|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 电子科技大学  UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA | | | | |
| **模式识别第一次作业**  **实验报告** | | | |
| 黑白色的标志  低可信度描述已自动生成 | | | |
|  | **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 -](#_Toc178336464)

[二、实验原理 - 1 -](#_Toc178336465)

[2.1 最大似然估计 - 1 -](#_Toc178336466)

[2.2 贝叶斯决策 - 1 -](#_Toc178336467)

[三、实验过程 - 2 -](#_Toc178336468)

[3.1 数据处理 - 2 -](#_Toc178336469)

[3.2 画图 - 2 -](#_Toc178336470)

[3.3 求最大似然估计参数 - 3 -](#_Toc178336471)

[3.4 求被贝叶斯估计参数 - 3 -](#_Toc178336472)

[3.5 决策 - 4 -](#_Toc178336473)

[四、实验结论 - 5 -](#_Toc178336474)

[五、实验总结 - 6 -](#_Toc178336475)

[附录 - 6 -](#_Toc178336476)

[①main.m - 6 -](#_Toc178336477)

[②process\_data.m - 7 -](#_Toc178336478)

[③plot\_weight.m - 8 -](#_Toc178336479)

[④max\_estimate.m - 9 -](#_Toc178336480)

[⑥plot\_decision.m - 11 -](#_Toc178336481)

# 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。

（2.1.1）

（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。

（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。

（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。

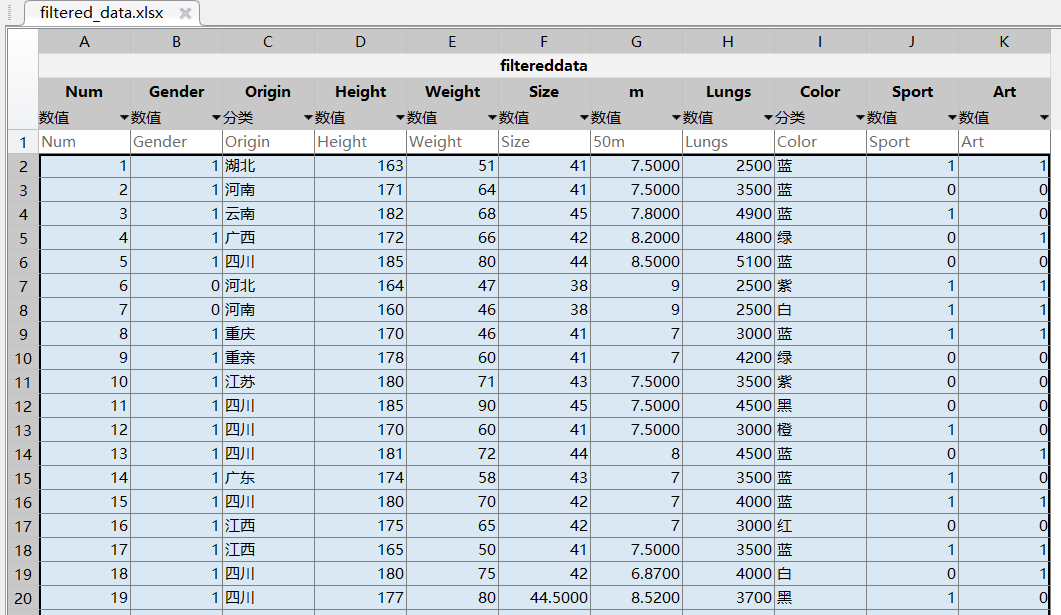


图1：数据预处理结果（仅展示前20行）

## 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。

图表, 直方图

描述已自动生成

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。

图片包含 文本

描述已自动生成

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)](https://www.sport.gov.cn/n315/n329/c24335066/content.html)。

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。

图示

描述已自动生成

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；

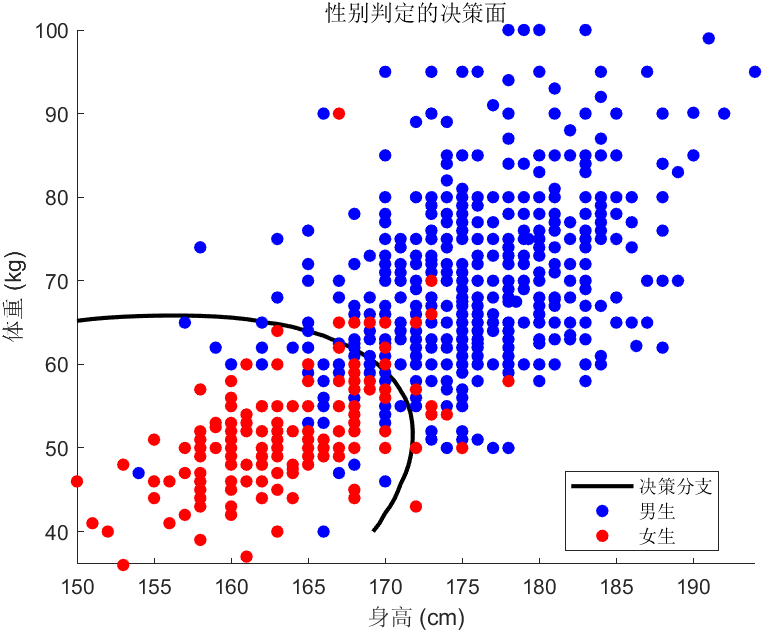


图5：男生和女生的散点图以及决策面

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。

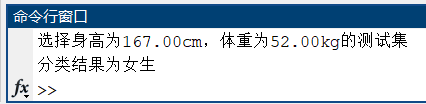


图6：分类决策结果

# 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. %% 2.画直方图
7. plot\_weight('filtered\_data.xlsx');
8. %% 3.1 求最大似然估计参数
9. [max\_male\_params, max\_female\_params] = max\_estimate('filtered\_data.xlsx');
10. %% 3.2 求贝叶斯估计参数
11. % 选定方差为1,先验均值，女生59 男生69.6
12. % 先验参数设置
13. female\_xy\_u0 = 59;%kg 女生均值
14. male\_xy\_u0 = 69.6;%kg 男生均值
15. variance = 1;%先验方差
16. % 计算
17. [bys\_male\_mean, bys\_male\_variance, bys\_female\_mean, bys\_female\_variance] = bayesian\_estimate('filtered\_data.xlsx',female\_xy\_u0,male\_xy\_u0,variance);
18. %% 4.输入样本决策
19. %样本数据
20. height = 167;
21. weight = 52;
22. %打印决策面
23. plot\_decision('filtered\_data.xlsx',height,weight);
24. clear;

## ②process\_data.m

1. function process\_data(input\_filename, output\_filename)
2. % 读取Excel文件
3. data\_l = readtable(input\_filename);
4. % 确认列的数量
5. num\_columns = width(data\_l); % 获取数据集列数
6. if num\_columns == 11
7. % 修改所有列名
8. data\_l.Properties.VariableNames(1:11) = {'Num','Gender','Origin','Height','Weight','Size','50m','Lungs','Color','Sport','Art'};
9. else
10. error('Number of new column names does not match the number of columns in the dataset.');
11. end
12. % 过滤性别、喜欢运动和喜欢文学的数据（只保留合法值 0 和 1）
13. data = data\_l(data\_l.Gender == 0 | data\_l.Gender == 1, :);
14. % data = data(data.Sport == 0 | data.Sport == 1, :);
15. % data = data(data.Art == 0 | data.Art == 1, :);
16. % 设定异常值的过滤阈值（对身高和体重的异常值进行过滤）
17. % 计算身高的均值和标准差
18. height\_mean = mean(data.Height, 'omitnan');
19. height\_std = std(data.Height, 'omitnan');
20. % 计算体重的均值和标准差
21. weight\_mean = mean(data.Weight, 'omitnan');
22. weight\_std = std(data.Weight, 'omitnan');
23. % 设置过滤条件，保留在3个标准差范围内的数据
24. height\_threshold = 3;
25. weight\_threshold = 3;
26. % 过滤掉超出范围的异常值
27. data = data(abs(data.Height - height\_mean) <= height\_threshold \* height\_std & ...
28. abs(data.Weight - weight\_mean) <= weight\_threshold \* weight\_std, :);
29. % 显示过滤后的数据
30. disp(data);
31. % 保存过滤后的数据到新的Excel文件
32. writetable(data, output\_filename);
33. end

## ③plot\_weight.m

1. function plot\_weight(input\_filename)
2. data = readtable(input\_filename);
3. % 分别获取男生和女生的体重数据
4. male\_weight = data.Weight(data.Gender == 1);
5. female\_weight = data.Weight(data.Gender == 0);
6. % 绘制直方图
7. figure;
8. hold on;
10. % 男生体重直方图
11. histogram(male\_weight, 'FaceColor', 'b', 'EdgeColor', 'k', 'FaceAlpha', 0.5);
13. % 女生体重直方图
14. histogram(female\_weight, 'FaceColor', 'r', 'EdgeColor', 'k', 'FaceAlpha', 0.5);
15. % 图表标题和标签
16. title('男女生体重直方图');
17. xlabel('体重(kg)');
18. ylabel('频数');
20. % 添加图例
21. legend('男生', '女生');
23. % 显示网格
24. grid on;
25. hold off;
26. 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. % 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. % 对男生体重进行最大似然估计（假设为正态分布）
8. male\_mean = mean(male\_weight);
9. male\_std = std(male\_weight);
10. male\_params = [male\_mean, male\_std];
11. % 对女生体重进行最大似然估计（假设为正态分布）
12. female\_mean = mean(female\_weight);
13. female\_std = std(female\_weight);
14. female\_params = [female\_mean, female\_std];
15. % 显示结果
16. fprintf('男生总体的最大似然估计(MLE): 均值 = %.2f, 方差 = %.2f\n', male\_mean, male\_std);
17. fprintf('女生总体的最大似然估计(MLE): 均值 = %.2f, 方差 = %.2f\n', female\_mean, female\_std);
18. end

## ⑤bayesian\_estimate.m

1. function [bys\_male\_mean, bys\_male\_variance, bys\_female\_mean, bys\_female\_variance] = bayesian\_estimate(input\_filename,female\_u0,male\_u0)
2. % 读取Excel文件中的数据
3. data = readtable(input\_filename);
4. % 分别提取男生和女生的体重数据
5. male\_weights = data.Weight(data.Gender == 1);
6. female\_weights = data.Weight(data.Gender == 0);
7. % 贝叶斯估计的固定先验方差为1
8. prior\_variance = 1;
9. %% 男生的贝叶斯参数估计
10. % 男生样本数量
11. n\_male = length(male\_weights);
12. % 男生样本均值和方差
13. male\_mean\_sample = mean(male\_weights);
14. male\_variance\_sample = var(male\_weights);
15. % 先验均值 (假设为固定值或输入参数)
16. mu0\_male\_prior = male\_u0;
17. % 计算男生的后验均值和方差
18. bys\_male\_mean = (mu0\_male\_prior / prior\_variance + n\_male \* male\_mean\_sample / male\_variance\_sample) / ...
19. (1 / prior\_variance + n\_male / male\_variance\_sample);
20. bys\_male\_variance = 1 / (1 / prior\_variance + n\_male / male\_variance\_sample);
21. %% 女生的贝叶斯参数估计
22. % 女生样本数量
23. n\_female = length(female\_weights);
24. % 女生样本均值和方差
25. female\_mean\_sample = mean(female\_weights);
26. female\_variance\_sample = var(female\_weights);
27. % 先验均值 (假设为固定值或输入参数)
28. mu0\_female\_prior = female\_u0;
29. % 计算女生的后验均值和方差
30. bys\_female\_mean = (mu0\_female\_prior / prior\_variance + n\_female \* female\_mean\_sample / female\_variance\_sample) / ...
31. (1 / prior\_variance + n\_female / female\_variance\_sample);
32. bys\_female\_variance = 1 / (1 / prior\_variance + n\_female / female\_variance\_sample);
33. % 显示计算结果
34. 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);
35. fprintf('男生的贝叶斯后验估计：均值: %.2f, 方差: %.2f\n', bys\_male\_mean, bys\_male\_variance);
36. fprintf('女生的贝叶斯后验估计：均值: %.2f, 方差: %.2f\n', bys\_female\_mean, bys\_female\_variance);
37. End

## ⑥plot\_decision.m

1. function plot\_decision(input\_filename,s\_hight,s\_weight)
2. data = readtable(input\_filename);
3. % 分别获取男生和女生的身高和体重数据
4. male\_data = data(data.Gender == 1, {'Height', 'Weight'});
5. female\_data = data(data.Gender == 0, {'Height', 'Weight'});
6. % 计算男生和女生的均值向量和协方差矩阵
7. mu\_male = mean(male\_data{:,:});  % 男生均值向量
8. mu\_female = mean(female\_data{:,:});  % 女生均值向量
9. sigma\_male = cov(male\_data{:,:});  % 男生协方差矩阵
10. sigma\_female = cov(female\_data{:,:});  % 女生协方差矩阵
11. % 手动计算多元正态分布PDF
12. function p = my\_mvnpdf(x, mu, sigma)
13. d = length(mu);  % 维度 (2维)
14. x\_mu = x - mu;   % (x - mu)
15. p = (1 / ((2\*pi)^(d/2) \* sqrt(det(sigma)))) \* exp(-0.5 \* (x\_mu / sigma) \* x\_mu');
16. end
17. % 绘制决策面
18. figure;
19. hold on;
20. % 生成网格数据
21. [x1Grid, x2Grid] = meshgrid(150:1:190, 40:1:80);
22. XGrid = [x1Grid(:), x2Grid(:)];  % 网格点
23. % 计算网格上男生和女生的判别值
24. g\_male = arrayfun(@(i) my\_mvnpdf(XGrid(i, :), mu\_male, sigma\_male), 1:size(XGrid, 1));  % 男生联合概率密度
25. g\_female = arrayfun(@(i) my\_mvnpdf(XGrid(i, :), mu\_female, sigma\_female), 1:size(XGrid, 1));  % 女生联合概率密度
26. % 计算决策面
27. decision\_surface = reshape(g\_male - g\_female, size(x1Grid));
28. % 绘制等高线决策面，决策面为等高线值为0的位置
29. contour(x1Grid, x2Grid, decision\_surface, [0 0], 'k', 'LineWidth', 2);
30. % 绘制男生和女生的散点图
31. scatter(male\_data.Height, male\_data.Weight, 'b', 'filled');
32. scatter(female\_data.Height, female\_data.Weight, 'r', 'filled');
33. % 添加标题和图例
34. title('性别判定的决策面');
35. xlabel('身高 (cm)');
36. ylabel('体重 (kg)');
37. legend('决策分支', '男生', '女生', 'Location', 'best');
38. hold off;
39. % 样本身高体重的分类
40. sample = [s\_hight, s\_weight];
41. fprintf('选择身高为%.2fcm，体重为%.2fkg的测试集\n',s\_hight,s\_weight);
42. % 计算样本属于男生和女生的概率
43. p\_male = my\_mvnpdf(sample, mu\_male, sigma\_male);  % 男生概率
44. p\_female = my\_mvnpdf(sample, mu\_female, sigma\_female);  % 女生概率
45. % 分类决策
46. if p\_male > p\_female
47. fprintf('分类结果为男生\n');
48. else
49. fprintf('分类结果为女生\n');
50. end
51. end