# 机器学习

# Homework3: Logistic Regression

# Logistic Regression（题1）

题1（88分）：使用逻辑回归模型来预测学生是否能被大学录取。

- 假设你是一个大学某系的管理员，你要根据每个申请者的两次考试成绩来决定该生是否被录取。你有以前申请者的历史数据（两次考试成绩和录取情况）可以用作logistic Regression的训练集。
- 你的任务是建立一个分类模型，以这两次考试的分数来估计申请人的录取的概率。
- 数据集：ex3data1.txt

# Logistic Regression（题1）

➢任务1：数据可视化。根据给出的训练集画出训练数据的图（你画出来的图应该如下图所示）。

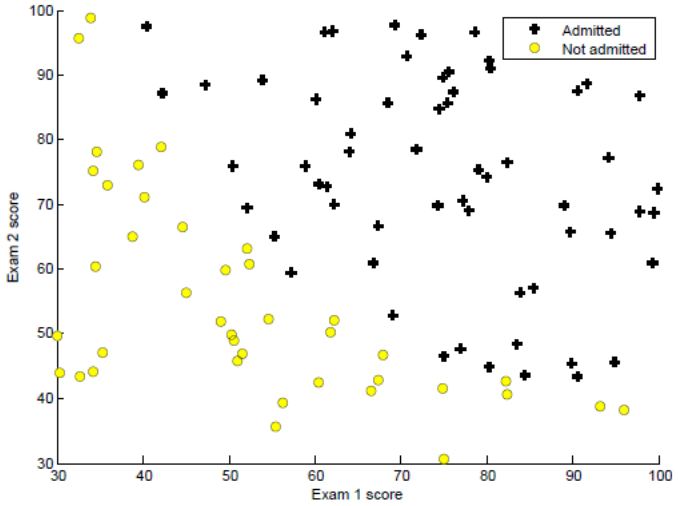


Figure 1: Scatter plot of training data

# Logistic Regression（题1）

- **任务2：**

**我们的模型是**$h_\theta(x) = g(\theta^T x)$

$$g(z) = \frac{1}{1 + e^{-z}}$$

- **任务2.1：定义一个函数sigmoid(z)，计算以上的sigmoid函数的值.**

# Logistic Regression（题1）

**我们使用的代价函数是：**

$$J(\theta)$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\ln\left(h_\theta\left(x^{(i)}\right)\right) + \left(1-y^{(i)}\right)\ln\left(1-h_\theta\left(x^{(i)}\right)\right)\right]$$

- **任务2.2：定义一个函数costFunction(X,Y, theta),(其中X是所有训练样本的特征，Y是训练样本的标签），该函数返回代价值和梯度**

$$\frac{\partial}{\partial\theta_j}J(\theta) = \frac{1}{m}\sum_{i=1}^{m}[h_\theta(x^{(i)}) - y^{(i)}]x_j{}^{(i)}$$

# Logistic Regression（题1）

**任务3：使用优化函数来求解theta。（注意，此处不需要自己写梯度下降法来求解theta）。**
**如果使用python编程，可使用scipy.optimize.fmin_bfgs函数来求解。**
**参考资料:**
**https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.fmin_bfgs.html**

**可使用如下方式调用:**

```
import scipy.optimize as opt
theta, cost, *unused = opt.fmin_bfgs(f=cost_func, fprime=grad_func,
x0=init_theta, maxiter=400, full_output=True, disp=False)
```

**其中cost_func是代价函数， grad_func是对应的梯度函数。 init_theta是theta的初始值。**
**该函数可返回最优theta值，对应的代价函数值。**

**你也可以使用其他求最优解的函数求theta。如果使用MATLAB编程，可使用fminunc函数。**

# Logistic Regression（题1）

任务4：画出训练数据集的数据图上决策边界（如下图所示）：

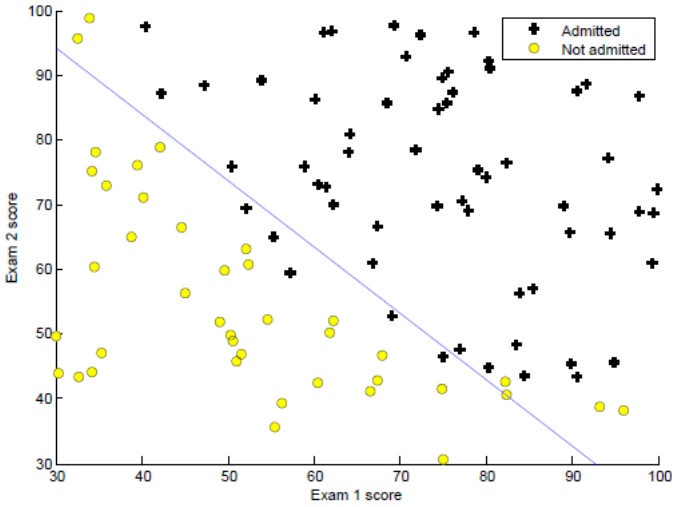


Figure 2: Training data with decision boundary

# Logistic Regression（题1）

任务5：使用求得的theta来对新数据进行预测。

例如，如果学生两名课的成绩是（45，85），用logistic Regression函数算出来的值约为0.776.

编写程序计算在训练数据集上的预测正确率。

# Logistic Regression（题2）

**题2（12分）：Regularized logistic regression（正则化逻辑回归）**

**注意：如果不完成题2，此次作业最多只能得良好；如果要拿优秀，就必须要完成题2 。**

**题2：使用正则化逻辑回归预测来自制造厂的微芯片是否通过质量检查（quality assurance (QA)）。在质量检查（QA)时，每个微芯片都要经过各种测试以确保它工作正常。**

**数据集：ex3data2.txt**

# Logistic Regression（题2）

- 任务1：数据可视化。根据给出的训练集画出训练数据的图（你画出来的图应该如下图所示）。

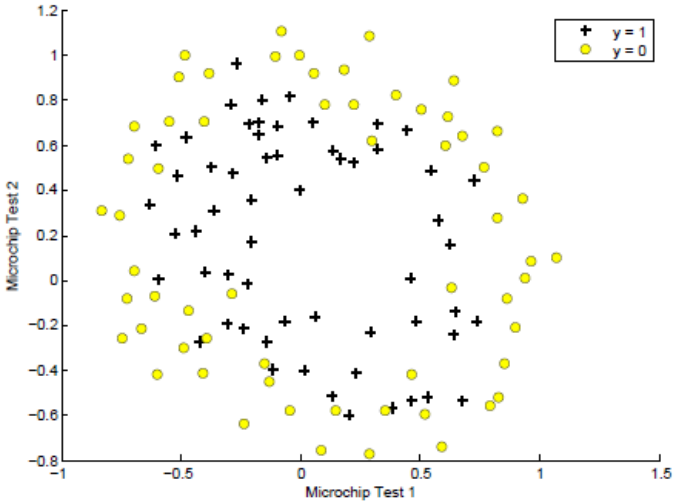


Figure 3: Plot of training data

# Logistic Regression（题2）

任务2：特征映射：使用以下特征映射把数据原始的二维特征映射成多维特征（最高次数是6次）。经过特征映射后，二维特征变成28维特征。

$$\text{mapFeature}(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \\ x_1^3 \\ \vdots \\ x_1 x_2^5 \\ x_2^6 \end{bmatrix}$$

# Logistic Regression（题2）

任务3：使用如下正则化的代价函数：

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left[-y^{(i)}\log(h_\theta(x^{(i)})) - (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] + \frac{\lambda}{2m}\sum_{j=1}^{n}\theta_j^2.$$

对应梯度函数：

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)} \qquad \text{for } j = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \left(\frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}\right) + \frac{\lambda}{m}\theta_j \quad \text{for } j \geq 1$$

定义一个函数costFunctionReg(X, Y, theta, lamda),(其中X是所有训练样本的特征，Y是训练样本的标签），该函数返回代价值和梯度.

theta初始值如果是0，则返回的代价值约为0.693，

# Logistic Regression（题2）

任务4：尝试不同的lamda值（例如λ =0，λ=1，λ=100），使用优化函数来求解theta。

（注意，此处不需要自己写梯度下降法来求解theta）。

如果使用python编程，可使用scipy.optimize.fmin_bfgs函数来求解。

# Logistic Regression（题2）

任务5：使用上一步求得的theta画出在训练数据集上的决策边界（如下图所示）：

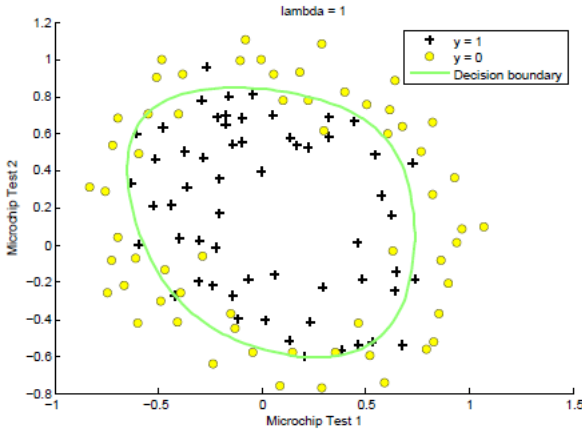


Figure 4: Training data with decision boundary ($\lambda = 1$)

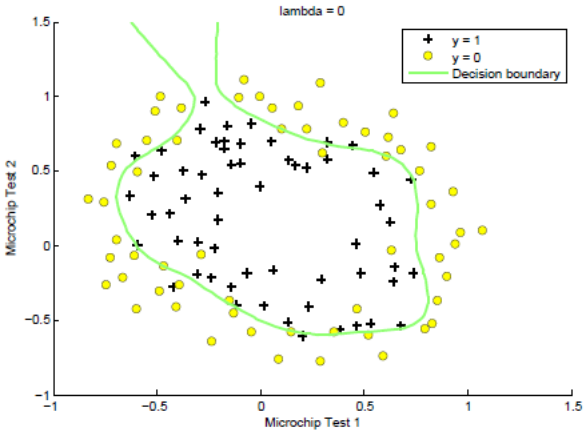


Figure 5: No regularization (Overfitting) ($\lambda = 0$)

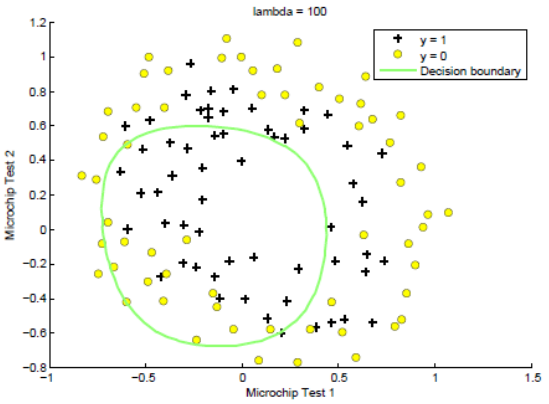


Figure 6: Too much regularization (Underfitting) ($\lambda = 100$)

# 实验报告内容

1、实验内容
　　实验要解决的问题、采用的模型或算法等
2、实验设置和实验结果
3、其它（其它你觉得需要写在实验报告中的内容）
4、实验过程中遇到的问题
5、实验心得体会。

注意事项：
1、实验报告请使用老师提供的实验模板，源代码作为单独的文件。
2、实验报告命名：完整学号_姓名_ML_project1.doc
3、建立个人文件夹放实验报告和源代码(源代码需加注释说明)，文件夹名“完整学号_姓名_ML_project1”
报告提交时间：10月28日周三下午3点前。