第三周任务

一、 数据预处理

首先获得鸢尾花数据集之后，发现数据中存在 4 个因变量和三个品种。此数据集包含了 150 个样本，这次的回归主要探求的是建立一个多元线性回归，并使用这个数据集对模型进行评估，

为了更加方便的处理这些变量，决定采用一些数学语言来表示数据集当中的一些变量。其中将 4 个因变量花萼长度（Sepal Length）、花萼宽度（SepalWidth）、花瓣长度（PetalLength）、花瓣宽度（PetalWidth）分别定义为 x_1, x_2, x_3, x_4，三个品种山鸢尾（Setosa）、变色鸢尾（Versicolor）、维吉尼亚鸢尾（Virginical）分别使用 1,2,3 来表示作为因变量。其最终数据如表 1:

表 **1**　符号说明

符号	代表意义
x_1	花萼长度
x_2	花萼宽度
x_3	花瓣长度
x_4	花瓣宽度
y	因变量代表 3 种品种
$\beta_i, i = 0, 1, 2, 3$	因变量的系数

在这里可以写出想要的多元线性回归的方程为:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_4 x_4$$

其中这里 β_0 表示的是截距项，$\beta_1 ... \beta_4$ 表示的是因变量前面的系数，而要求解的就是这个未知数并判断这个模型是否有效。

二、 求解方法及其过程

首先由于只有这 150 个数据，所以要将数据分为测试集和训练集，这里将随机抽取数据中 80% 的数据作为测试集，剩下的 20% 测试集来验证这个模型是否有效。由于篇幅的问题，我将会把随机抽取的数据放在附件里面，这样便于查看。

这里我采用了正规矩阵法来求解这个线性模型。对于正规矩阵法：

假设一些数据，假设 X 是一个 m*(n+1) 的矩阵，其中每一行都对应了一个单独的训练样本，X 矩阵如下：

$$X = \begin{bmatrix} \left(x^{(1)}\right)^T \\ . \\ . \\ . \\ \left(x^{(m)}\right)^T \end{bmatrix}$$

假设 y 是一个 m 维的向量，包含了所有训练集中的标签（即因变量），y 向量如下：

$$y = \begin{bmatrix} \left(y^{(1)}\right) \\ . \\ . \\ . \\ \left(y^{(m)}\right) \end{bmatrix}$$

接下来就是代价函数:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

为了方便运算，要将代价函数转化为矩阵的形式才可以，转化为矩阵的形式如下：

$$J(\theta) = \frac{1}{2m} (X\theta - y)^T (X\theta - y)$$

其中里面的 θ 向量包含了需要的截距和因变量前面的系数。最终采用矩阵求导法则可以得出：

$$\theta = \left(X^T X\right)^{-1} X^T y$$

这个样子就可以求出截距和所需要的系数了。

三、 模型的求解

将数据导入编程语言中后，随机将数据 80% 的样本作为测试集用于训练模型，最终算得的系数如表 2:

表 **2**　截距与系数

变量	结果
β_0	1.201495
β_1	-0.07571
β_2	-0.0719
β_3	0.154351
β_4	0.740874

所以可以得出最后的回归方程为：

$$y = 1.201495 - 0.07571x_1 - 0.0719x_2 + 0.154351x_3 + 0.740874x_4$$

四、 模型的检验

上述随机抽取了 80% 的数据作为训练集，最终得到了多元线性回归的结果，接下来就是要将剩下的 20% 的数据作为测试集，来检验最终得到的这个模型合不合理。

将剩余的 20% 的数据带入到上述的回归方程中去，由于篇幅原因，这里只展示部分数据，剩余的数据在附录中。最终结果如表 3:

表 **3**　测试集的回归数据

测试集样本	模拟后数据	原本的分类
3	0.964406	1
4	1.010038	1
13	0.912564	1
21	0.95877	1
22	1.077217	1
47	0.937287	1
55	2.329386	2

在将数据带回回归方程后，进一步确定属于那个分类，这里采用四舍五入的方法来做最终的确定，最终的部分分类如下表 4:

表 **4**　最终的结果

样本	回归结果	原始分类	回归后分类
113	2.875733	3	3
117	2.676184	3	3
123	2.93311	3	3
131	2.789124	3	3
135	2.454314	3	2
137	3.122524	3	3
140	2.845537	3	3
141	3.113811	3	3

接下来就是判断模型的准确性了，采用最终的结果与原先的分类进行对比，看看符不符合原先的分类，并将最后的结果除以测试集样本的个数即可，最终算出模型的准确率为 96.77%。模型的准确率较高，证明了模型的有效。

附录 A 回归代码（R 语言）

```
library(openxlsx)
library(pracma) #矩阵的处理以后可以加上这个包。
library(openxlsx)
#导入处理后的鸢尾花数据集。
iris_data<-read.xlsx("iris数据集.xlsx",
colNames = TRUE,
rowNames = TRUE)
head(iris_data)
#将数据随机分成测试集和训练集。
a<-numeric(100) #用于生成0向量或者矩阵。
n<-nrow(iris_data)
T_index<-sample(n,0.8*n,replace = FALSE)
train_data<-iris_data[T_index,]
lab_data<-iris_data[-T_index,]
one_data<-ones(n*0.8,1)
one2_data<-ones(n*0.2,1) #这里是要使用1矩阵来求和。
#合并数据
dim(B)
dim(lab_data) #查看数据的大小。
B<-cbind(one_data,train_data) #B表示的是训练集。
B<-as.matrix(B) #将数据转化为矩阵的形式。
C<-cbind(one2_data,lab_data) #C表示的是测试集的数据
C<-as.matrix(C) #将数据转化为矩阵的形式。
dim(C)
#开始求解系数和常量。
B1<-B[,-6] #首先提取出所有的自变量。这里一定要注意提取。
finish<-solve(t(B1)%*%B1)%*%t(B1)%*%as.matrix(train_data[,5])
#最终算出来的finish是前面的系数。

#接下来就是使用训练集来预测测试集的数据准不准确。
finish2<-C[,-6]%*%finish #其中finish2表示的是测试集预测的数据。

finish3<-cbind(finish2,as.matrix(C[,6]))
#finish3表示的是合并数据来进行对比
write.csv(finish,file = '截距与数据.csv',
fileEncoding = 'utf-8',row.names=T)
dim(finish3)
finish4<-finish2
#这里的for循环要主动修改，若使用其他的数据一定要修改。
#但是这里的代码只是可以使用这个iris数据集。
#总共有3个分类所以写3个if即可。
for (i in 1:30)
{
  if(finish4[i]<1.5)
```

```
  {
    finish4[i]=1
  }
  else if (finish4[i]>=1.5&finish4[i]<2.5)
  {
    finish4[i]=2
  }
  if(finish4[i]>2.5)
  {
    finish4[i]=3
  }

}
finish5<-cbind(finish3,finish4)
write.csv(finish5,file = '最终结果.csv',
fileEncoding = 'utf-8',row.names=T)

#接下来就是数据可视化和模型评估的环节。
#模型评估，这里主要是看看使用回归方程预测后的品种
#和实际品种符不符合即可。
p=0#初始化。
for(i in 1:30)
{
  if(finish4[i]==as.matrix(C[,6])[i])
  {
    p=p+1
  }
}
accu=p/30
#其中accu表示的是最终的准确率。
#这里算出来的是这一次随机抽取样本的准确率。
#上面的结果算出来最终的准确率为0.97，可以说明这个模型训练的
#还是比较好的。
```