程序设计练习2：逻辑回归

机器学习

# 介绍

在本练习中，您将实现逻辑回归，并将其应用于两个不同的数据集。在开始编程练习之前，我们强烈建议观看视频讲座并完成相关主题的复习问题。

要开始练习，您需要下载起始代码并将其内容解压缩到您希望完成练习的目录中。如果需要，在开始本练习之前，请使用octave/matlab中的cd命令更改到此目录。

您也可以在课程网站的“环境设置说明”中找到安装octave/matlab的说明。

## 本练习中包含的文件

ex2.m-八度/matlab脚本，它将引导您完成练习

ex2 reg.m-八度/matlab脚本（练习的后半部分）

ex2data1.txt-练习前半部分的训练集

ex2data2.txt-练习后半部分的训练集

submit.m-将解决方案发送到服务器的提交脚本

mapfeature.m-有趣生成多项式特征的作用

plot decision boundary.m-函数绘制分类器的决策边界

[]plot data.m-绘制二维分类数据的函数

[]sigmoid.m-sigmoid函数*？？*

[]cost function.m-逻辑回归成本函数*？*

[]Predict.m-逻辑回归预测函数*？*

[]CostFunctionReg.m-规范化逻辑回归成本*？*

*？*指示需要完成的文件

在整个练习中，您将使用脚本ex2.m和ex2D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image001.gif这些脚本为问题设置数据集，并调用将要编写的函数。您不需要修改其中任何一个。您只需要按照本分配中的说明修改其他文件中的函数。

## 在哪里寻求帮助

本课程的练习使用Octaveor Matlab，这是一种高级编程语言，非常适合数值计算。如果您没有安装八度音阶或matlab，请参考课程网站“环境设置说明”中的安装说明。[〔1〕](" \l "_ftn1" \o ")

在octave/matlab命令行中，键入帮助，然后键入函数名，显示内置函数的文档。例如，帮助图将显示用于绘制的帮助信息。有关八度函数的更多文档，请参见八度文档页面。matlab文档可以在matlab文档页面中找到。

我们还强烈鼓励使用在线讨论与其他学生讨论练习。但是，不要查看其他人编写的任何源代码，也不要与其他人共享源代码。

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image002.gif

# 1             对数几率回归

在练习的这一部分中，您将构建一个逻辑回归模型来预测学生是否被大学录取。

假设你是一个大学系的管理者，你想根据每个申请者在两次考试中的成绩来确定他们的入学机会。您有以前申请者的历史数据，可以用作逻辑回归的培训集。对于每个培训示例，您都有申请人在两次考试中的分数和录取决定。

你的任务是建立一个分类模型，根据这两个考试的分数来估计申请人的入学概率。这个大纲和ex2.m中的框架代码将指导您完成这个练习。

## 1.1           可视化数据

在开始实施任何学习算法之前，如果可能的话，最好是可视化数据。在ex2.m的第一部分中，代码将通过调用函数plot data加载数据并将其显示在二维绘图上。

现在您将完成plotdata中的代码，以便它显示如图1所示的图，其中轴是两个考试分数，正负示例用不同的标记显示。



图1：训练数据散点图

为了帮助您更熟悉绘图，我们将plotdata.m留空，以便您自己尝试实现它。但是，这是一个可选（未分级）练习。我们还提供下面的实现，以便您复制或引用它。如果您选择复制我们的示例，请通过查阅octave/matlab文档，确保您了解每个命令都在做什么。

|  |
| --- |
| %查找正负示例的索引pos=find（y==1）；neg=find（y==0）；  %绘图示例绘图（x（pos，1），x（pos，2），“k+”，“线宽”，2，……  “markersize”，7）；绘图（x（neg，1），x（neg，2），“ko”，“markerfacecolor”，“y”，…  “标记化”，7）； |

## 1.2           实施

### 1.2.1          热身运动：乙状结肠功能

在开始使用实际成本函数之前，请记住逻辑回归假设定义为：

*H时*（x）＝g（th-tx），

其中，函数g是乙状结肠函数。乙状结肠功能定义如下：

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image004.gif*.*

第一步是在sigmoid.m中实现这个函数，这样程序的其余部分就可以调用它。完成后，尝试通过在octave/matlab命令行调用sigmoid（x）来测试一些值。对于x的大正值，乙状结肠应接近1，而对于大负值，乙状结肠应接近0。评估乙状结肠（0）应该给你准确的0.5。您的代码还应该使用向量和矩阵。对于矩阵，您的函数应该对每个元素执行sigmoid函数。

您可以通过在octave/matlab命令行中键入submit提交解决方案进行评分。提交脚本将提示您输入登录电子邮件和提交令牌，并询问您要提交哪些文件。您可以从网页获取分配的提交令牌。

*现在您应该提交解决方案。*

### 1.2.2          成本函数和梯度

现在，您将实现逻辑回归的成本函数和梯度。

在costfunction.m中完成代码以返回成本和梯度。

回想一下，逻辑回归中的成本函数是

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image005.gif*,*

成本梯度是与θ长度相同的向量，其中jth元素（j=0,1，…，n）定义如下：

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image006.gif

注意，虽然这个梯度看起来与线性回归梯度相同，但是公式实际上是不同的，因为线性回归和逻辑回归对hθ（x）的定义不同。

完成后，ex2.m将使用θ的初始参数调用costFunction。你应该看到成本大约是0.693。

*现在您应该提交解决方案。*

### 1.2.3          使用fminunc学习参数

在上一个任务中，通过实现梯度下降找到线性回归模型的最佳参数。编写了一个成本函数并计算了它的梯度，然后相应地进行了一个梯度下降步骤。这次，您将使用一个名为fminunc的八度/matlab内置函数，而不是执行梯度下降步骤。

Octave/Matlab的fminunc是一个优化解算器，可以找到未经训练函数的最小值。对于逻辑回归，需要使用参数θ优化成本函数j（θ）。[〔2〕](" \l "_ftn2" \o ")

具体来说，在给定一个固定的数据集（x和y值）的情况下，您将使用fminunc为逻辑回归成本函数找到最佳参数θ。您将向fminunc传递以下输入：

•我们试图优化的参数的初始值。

•当给定训练集和特定θ时，计算与数据集θ相关的逻辑回归成本和梯度的函数（x，y）

在ex2.m中，我们已经编写了使用正确参数调用fminunc的代码。

|  |
| --- |
| %设置fminunc选项的选项=optimset（“gradobj”，“on”，“maxiter”，400）；  %运行fminunc获得最佳的theta%这个函数将返回theta和cost[theta，cost]=…  fminunc（@（t）（CostFunction（t，x，y）），初始D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image007.giftheta，选项）； |

在这个代码片段中，我们首先定义了要与fminunc一起使用的选项。具体来说，我们将gradobj选项设置为on，这告诉fminunc我们的函数返回成本和梯度。这允许fminunc在最小化函数时使用渐变。此外，我们将maxiter选项设置为400，以便fminunc在终止前最多运行400步。

为了指定我们要最小化的实际函数，我们使用“short hand”来指定带有@（t）（costfunction（t，x，y））的函数。这将创建一个带有参数t的函数，它调用costFunction。这允许我们包装costFunction以用于fminunc。

如果您正确地完成了costFunction，fminunc将收敛于正确的优化参数，并返回成本和θ的最终值。注意，通过使用fminunc，您不必自己编写任何循环，也不必像使用梯度下降那样设置学习速率。这一切都是由fminunc完成的：您只需要提供一个计算成本和梯度的函数。

一旦fminunc完成，ex2.m将使用θ的最佳参数调用costFunction函数。你应该知道费用大概是

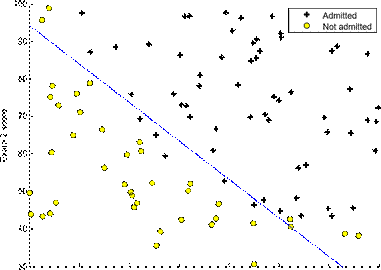
0.203.

最后的θ值将用于绘制训练数据上的决策边界，从而得到与图2类似的图。我们还鼓励您查看plotdecisionboundary.m中的代码，了解如何使用θ值绘制这样的边界。

### 1.2.4          评价逻辑回归

学习完参数后，您可以使用该模型预测某个学生是否会被录取。对于考试1分为45分，考试2分为85分的学生，您应该看到0.776的入学概率。

评估我们发现的参数质量的另一种方法是查看学习的模型对我们的训练集的预测有多好。在这



                               30             40             50             60             70             80             90             100

考试1分

图2：具有决策边界的培训数据

部分，您的任务是完成predict.m中的代码。predict函数将根据给定的数据集和学习的参数向量θ生成“1”或“0”预测。

在完成predict.m中的代码之后，ex2.m脚本将通过计算正确的示例百分比来报告分类器的培训精度。

*现在您应该提交解决方案。*

# 2             正则逻辑回归

在本部分练习中，您将实现规则化逻辑回归，以预测制造厂的微芯片是否通过质量保证（QA）。在质量保证期间，每个微芯片都要经过各种测试，以确保其正常工作。

假设你是这家工厂的产品经理，你有两个不同测试中一些微芯片的测试结果。从这两个测试中，您要确定是接受还是拒绝微芯片。为了帮助您做出决定，您在过去的微芯片上有一组测试结果，从中您可以建立逻辑回归模型。

您将使用另一个脚本ex2 reg.m来完成这部分练习。

## 2.1           可视化数据

与本练习前面的部分类似，plotdata用于生成如图3所示的图，其中轴是两个测试分数，正（y=1，接受）和负（y=0，拒绝）示例用不同的标记显示。

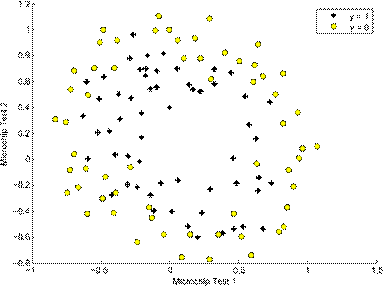


图3：培训数据图

图3显示了我们的数据集不能通过图中的直线分为正负示例。因此，逻辑回归的直接应用在这个数据集上不能很好地执行，因为逻辑回归只能找到一个线性决策边界。

## 2.2           特征映射

一种更好地适应数据的方法是从每个数据点创建更多的特性。在提供的函数mapfeature.m中，我们将把这些特征映射到x1和x2的所有多项式项中，直到第六次幂。

|  |  |  |
| --- | --- | --- |
| γ  γ  γ  γ  γ  γ | 1 X1 X2  *X*21 | γ  γ  γ  γ  γ  γ |

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image010.gif                                                                                                 

地图特征

（3）

x\*1

…阿尔法

γ

γ

x152\_\_*X*

*X*62

由于这个映射，我们的两个特征向量（两个QA测试的分数）已经转换为28维向量。在高维特征向量上训练的逻辑回归分类器具有更复杂的决策边界，在二维图中绘制时会出现非线性。

虽然特征映射允许我们构建更具表现力的分类器，但它也更容易被过度拟合。在练习的下一部分中，您将实现正则化逻辑回归来拟合数据，并亲自了解正则化如何帮助解决过度拟合问题。

## 2.3           成本函数和梯度

现在，您将实现代码来计算正则化逻辑回归的成本函数和梯度。在CostFunctionReg.m中完成代码以返回成本和渐变。

回想一下，逻辑回归中的正规化成本函数是

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image011.gif*.*

请注意，不应将参数θ0正规化。在octave/matlab中，回想一下索引从1开始，因此，不应该调整代码中的theta（1）参数（对应于θ0）。成本函数的梯度是一个向量，其中jth元素定义如下：

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image012.gifj＝0

D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image013.gifj＝1

完成后，ex2D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image001.gifm将使用θ（初始化为所有零）的初始值调用costFunctionReg函数。你应该看到成本大约是0.693。

*现在您应该提交解决方案。*

### 2.3.1          使用fminunc学习参数

与前面的部分类似，您将使用fminunc来学习最佳参数θ。如果您正确地完成了正则化逻辑回归（costFunctionReg.m）的成本和梯度，那么您应该能够使用fminunc逐步完成ex2 reg.m的下一部分来学习参数θ。

## 2.4           绘制决策边界

为了帮助您可视化这个分类器所学习的模型，我们提供了函数plotdecisionboundary.m，它绘制了（非线性）决策边界，将正负示例分开。在plot decision boundary.m中，我们通过在一个均匀分布的网格上计算分类器的预测来绘制非线性决策边界，然后绘制预测从y=0到y=1的等高线图。

在学习参数θ之后，ex中的下一步D:\document\convert_tasks\transweb\1578345_1590554\1578345.docx.files\image001.gifReg.m将绘制一个类似于图4的决策边界。

## 2.5           可选（未分级）练习

在本部分练习中，您将尝试数据集的不同正则化参数，以了解正则化如何防止过度拟合。

注意决策边界的变化，因为你改变了λ。对于一个小的λ，您应该发现分类器几乎可以正确地获取每个训练示例，但绘制了一个非常复杂的边界，从而对数据进行了过度拟合（图5）。这不是一个好的决策边界：例如，它预测X=（-0.25,1.5）处的一个点被接受（Y=1），这似乎是给定训练集的错误决策。

对于较大的λ，您应该看到一个图，它显示了一个更简单的决策边界，它仍然能够很好地区分正负。但是，如果将λ设置为太高的值，将无法获得良好的拟合，并且决策边界将不会很好地跟随数据，从而使数据下溢（图

6）

*您不需要为这些可选（未分级）练习提交任何解决方案。*

λ＝1

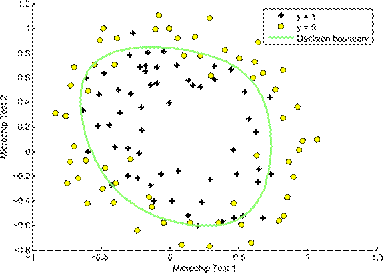


图4：具有决策边界的训练数据（λ=1）

λ＝0

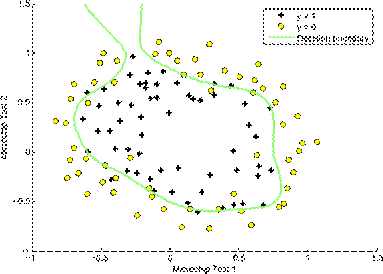


图5：无正则化（过度拟合）（λ=0）

λ＝100

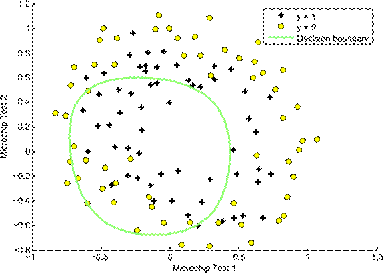


图6：太多的正则化（下装）（λ=100）

# 提交和评分

完成任务的各个部分后，请务必使用提交功能系统将您的解决方案提交给我们的服务器。下面是这个练习的每个部分的评分方式的细目。

|  |  |  |
| --- | --- | --- |
| **部分** | **提交文件** | **点** |
| 乙状结肠功能 | 乙状结肠 | 5点 |
| 计算逻辑回归的成本 | 成本函数 | 30点 |
| 逻辑回归梯度 | 成本函数 | 30点 |
| 预测函数 | 预测，M | 5点 |
| 计算正规化LR的成本 | 成本函数Reg.m | 15点 |
| 规则化LR的梯度 | 成本函数Reg.m | 15点 |
| 总分 |  | 100点 |

您可以多次提交解决方案，我们只考虑最高分数。

[〔1〕](" \l "_ftnref1" \o ")八度是matlab的免费替代品。对于编程练习，您可以自由使用八度或matlab。

[〔2〕](" \l "_ftnref2" \o ")优化中的约束通常指参数上的约束，例如，约束θ可以取的可能值的约束（例如，θ≤1）。逻辑回归没有这样的约束，因为θ可以取任何实值。