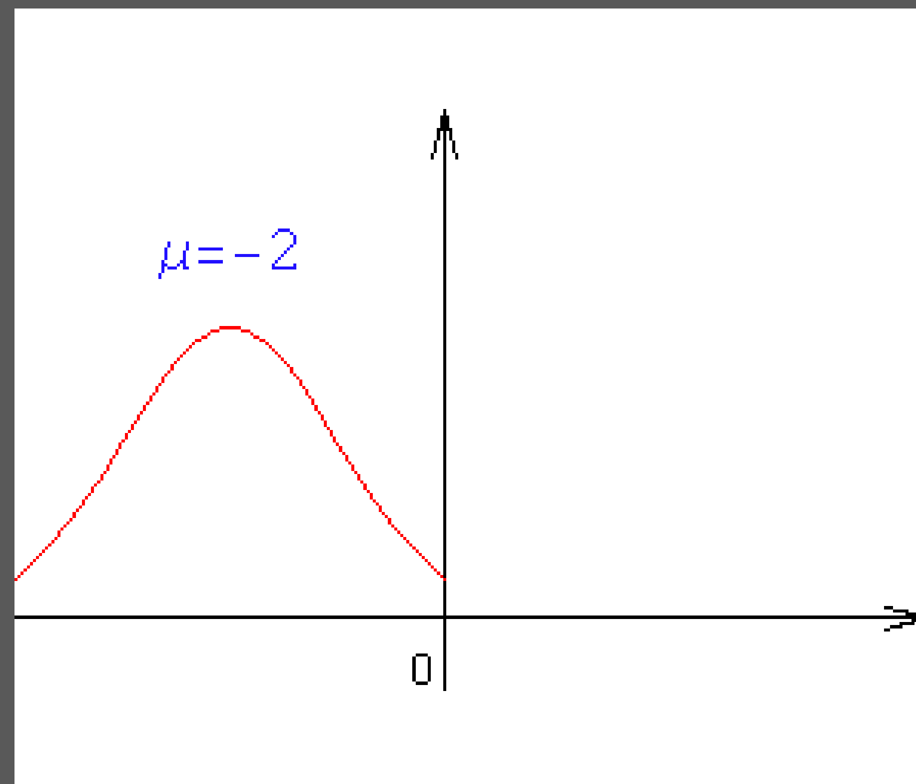
参数估计：*基于分布的假设得出曲线*

非参数估计：**直接得出一条曲线**

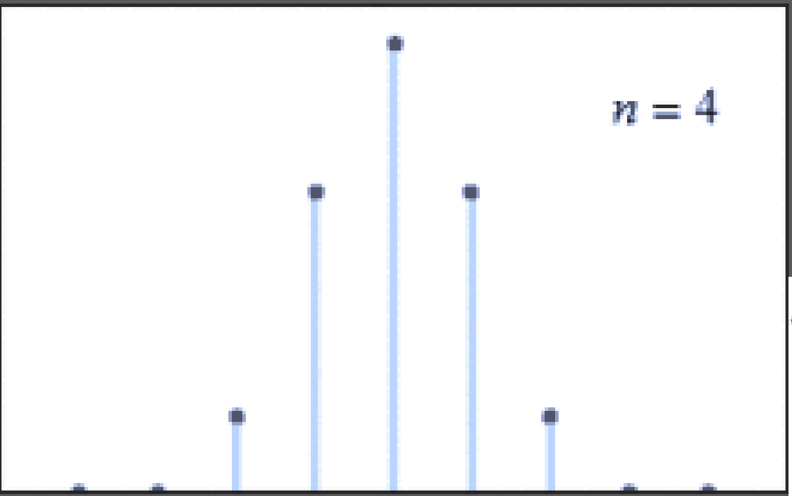
方法	样本类别	数据分布形式	目标
监督参数估计	已知	已知	求分布的参数
非监督参数估计	未知	已知	求分布的参数
非参数估计	已知	未知	求密度函数

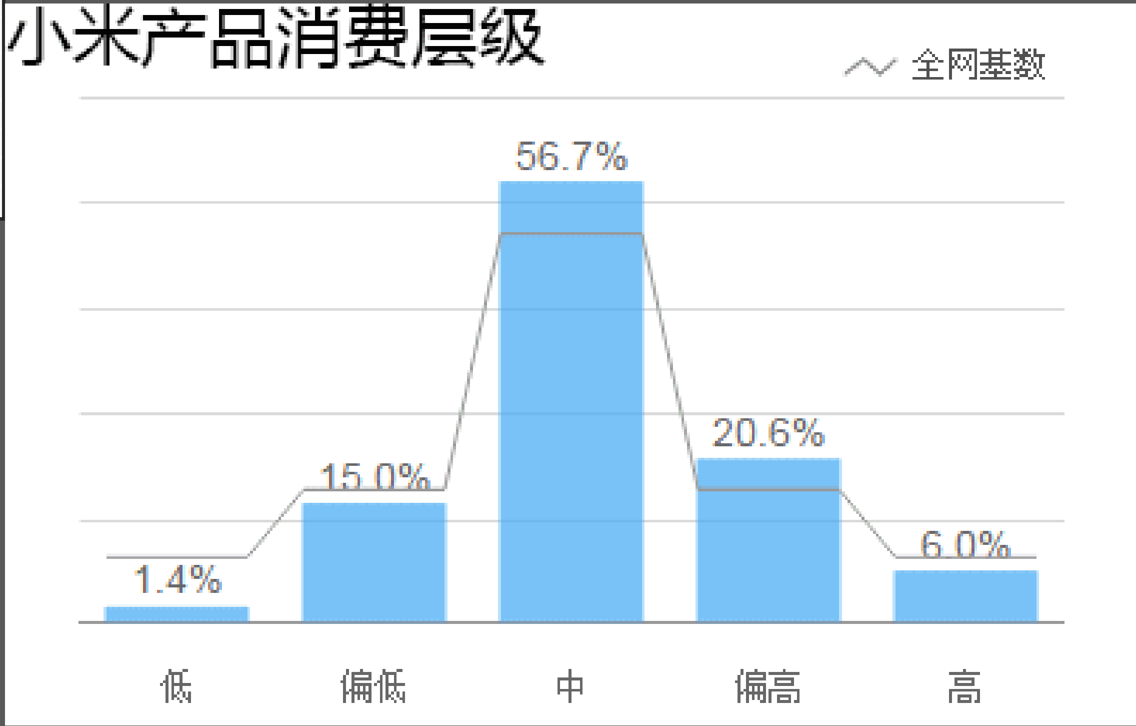
最常见的数据分布形式： 正态分布





最常见的数据分布形式：正态分布





估计方法的数学原理：

参数估计的数学原理：

- 最大似然估计方法与Bayes估计方法

非参数估计的数学原理：

- Parzen窗法与 k_N 近邻法

本章讲解的重点内容：

1 监督参数估计（估计类条件概率密度的参数）

2 非参数估计（估计类条件概率密度本身）

5.2 参数估计的基本概念

1 统计量

2 参数空间

3 点估计、估计量（估计子）、估计值

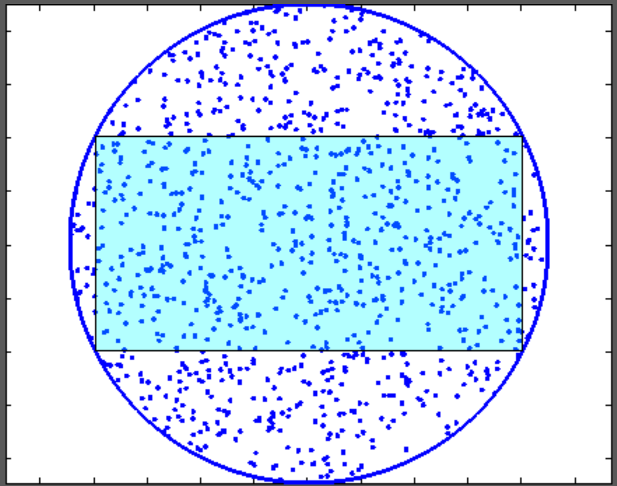
1 统计量

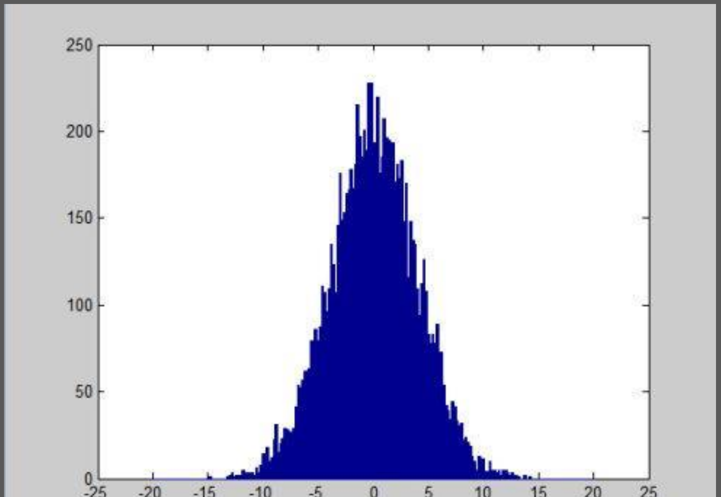
目的：样本中包含着总体的信息，希望有一种数学手段将样本集中的有关信息抽取出来

统计量：针对不同要求构造出的关于样本的某种函数，这种函数在统计学中称为**统计量**

统计

- 个体信息----> 总体信息
- 个体的随机性---> 总体的规律性





5.3 最大似然估计与正态分布的参数估计

本章假设类条件概率密度服从正态分布，

利用最大似然估计得出正态分布的参数 ⇒ $p(\mathbf{x}\mid\omega_i)$

用身高、体重区分男女生的例子

3.3.1 最大似然估计的基本理论

假设条件:

①待估计参数θ 是确定性的未知量

②按类别将样本划分 c 类，第 i 样本都是从类概率密度 $p(x|\omega_i)$ 的总体中独立地抽取出来的

③类条件概率密度 $p(x|\omega_i)$ 的函数形式是确定的，

但是其中的某些参数是未知的

④第 i 类的样本不包含有关 θ_j $(i \neq j)$的信息。不同

类别的参数在函数上相互独立，每一类样本可

以独立进行处理

在满足四个假设条件下，可以将 c 类概率密度估计问题转化为 c 个独立的密度估计问题，分别单独进行处理

记号：$\boldsymbol{\theta}, p(\mathbf{x} \mid \boldsymbol{\theta})$

待求的参数向量

待求的概率密度，与 $\boldsymbol{\theta}$ 有关

最大似然函数估计的思想

观察到的行为（品行）一般会经常出现

最大似然函数估计的思想

假设某一类的样本（数据）服从正态分布

则均值与方差（协方差矩阵）的最优估计应该使似然取得最大值！

在统计学中**似然函数**的定义

N 个随机变量 $x_1,\ldots x_N$ 的似然函数是 N 个随机变量的联合密度

$$l(\theta) = p(x_1, \ldots, x_N \mid \theta)$$

这是 θ 的函数

设某一类样本集有 N 个样本

$$X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

它们是独立地按照概率密度 $p(\mathbf{x} \mid \theta)$ 抽取出来的

（独立同分布样本）

似然函数为

$$
\begin{aligned}
l(\theta) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_N \mid \theta) \\
&= p(\mathbf{x}_1 \mid \theta) \ldots p(\mathbf{x}_N \mid \theta) \\
&= \prod_{k=1}^{N} p(\mathbf{x}_k \mid \theta)
\end{aligned}
$$

含义：从总体中抽取 N 个样本，这 N 个样本正好是 $\mathbf{x}_1, \ldots \mathbf{x}_N$ 的概率（可能性）

最大似然估计的主要思想：如果在一次观察中一个事件出现了，则我们可以认为这一事件出现的可能性很大。现在，事件（$x_1,\ldots x_N$）在一次观察（从概率总体中抽取一组样本）中居然出现了，则我们认为相应的参数θ和$x_1,\ldots x_N$使似然函数 $l(\theta)$ 达到了最大值

身高分布估计的例子

- （1）对从大街上某一段时间出现的成年男性测量身高
- （2）假设成年男性的身高服从正态分布
- （3）大街上出现的成年男性具有随机性（不同时刻不同地点出现的人是不一样的）
- **最大似然估计的思想**：既然能正好观察到这批成年男性，说明这些身高值出现的概率最大，即似然函数 $l(\theta)$对应着最大值，据此可估计出正态分布的参数（即均值与期望）

身高分布估计的例子

- 当然，单个人的身高值对整个男性身高分布的估计意义不大---从一个人的信息推测整个群体的信息是非常片面的
- 观察次数越多（个体信息越多），样本越具有代表性，最大似然估计的结果越准确---因此，适用于样本数较多的情况
- 生活中的例子

最大似然估计量：设 $l(\theta)$ 是样本集

$$X=\{ \mathbf{x}_1, \ldots, \mathbf{x}_N \}$$

的似然函数，如果

$$\hat{\theta} = d(X) = d(\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

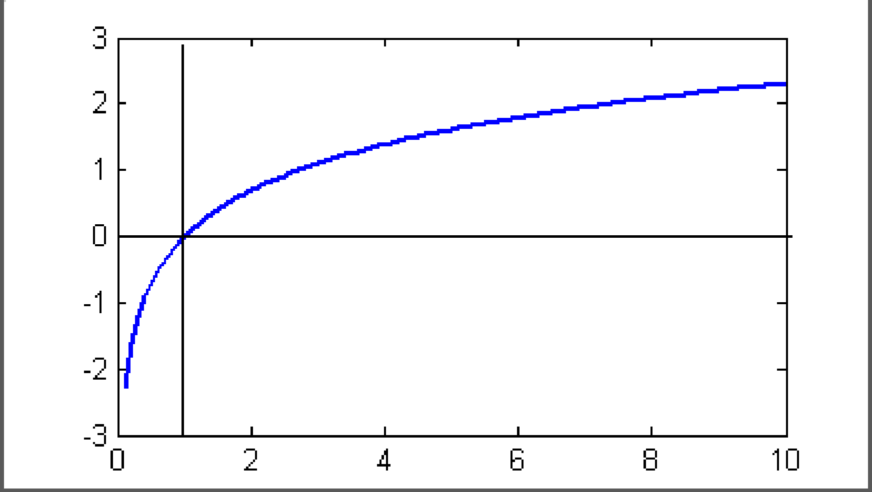
是参数空间 Θ 中使似然函数 $l(\theta)$ 极大化的 θ 值，

则称 $\hat{\theta}$ 是 θ 的最大似然估计量（估计子）

便于分析，可以取**似然函数**的对数，即

$$\boxed{H(\theta) = \ln l(\theta)}$$

对数函数是单调增函数，$H(\theta)$ 与 $l(\theta)$ 的最大点相同



求最大似然估计量的方法

如果$H(\theta)$ 满足一定数学性质（连续可微），可以直接应用高等数学的知识来求**最大点**，即求梯度（**偏导数**），令其等于**零**，解线性或者非线性方程组得到**估计量**

设 $\boldsymbol{\theta} = [\theta_1, ..., \theta_s]^{\mathrm{T}}$

梯度算子 $\nabla_\theta = \begin{bmatrix} \dfrac{\partial}{\partial \theta_1} \\ \ldots \\ \dfrac{\partial}{\partial \theta_s} \end{bmatrix}$

$$l(\boldsymbol{\theta}) = \prod_{k=1}^{N} p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$H(\boldsymbol{\theta}) = \ln l(\boldsymbol{\theta}) = \sum_{k=1}^{N} \ln p(\mathbf{x}_k \mid \boldsymbol{\theta})$$

$$H(\theta) = \ln l(\theta) = \sum_{k=1}^{N} \ln p(\mathbf{x}_k \mid \theta)$$

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \nabla_\theta \ln p(\mathbf{x}_k \mid \theta)$$

$$\nabla_\theta H(\theta) = 0$$

从中求解出 θ 的最大似然估计量

5.3.2 正态分布参数的最大似然估计值

单变量正态分布的概率密度函数

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

要求的未知参数（均值与方差）

$$\theta = [\theta_1, \theta_2]^{\mathrm{T}} = [\mu, \sigma^2]^{\mathrm{T}}$$

最大似然估计的结果：

$$\widehat{\theta}_1 = \widehat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\widehat{\theta}_2 = \widehat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \widehat{\mu})^2$$

最大似然估计结果的说明：

- 中学时，我们就学过：已知一个物理量的多个观察值 $X_1, \ldots, X_N$ ，*其均值和方差计算如下：*

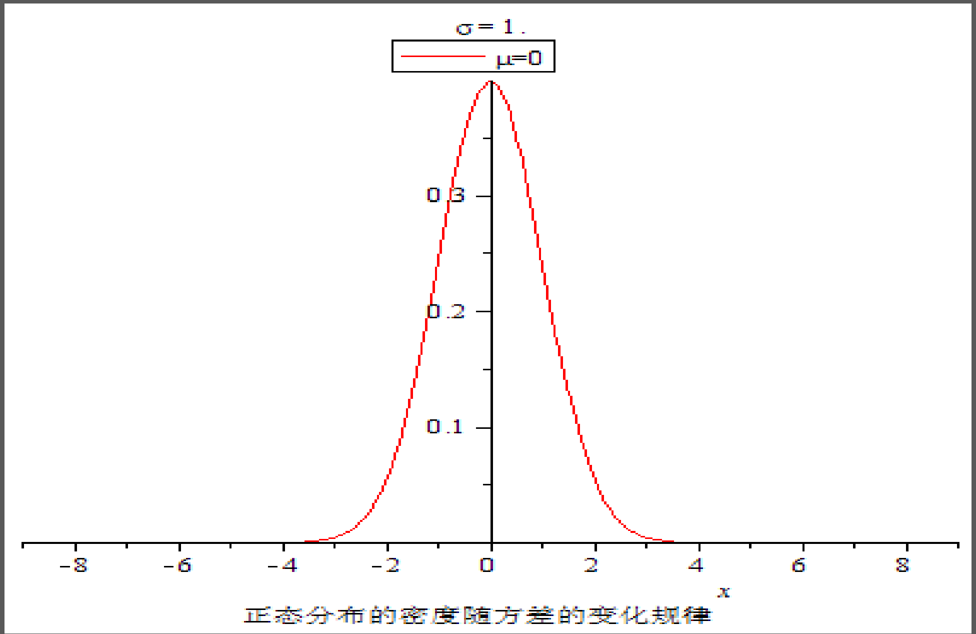
$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu})^2$$

- *本课程中，新的内容：不仅给出物理量的均值和方差，而且还给出了物理量的分布（成年男性身高这一物理量可取得不同的值，但不同值出现的概率不一样，身高在平均值附件的成年男性最多！）*

最大似然估计结果的说明：

- *本课程中，新的内容：不仅给出物理量的均值和方差，而且还给出了物理量的分布（成年男性身高这一物理量可取得不同的值，但不同值出现的概率不一样，身高在平均值附件的成年男性最多！）*



我们已知 N 个一维样本集

$$X = \{x_1, x_2, \ldots, x_N\}$$

问题：利用最大似然估计法，针对上述样本集，求出均值与方差的估计值

$$\widehat{\boldsymbol{\theta}} = [\widehat{\theta}_1, \widehat{\theta}_2]^{\mathrm{T}} = [\widehat{\mu}, \widehat{\sigma}^2]^{\mathrm{T}}$$

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \nabla_\theta \ln p(x_k \mid \theta) = 0$$

$$p(x_k \mid \theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\tfrac{1}{2} \left(\tfrac{x_k - \mu}{\sigma} \right)^2 \right]$$

$$\ln p(x_k \mid \theta) = -\tfrac{1}{2} \ln(2\pi\sigma^2) - \tfrac{1}{2} \left(\frac{x_k - \mu}{\sigma} \right)^2$$

$$= -\tfrac{1}{2} \ln(2\pi\theta_2) - \tfrac{1}{2} \frac{(x_k - \theta_1)^2}{\theta_2}$$

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \underline{\nabla_\theta \ln p(x_k \mid \theta)} = 0$$

$$\ln p(x_k \mid \theta) = -\tfrac{1}{2}\ln(2\pi) - \tfrac{1}{2}\ln(\theta_2) - \tfrac{1}{2}\frac{(x_k - \theta_1)^2}{\theta_2}$$

$$\nabla_\theta \ln p(x_k \mid \theta) = \begin{bmatrix} \dfrac{(x_k - \theta_1)}{\theta_2} \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\nabla_\theta H(\theta) = \sum_{k=1}^{N} \underline{\nabla_\theta \ln p(x_k \mid \theta)} = 0$$

$$\nabla_\theta \ln p(x_k \mid \theta) = \begin{bmatrix} \dfrac{(x_k - \theta_1)}{\theta_2} \\ -\dfrac{1}{2\theta_2} + \dfrac{(x_k - \theta_1)^2}{2\theta_2^2} \end{bmatrix}$$

$$\begin{cases} \displaystyle\sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)}{\hat{\theta}_2} = 0 \\ \displaystyle -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \end{cases}$$

最大似然估计量

满足的方程

$$\begin{cases} \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)}{\hat{\theta}_2} = 0 \\ -\sum_{k=1}^{N} \frac{1}{\hat{\theta}_2} + \sum_{k=1}^{N} \frac{(x_k - \hat{\theta}_1)^2}{\hat{\theta}_2^2} = 0 \end{cases}$$

$$\hat{\theta}_1 = \hat{\mu} = \frac{1}{N} \sum_{k=1}^{N} x_k$$ 均值

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{1}{N} \sum_{k=1}^{N} (x_k - \hat{\mu})^2$$ 方差

5.3.3 用身高、体重区分男女生的例子

到现在为止，我们**知道**：

- Bayes决策理论
- 概率密度参数的最大似然估计

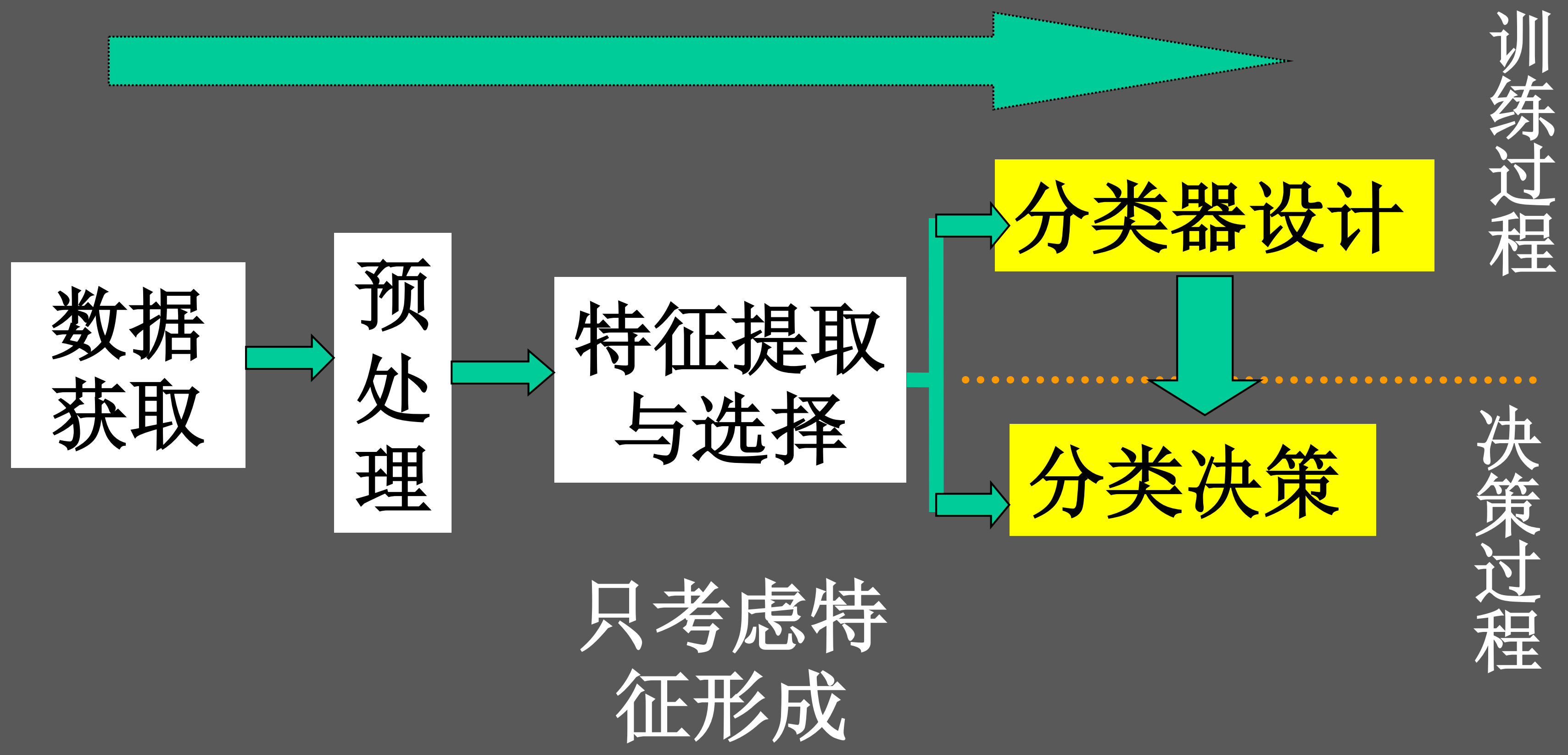
下面讲一个简单的**应用**

我们的任务可能是：

- 大学生男女同学在身高、体重方面的差别？
- 大学生男女同学在身高、体重方面是否存在明显的界限？
- 用同学们的身高、体重来区分男女同学？

解决的方案：已讲的分类方法来处理

模式识别系统的基本构造



数据获取：

给每一个同学发一张小纸条，要求同学将

自己的身高(cm)、体重(kg)、性别(男、女）

资料写在上面，最后收集小纸条

数据预处理：

- 检查身高数据与单位、体重数据与单位是否有问题，如身高以 m 为单位，体重以斤为单位，如有则统一改成 cm 和 kg
- 是否有野值数据，如，身高 220 cm

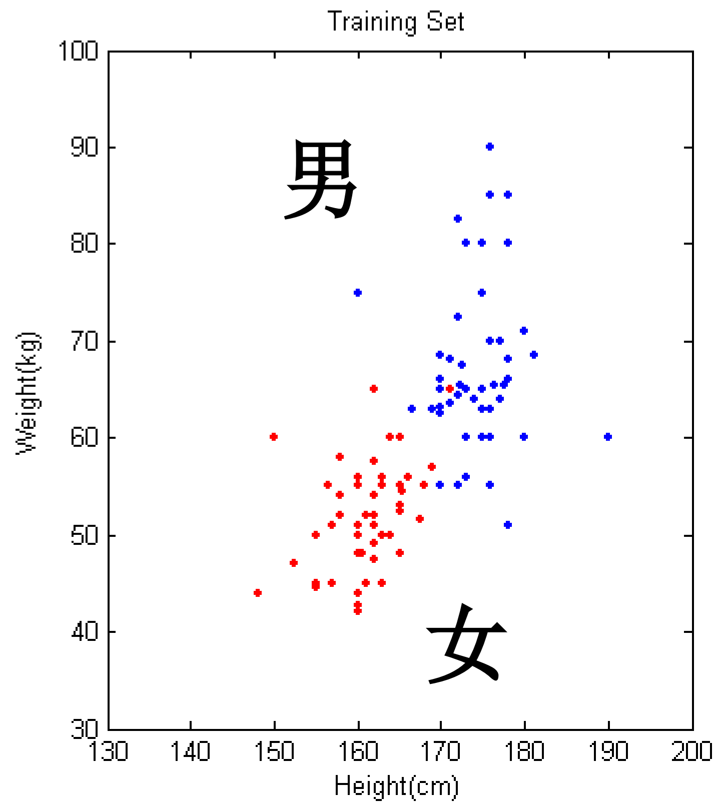
特征形成：

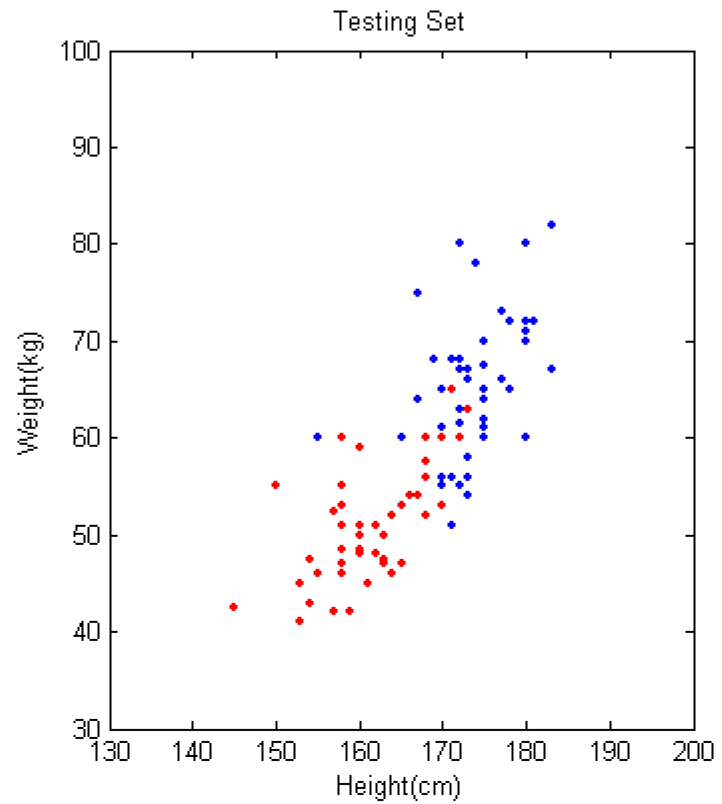
每一个同学有三个数据：

- ✓性别（类别标识）
- ✓身高（第一个特征）
- ✓体重（第二个特征）

收集整理的样本构成两个样本集，各包含50个男女同学的数据：

- ✓ **样本集1**（50个男生、50个女生）：作为训练样本集
- ✓ **样本集2**（50个男生、50个女生）：作为测试样本集


Training Set
100
90
80
70
60
50
40
30
130
140
150
160
170
180
190
200
Height(cm)
Weight(kg)
男
女


Testing Set
100
90
80
70
60
50
40
30
130
140
150
160
170
180
190
200
Height(cm)
Weight(kg)

样本集1

样本集2

Byes分类器设计

假设男女生样本分别满足各自的正态分布，

针对样本集1，利用最大似然估计方法分

别求出男女生的均值向量和协方差矩阵

$$\widehat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x}_k$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x}_k - \widehat{\boldsymbol{\mu}})(\mathbf{x}_k - \widehat{\boldsymbol{\mu}})^{\mathrm{T}}$$

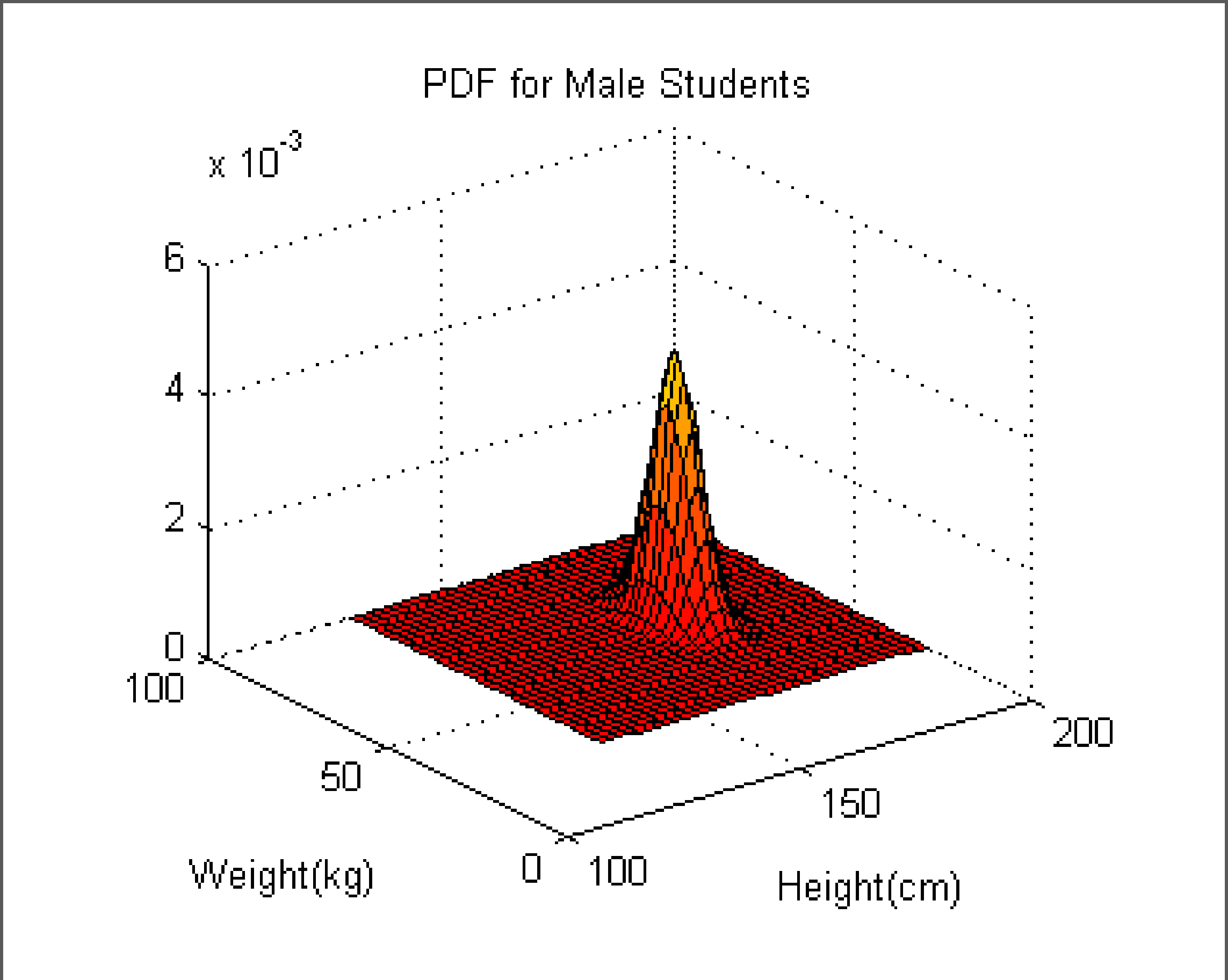
男生：均值向量和协方差矩阵

$$\boldsymbol{\mu}_m = \begin{bmatrix} 174.13 \\ 66.61 \end{bmatrix}$$

$$\Sigma_m = \begin{bmatrix} 20.91 & 2.25 \\ 2.25 & 72.01 \end{bmatrix}$$

$$|\Sigma_m| = 1500.9$$

$$\Sigma_m^{-1} = \begin{bmatrix} 0.0480 & -0.0015 \\ -0.0015 & 0.0139 \end{bmatrix}$$


PDF for Male Students
x 10^-3
6
4
2
0
100
50
0
Weight(kg)
100
150
200
Height(cm)

概率密度函数（男生）

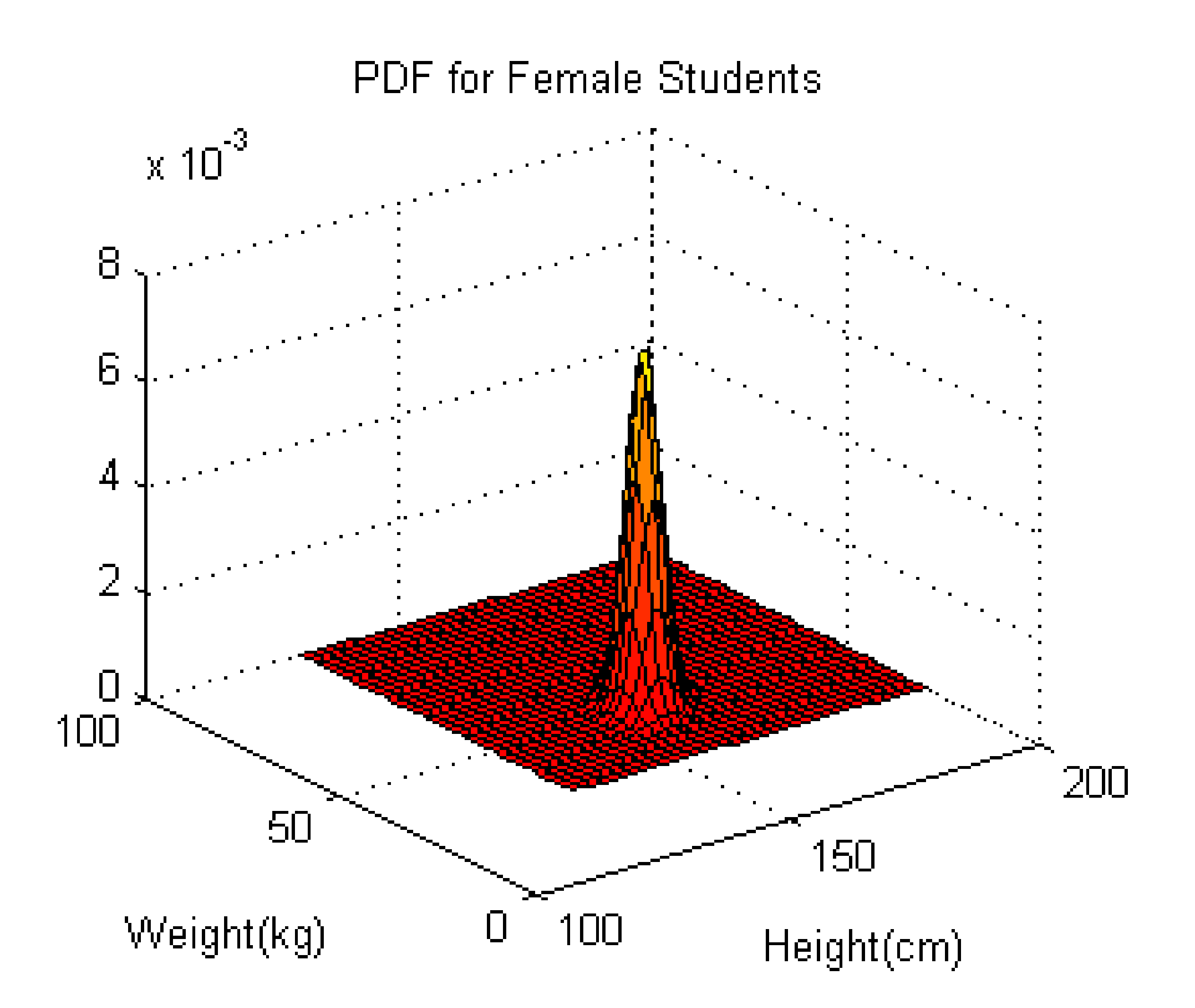
女生：均值向量和协方差矩阵

$$\boldsymbol{\mu}_f = \begin{bmatrix} 161.03 \\ 51.96 \end{bmatrix}$$

$$\Sigma_f = \begin{bmatrix} 20.06 & 9.52 \\ 9.52 & 29.78 \end{bmatrix}$$

$$\left|\Sigma_f\right| = 509.79$$

$$\Sigma_f^{-1} = \begin{bmatrix} 0.0588 & -0.0188 \\ -0.0188 & 0.0396 \end{bmatrix}$$


PDF for Female Students
x 10^-3
8
6
4
2
0
100
50
0
Weight(kg)
100
150
200
Height(cm)

概率密度函数（女生）

（最小错误率）Bayes决策规则：

$$\text{if} \quad p(\mathbf{x} \mid \omega_m) P(\omega_m) \gtrless p(\mathbf{x} \mid \omega_f) P(\omega_f)$$

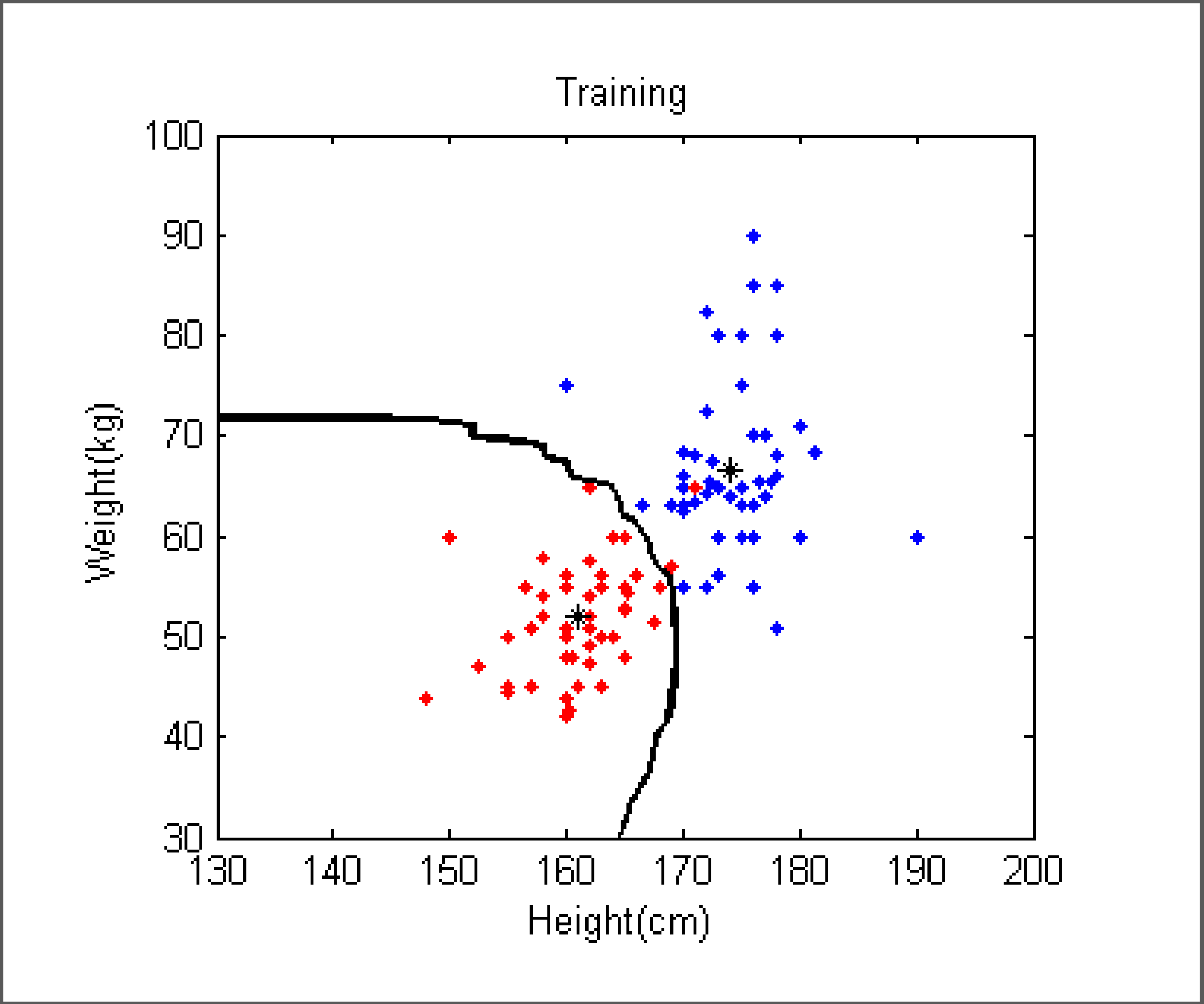
$$\textbf{then } \mathbf{x} \in \begin{cases} \omega_m & \text{男生} \\ \omega_f & \text{女生} \end{cases}$$

这里，我们假设两类先验概率相等

决策面方程：

$$p(\mathbf{x}|\omega_m)P(\omega_m) = p(\mathbf{x}|\omega_f)P(\omega_f)$$

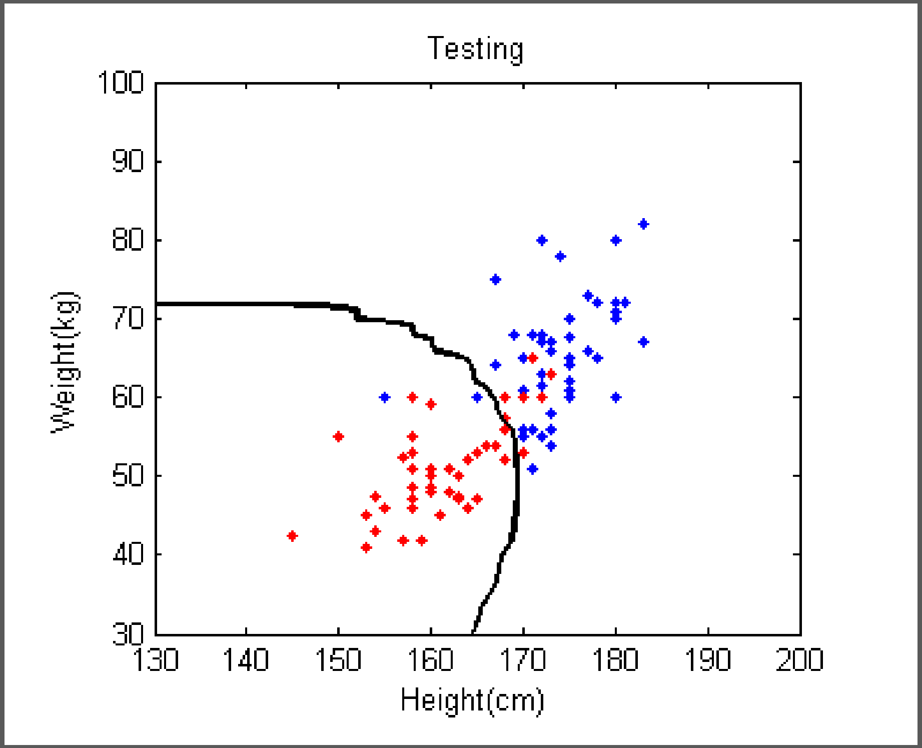
$$p(\mathbf{x}|\omega_m) = p(\mathbf{x}|\omega_f)$$

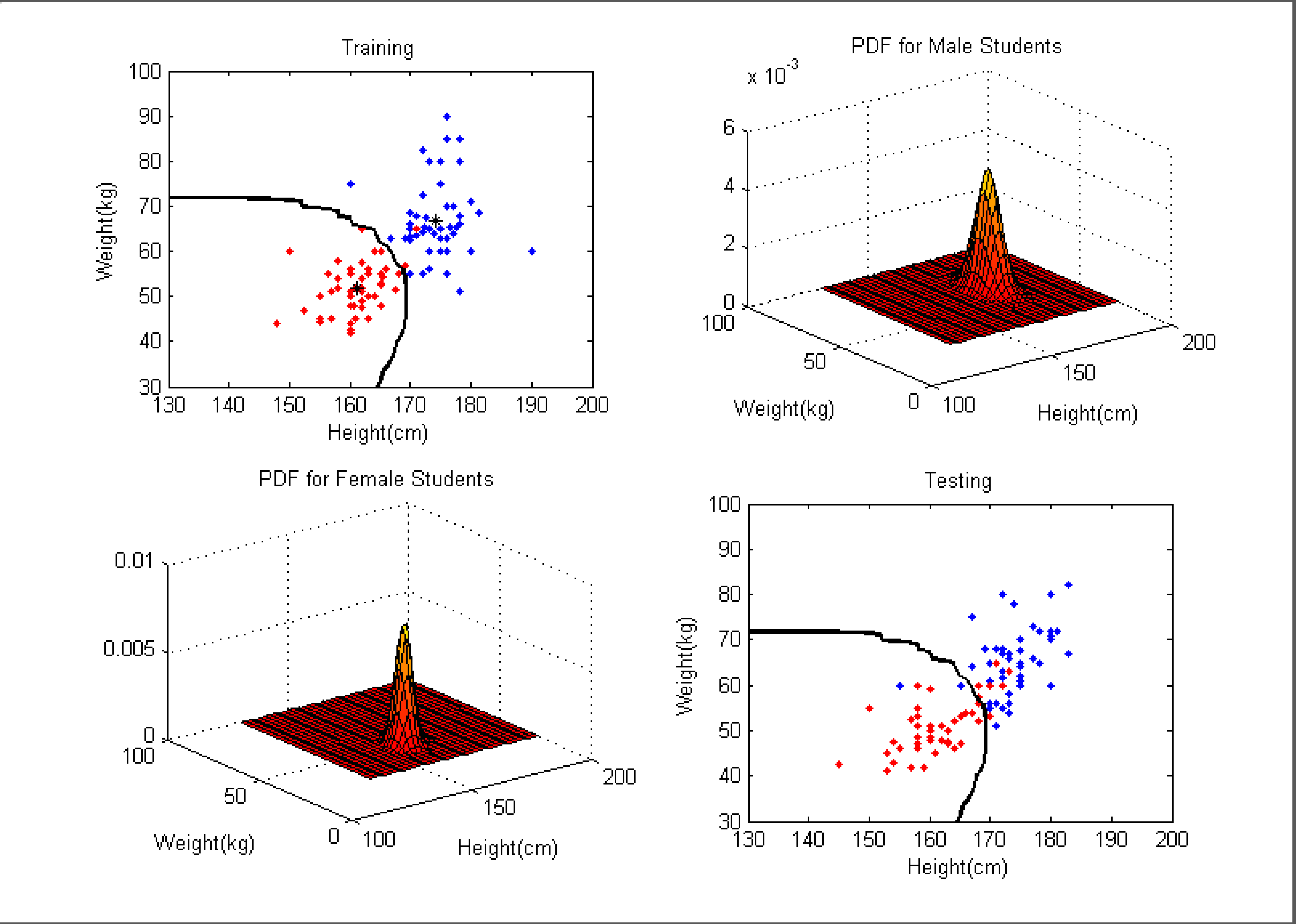

Training
100
90
80
70
60
50
40
30
130
140
150
160
170
180
190
200
Height(cm)
Weight(kg)

决策面

分类决策过程：

我们将样本集 2 作为待分类的新样本，判断每一学生的性别





Training
Weight(kg)
Height(cm)
PDF for Male Students
x 10^-3
Weight(kg)
Height(cm)
PDF for Female Students
Weight(kg)
Height(cm)
Testing
Weight(kg)
Height(cm)