

模式识别第一次作业 实验报告



专 业	控制工程
学 号	202422280516
姓 名	陈劭杰
承担内容	报告撰写

专 业	控制工程
学 号	202422280540
姓 名	郭 昊
承担内容	代码编写、报告修改

目录

一、实验目的.....	- 1 -
二、实验原理.....	- 1 -
2.1 最大似然估计.....	- 1 -
2.2 贝叶斯决策.....	- 1 -
三、实验过程.....	- 2 -
3.1 数据处理.....	- 2 -
3.2 画图.....	- 2 -
3.3 求最大似然估计参数.....	- 3 -
3.4 求被贝叶斯估计参数.....	- 3 -
3.5 决策.....	- 4 -
四、实验结论.....	- 5 -
五、实验总结.....	- 6 -
附录.....	- 6 -
①main.m.....	- 6 -
②process_data.m.....	- 7 -
③plot_weight.m.....	- 8 -
④max_estimate.m.....	- 9 -
⑥plot_decision.m.....	- 11 -

一、实验目的

- ①学习最大似然估计和贝叶斯估计的参数估计方法。
- ②掌握贝叶斯最小错误率决策方法，并通过给定的数据集进行分析预测，深刻理解统计方法在实际问题中的应用。
- ③学习 MATLAB, Python 等编程语言的使用，掌握常用的接口函数。

二、实验原理

2.1 最大似然估计

最大似然估计（Maximum Likelihood Estimation, MLE）是一种用于估计统计模型参数的方法。其基本思想是通过最大化样本数据的似然函数，找到最有可能产生观测数据的参数值。在本实验中，我们假设男生和女生的体重服从正态分布，通过最大似然估计方法求出其均值和方差。

$$\hat{\mu} = \hat{\theta}_1 = \frac{1}{N} \sum_{l=1}^N x_k \quad (2.1.1)$$

$$\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^N (x_k - \hat{\mu})^2 \quad (2.1.2)$$

2.2 贝叶斯估计

贝叶斯估计（Bayesian Estimation）是一种结合先验知识和样本数据来估计参数的方法。贝叶斯估计通过贝叶斯定理，将先验分布与样本数据的似然函数结合，得到后验分布。在本实验中，我们假设先验分布为已知的正态分布，并结合样本数据，计算男女生体重的后验均值和方差。

$$\begin{aligned} \hat{\mu} &= \mu_N \\ &= \frac{\sigma_0^2}{N\sigma_0^2 + \sigma^2} \sum_{k=1}^N x_k + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \\ &= \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0 \end{aligned} \quad (2.2.1)$$

2.3 最小错误率贝叶斯决策

贝叶斯决策（Bayesian Decision Theory）是一种基于概率论的决策方法。它通过结合先验概率和样本数据，计算后验概率，并基于最小化期望损失的原则进行决策。在本实验中，我们将使用贝叶斯估计方法求出男女生体重的分布参数，并基于这些参数进行分类决策。

$$\begin{aligned} g(x) &= g_2(x) - g_1(x) \\ &= \frac{1}{2} (x - \bar{x}_1)^T \Sigma_1^{-1} (x - \bar{x}_1) \\ &\quad - \frac{1}{2} (x - \bar{x}_2)^T \Sigma_2^{-1} (x - \bar{x}_2) \\ &\quad + \frac{1}{2} \ln \frac{|\Sigma_1|}{|\Sigma_2|} - \ln \frac{P(\omega_1)}{P(\omega_2)} \stackrel{<}{>} 0 \Rightarrow x \in \begin{matrix} \omega_1 \\ \omega_2 \end{matrix} \end{aligned} \quad (2.3.1)$$

三、实验过程

3.1 数据预处理

首先，对原始数据进行预处理，去除异常值和噪声数据，以确保数据的准确性和可靠性。

①读取数据：使用 `readtable` 函数读取 Excel 文件中的数据。确认数据集的列数是否为预期的 11 列。如果列数不匹配，则打印错误信息。

②修改列名：将数据集的列名进行编码，改为更具描述性的名称，包括编号、性别、来源地、身高、体重、鞋码、50 米成绩、肺活量、颜色、喜欢运动和喜欢文学等。

③过滤数据：只保留性别为 0 或 1 的数据（即合法值），并对喜欢运动和喜欢文学的数据进行类似的过滤（代码中暂时注释掉）。

④设定异常值过滤阈值：计算身高和体重的均值和标准差，根据 3 标准差准则进行过滤，保留在此范围内的数据。

⑤过滤异常值：过滤掉超出 3 个标准差范围的异常值，确保合理性。

⑥保存数据：处理后的数据保存在 `filtered_data.xlsx` 文件中。

filtered_data.xlsx										
filtereddata										
Num	Gender	Origin	Height	Weight	Size	m	Lungs	Color	Sport	Art
数值	▼数值	▼分类	▼数值	▼数值	▼数值	▼数值	▼数值	▼分类	▼数值	▼数值
1 Num	Gender	Origin	Height	Weight	Size	50m	Lungs	Color	Sport	Art
2	1	1 湖北	163	51	41	7.5000	2500	蓝	1	1
3	2	1 河南	171	64	41	7.5000	3500	蓝	0	0
4	3	1 云南	182	68	45	7.8000	4900	蓝	1	0
5	4	1 广西	172	66	42	8.2000	4800	绿	0	1
6	5	1 四川	185	80	44	8.5000	5100	蓝	0	0
7	6	0 河北	164	47	38	9	2500	紫	1	1
8	7	0 河南	160	46	38	9	2500	白	1	1
9	8	1 重庆	170	46	41	7	3000	蓝	1	1
10	9	1 重庆	178	60	41	7	4200	绿	0	0
11	10	1 江苏	180	71	43	7.5000	3500	紫	0	0
12	11	1 四川	185	90	45	7.5000	4500	黑	0	0
13	12	1 四川	170	60	41	7.5000	3000	橙	1	0
14	13	1 四川	181	72	44	8	4500	蓝	0	1
15	14	1 广东	174	58	43	7	3500	蓝	1	0
16	15	1 四川	180	70	42	7	4000	蓝	1	1
17	16	1 江西	175	65	42	7	3000	红	0	0
18	17	1 江西	165	50	41	7.5000	3500	蓝	1	1
19	18	1 四川	180	75	42	6.8700	4000	白	0	1
20	19	1 四川	177	80	44.5000	8.5200	3700	黑	1	0

图 1：数据预处理结果（仅展示前 20 行）

3.2 体重分布脂肪图

在数据处理完成后，绘制男女生体重的直方图，直观展示数据的分布情况。

读取数据：使用 `readtable` 函数读取处理后的 Excel 文件中的数据。

①获取体重数据：分别提取男生和女生的体重数据，便于后续绘图。

②绘制直方图：使用 `histogram` 函数分别绘制男生和女生的体重直方图，并设置不同的颜色和透明度，以便于对比。

③设置图表属性：添加图表标题、坐标轴标签和图例，确保图表信息清晰易读。

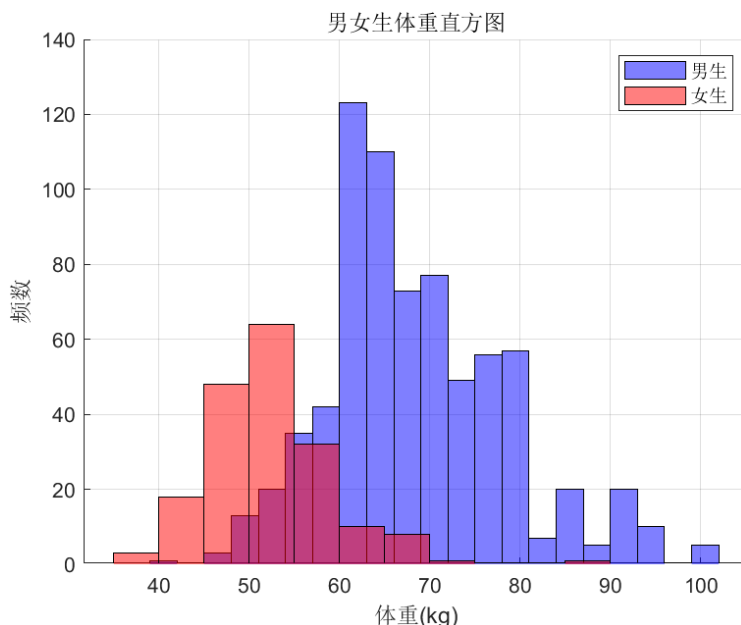


图 2：男女生体重的直方图

3.3 求最大似然估计参数

使用最大似然估计方法，计算男女生体重的均值和方差。

在数据处理和绘图之后，我们使用最大似然估计（MLE）方法来计算男女生体重的分布参数。具体步骤如下：

①读取数据：使用 `readtable` 函数读取处理后的 Excel 文件中的数据。

②获取体重数据：分别提取男生和女生的体重数据，便于后续计算。

③计算最大似然估计参数：假设体重数据服从正态分布，分别计算男生和女生体重数据的均值和标准差，作为最大似然估计的参数。

④显示结果：使用 `fprintf` 函数输出计算结果，显示男生和女生体重的均值和方差。

命令行窗口

```
男生总体的最大似然估计 (MLE)： 均值 = 67.97， 方差 = 10.03  
女生总体的最大似然估计 (MLE)： 均值 = 51.45， 方差 = 6.66
```

图 3：男生和女生体重的均值和方差。

3.4 求被贝叶斯估计参数

在已知方差的情况下，使用贝叶斯估计方法计算男女生体重的均值和方差。

①读取数据：使用 `readtable` 函数读取处理后的 Excel 文件中的数据。

②获取体重数据：分别提取男生和女生的体重数据，便于后续计算。

③设定先验参数：分别提取的男生和女生的身高和体重数据，计算均值向量和协方差矩阵，用于多元正态分布的概率密度函数计算。

假设先验方差为 1，女生的先验均值为 59，男生的先验均值为 69.6。数据来源如下：[国家国民体质监测中心发布《第五次国民体质监测公报》](#) [国家体育总局 \(sport.gov.cn\)](#)。

④计算贝叶斯估计参数：对男生、女生分别进行：计算样本数量、样本均值和样本方差；使用先验均值和样本数据，计算后验均值和方差。并显示结果，使用 `fprintf` 函数输出计算结果，显示男生和女生体重的后验均值和方差。

```
命令行窗口
选取男生先验均值： 69.60， 方差： 1.00， 女生先验均值： 59.00， 方差： 1.00
男生的贝叶斯后验估计： 均值： 68.17， 方差： 0.12
女生的贝叶斯后验估计： 均值： 52.91， 方差： 0.19
```

图 4：男生和女生体重的后验均值和方差。

3.5 最小错误率贝叶斯决策

在求得最大似然估计和贝叶斯估计参数后，使用最小错误率贝叶斯决策方法，基于身高和体重数据，绘制决策面并进行分类决策。

①读取数据：使用 `readtable` 函数读取处理后的 Excel 文件中的数据。

②获取身高和体重数据：分别提取男生和女生的身高和体重数据，便于后续计算。

③计算均值向量和协方差矩阵：分别计算男生和女生的均值向量和协方差矩阵，用于多元正态分布的概率密度函数计算。

④定义多元正态分布 PDF 函数：手动定义一个函数 `my_mvnpdf`，用于计算多元正态分布的概率密度函数值。

⑤绘制决策面：

生成网格数据，计算网格上男生和女生的判别值；

计算决策面，并绘制等高线决策面，决策面为等高线值为 0 的位置；

绘制男生和女生的散点图，添加标题和图例；

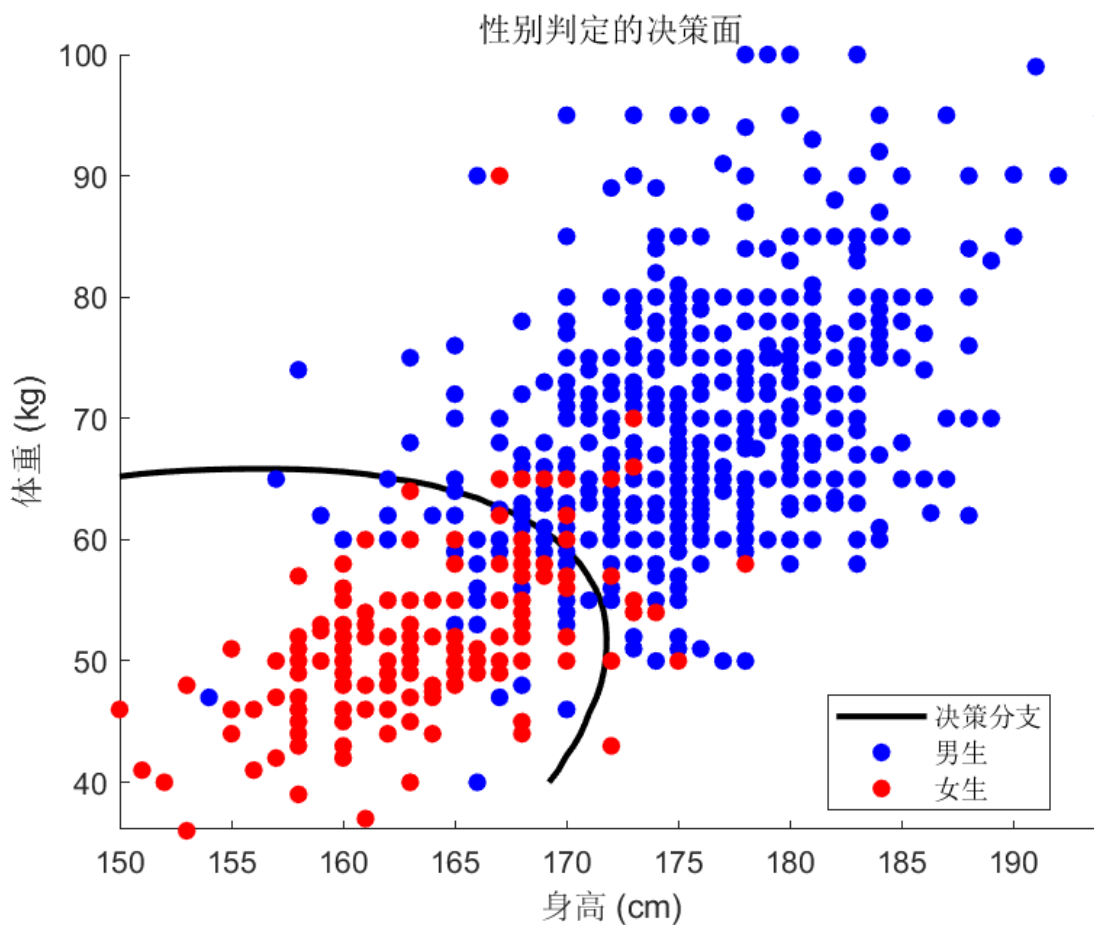


图 5：男生和女生的散点图以及决策面

⑥样本分类：

输入样本的身高和体重，计算样本属于男生和女生的概率；根据概率大小进行分类决策，判断样本属于男生还是女生。



图 6：分类决策结果

四、实验结论

①男女生的直方图绘制结果近似正态分布。

②使用最大似然估计（MLE）方法来计算男女生体重的分布参数，得到男生和女生体重的均值和方差分别为：67.97,10.03;51.45,6.66；

使用贝叶斯估计方法来计算男女生体重的分布参数，得到男生和女生体重

的均值和方差分别为：68.17,0.12;52.91,0.19;

③用最小错误率贝叶斯决策方法，基于身高和体重数据，绘制决策面并进行分类决策，判断身高为 167.00cm，体重为 52.00kg 的样本属于女生。

五、实验总结

本实验通过具体的数据分析任务，深入理解了最大似然估计和贝叶斯决策方法的应用。实验结果表明，这些方法在处理实际问题时具有较高的准确性和可靠性。未来的工作可以进一步优化数据处理和模型参数，以提高分类决策的精度。

附录

代码：

①main.m

```
1. clear;clc
2. %% 1.处理异常数据
3. process_data('data.xlsx', 'filtered_data.xlsx');
4. % 重新加载过滤后的数据
5. data = readtable('filtered_data.xlsx');
6.
7. %% 2.画直方图
8. plot_weight('filtered_data.xlsx');
9.
10.%% 3.1 求最大似然估计参数
11.[max_male_params, max_female_params] = max_estimate('filtered_data.xlsx');
12.
13.%% 3.2 求贝叶斯估计参数
14.% 选定方差为 1,先验均值，女生 59 男生 69.6
15.% 先验参数设置
16.female_xy_u0 = 59;%kg 女生均值
17.male_xy_u0 = 69.6;%kg 男生均值
18.variance = 1;%先验方差
19.% 计算
20.[bys_male_mean, bys_male_variance, bys_female_mean, bys_female_variance] = bayesian_estimate('filtered_data.xlsx',female_xy_u0,male_xy_u0,variance);
21.
22.%% 4.输入样本决策
```



```

23.%样本数据
24.height = 167;
25.weight = 52;
26.%打印决策面
27.plot_decision('filtered_data.xlsx',height,weight);
28.
29.clear;

```

②process_data.m

```

1. function process_data(input_filename, output_filename)
2.     % 读取 Excel 文件
3.     data_1 = readtable(input_filename);
4.
5.     % 确认列的数量
6.     num_columns = width(data_1); % 获取数据集列数
7.
8.     if num_columns == 11
9.         % 修改所有列名
10.        data_1.Properties.VariableNames(1:11) = {'Num','Gender','Origin','Height','Weight','Size','50m','Lungs','Color','Sport','Art'};
11.    else
12.        error('Number of new column names does not match the number of columns in the dataset.');

```

```

33.
34.    % 过滤掉超出范围的异常值
35.    data = data(abs(data.Height - height_mean) <= height_th
        reshold * height_std & ...
36.                abs(data.Weight - weight_mean) <= weight_th
        reshold * weight_std, :);
37.
38.    % 显示过滤后的数据
39.    disp(data);
40.
41.    % 保存过滤后的数据到新的 Excel 文件
42.    writetable(data, output_filename);
43.end

```

③plot_weight.m

```

1. function plot_weight(input_filename)
2.
3.    data = readtable(input_filename);
4.    % 分别获取男生和女生的体重数据
5.    male_weight = data.Weight(data.Gender == 1);
6.    female_weight = data.Weight(data.Gender == 0);
7.
8.    % 绘制直方图
9.    figure;
10.   hold on;
11.
12.   % 男生体重直方图
13.   histogram(male_weight, 'FaceColor', 'b', 'EdgeColor', '
        k', 'FaceAlpha', 0.5);
14.
15.   % 女生体重直方图
16.   histogram(female_weight, 'FaceColor', 'r', 'EdgeColor',
        'k', 'FaceAlpha', 0.5);
17.
18.   % 图表标题和标签
19.   title('男女生体重直方图');
20.   xlabel('体重(kg)');
21.   ylabel('频数');
22.
23.   % 添加图例
24.   legend('男生', '女生');
25.
26.   % 显示网格
27.   grid on;
28.   hold off;

```

```
29.end
```

④max_estimate.m

```
1. function [male_params, female_params] = max_estimate(input_
   filename)
2.     % 读取 Excel 文件
3.     data = readtable(input_filename);
4.
5.     % 分别获取男生和女生的体重数据
6.     male_weight = data.Weight(data.Gender == 1);
7.     female_weight = data.Weight(data.Gender == 0);
8.
9.     % 对男生体重进行最大似然估计（假设为正态分布）
10.    male_mean = mean(male_weight);
11.    male_std = std(male_weight);
12.    male_params = [male_mean, male_std];
13.
14.    % 对女生体重进行最大似然估计（假设为正态分布）
15.    female_mean = mean(female_weight);
16.    female_std = std(female_weight);
17.    female_params = [female_mean, female_std];
18.
19.    % 显示结果
20.    fprintf('男生总体的最大似然估计(MLE): 均值 = %.2f, 方
       差 = %.2f\n', male_mean, male_std);
21.    fprintf('女生总体的最大似然估计(MLE): 均值 = %.2f, 方
       差 = %.2f\n', female_mean, female_std);
22.end
```

⑤bayesian_estimate.m

```
1. function [bys_male_mean, bys_male_variance, bys_female_mean
   , bys_female_variance] = bayesian_estimate(input_filename,f
   emale_u0,male_u0)
2.     % 读取 Excel 文件中的数据
3.     data = readtable(input_filename);
4.
5.     % 分别提取男生和女生的体重数据
6.     male_weights = data.Weight(data.Gender == 1);
7.     female_weights = data.Weight(data.Gender == 0);
8.
9.     % 贝叶斯估计的固定先验方差为 1
10.    prior_variance = 1;
11.
12.    %% 男生的贝叶斯参数估计
13.    % 男生样本数量
14.    n_male = length(male_weights);
```

```

15.
16.     % 男生样本均值和方差
17.     male_mean_sample = mean(male_weights);
18.     male_variance_sample = var(male_weights);
19.
20.     % 先验均值 (假设为固定值或输入参数)
21.     mu0_male_prior = male_u0;
22.
23.     % 计算男生的后验均值和方差
24.     bys_male_mean = (mu0_male_prior / prior_variance + n_ma
        le * male_mean_sample / male_variance_sample) / ...
25.                     (1 / prior_variance + n_male / male_var
        iance_sample);
26.     bys_male_variance = 1 / (1 / prior_variance + n_male /
        male_variance_sample);
27.
28.     %% 女生的贝叶斯参数估计
29.     % 女生样本数量
30.     n_female = length(female_weights);
31.
32.     % 女生样本均值和方差
33.     female_mean_sample = mean(female_weights);
34.     female_variance_sample = var(female_weights);
35.
36.     % 先验均值 (假设为固定值或输入参数)
37.     mu0_female_prior = female_u0;
38.
39.     % 计算女生的后验均值和方差
40.     bys_female_mean = (mu0_female_prior / prior_variance +
        n_female * female_mean_sample / female_variance_sample) / .
        ..
41.                     (1 / prior_variance + n_female / fema
        le_variance_sample);
42.     bys_female_variance = 1 / (1 / prior_variance + n_femal
        e / female_variance_sample);
43.
44.     % 显示计算结果
45.     fprintf('选取男生先验均值: %.2f, 方差: %.2f, 女生先验均
        值: %.2f, 方
        差: %.2f\n', mu0_male_prior, prior_variance, mu0_female_prior
        , prior_variance);
46.     fprintf('男生的贝叶斯后验估计: 均值: %.2f, 方
        差: %.2f\n', bys_male_mean, bys_male_variance);

```

```

47.     fprintf('女生的贝叶斯后验估计: 均值: %.2f, 方
        差: %.2f\n', bys_female_mean, bys_female_variance);
48. End

```

⑥plot_decision.m

```

1. function plot_decision(input_filename,s_height,s_weight)
2.
3. data = readtable(input_filename);
4.
5. % 分别获取男生和女生的身高和体重数据
6. male_data = data(data.Gender == 1, {'Height', 'Weight'});
7. female_data = data(data.Gender == 0, {'Height', 'Weight'});
8.
9. % 计算男生和女生的均值向量和协方差矩阵
10. mu_male = mean(male_data{:,:}); % 男生均值向量
11. mu_female = mean(female_data{:,:}); % 女生均值向量
12.
13. sigma_male = cov(male_data{:,:}); % 男生协方差矩阵
14. sigma_female = cov(female_data{:,:}); % 女生协方差矩阵
15.
16. % 手动计算多元正态分布 PDF
17. function p = my_mvnpdf(x, mu, sigma)
18.     d = length(mu); % 维度 (2 维)
19.     x_mu = x - mu; % (x - mu)
20.     p = (1 / ((2*pi)^(d/2) * sqrt(det(sigma)))) * exp(-
        0.5 * (x_mu / sigma) * x_mu');
21. end
22.
23. % 绘制决策面
24. figure;
25. hold on;
26.
27. % 生成网格数据
28. [x1Grid, x2Grid] = meshgrid(150:1:190, 40:1:80);
29. XGrid = [x1Grid(:), x2Grid(:)]; % 网格点
30.
31. % 计算网格上男生和女生的判别值
32. g_male = arrayfun(@(i) my_mvnpdf(XGrid(i, :), mu_male, sigma_
        a_male), 1:size(XGrid, 1)); % 男生联合概率密度
33. g_female = arrayfun(@(i) my_mvnpdf(XGrid(i, :), mu_female,
        sigma_female), 1:size(XGrid, 1)); % 女生联合概率密度
34.
35. % 计算决策面
36. decision_surface = reshape(g_male - g_female, size(x1Grid))
    ;

```

```

37.
38.% 绘制等高线决策面，决策面为等高线值为 0 的位置
39.contour(x1Grid, x2Grid, decision_surface, [0 0], 'k', 'Line
    Width', 2);
40.
41.% 绘制男生和女生的散点图
42.scatter(male_data.Height, male_data.Weight, 'b', 'filled');
43.scatter(female_data.Height, female_data.Weight, 'r', 'filled');
44.
45.% 添加标题和图例
46.title('性别判定的决策面');
47.xlabel('身高 (cm)');
48.ylabel('体重 (kg)');
49.legend('决策分支', '男生', '女生', 'Location', 'best');
50.
51.hold off;
52.
53.% 样本身高体重的分类
54.sample = [s_hight, s_weight];
55.fprintf('选择身高为%.2fcm，体重为%.2fkg 的测试集
    \n', s_hight, s_weight);
56.
57.% 计算样本属于男生和女生的概率
58.p_male = my_mvnpdf(sample, mu_male, sigma_male); % 男生概率
59.p_female = my_mvnpdf(sample, mu_female, sigma_female); % 女生概率
60.
61.% 分类决策
62.if p_male > p_female
63.    fprintf('分类结果为男生\n');
64.else
65.    fprintf('分类结果为女生\n');
66.end
67.
68.end

```