作业2：线性分类模型

本次作业deadline为2019.3.26，大家有两周的时间来完成此次作业。任何编程问题请提供源代码，作业有任何问题请及时联系助教。

本次作业负责助教：张威

- 1. Fisher准则的最小二乘法推导 (20p)
- 2. 二分类问题：乳腺癌数据集分类 (40p)
- 3. Logistic Regression的多类推广: Softmax Regression (40p)

1. Fisher准则的最小二乘法推导 (20p)

在某些情况下，Fisher准则可以通过最小二乘法得到。这里考虑二分类问题。假设C_1类有n_1样本，C_2类有n_2样本，两类别的均值向量如下：

$$\boldsymbol{m}_1 = \frac{1}{n_1}\sum_{i\in C_1}\boldsymbol{x}_i, \boldsymbol{m}_2 = \frac{1}{n_2}\sum_{i\in C_2}\boldsymbol{x}_i \tag{1}$$

类别间方差矩阵和类别内方差矩阵分别为：

$$S_B = (\boldsymbol{m}_2 - \boldsymbol{m}_1)(\boldsymbol{m}_2 - \boldsymbol{m}_1)^T \tag{2}$$

$$S_w = \sum_{i\in C_1}(\boldsymbol{x}_i - \boldsymbol{m}_1)(\boldsymbol{x}_i - \boldsymbol{m}_1)^T + \sum_{i\in C_2}(\boldsymbol{x}_i - \boldsymbol{m}_2)(\boldsymbol{x}_i - \boldsymbol{m}_2)^T \tag{3}$$

我们将$\frac{n}{n_1}$,$\frac{-n}{n_2}$分别作为类别C_1,C_2的目标，这里$n = n_1 + n_2$，那么误差平方和函数可以表示为：

$$E = \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{x}_i + w_0 - t_i)^2 \tag{4}$$

其中，$(\boldsymbol{x}_i, t_i)$是我们已知的点，t_i根据类别等于$\frac{n}{n_1}$或$\frac{-n}{n_2}$，我们的目标是确定$\boldsymbol{w}$和w_0。

（1）证明最优化$w_0 = -\boldsymbol{w}^T\boldsymbol{m}$，$\boldsymbol{m}$为所有样本的均值。

（2）推导最优化$\boldsymbol{w}$服从下方等式：

$$(S_w + \frac{n_1 n_2}{n}S_B)\boldsymbol{w} = n(\boldsymbol{m}_1 - \boldsymbol{m}_2) \tag{5}$$

（3）通过公式(5)推导出$\boldsymbol{w} \propto S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$，这意味着我们得到了与Fisher准则相同的形式。

2. 二分类问题：乳腺癌数据集分类 (40p)

请使用Logistic回归和Fisher线性判别设计分类器，实现乳腺癌数据集（数据来自于UCIML，附件为breast-cancer-wisconsin.txt，数据经过简单处理）的分类 。我们使用的威斯康辛州乳腺癌诊断数据集是699×11维的矩阵，11维的特征信息如下：

Attribute Domain

```
1. Sample code number            id number
2. Clump Thickness               1 - 10
3. Uniformity of Cell Size       1 - 10
4. Uniformity of Cell Shape      1 - 10
5. Marginal Adhesion             1 - 10
6. Single Epithelial Cell Size   1 - 10
7. Bare Nuclei                   1 - 10
8. Bland Chromatin               1 - 10
9. Normal Nucleoli               1 - 10
10. Mitoses                      1 - 10
11. Class:                       (0 for benign, 1 for malignant)
```

我们的目标是实现对良性和恶性肿瘤的分类和预测。请随机使用75%的数据作为训练集，25%数据作为测试集，给出两种算法的测试集准确率。
要求：Logistic回归可以调用函数，Fisher线性判别请**自行编写程序实现**。

3. Logistic Regression的多类推广: Softmax Regression (40p)

在课上我们已经学习了如何使用Logistic Regression进行二分类。请大家阅读Softmax Regression并回答以下问题。

（1）Softmax Regression是线性分类器还是非线性分类器？请说明你的理由。

提示：可以从位于分界面上的点入手分析。

（2）在附件中我们给出了10个人的人脸图像（数据来自于VGGface2，附件为Pictures.rar）。请使用Softmax Regression设计分类器，实现以下要求。

- 请**随机**取出2个人的图像，75%作为训练集，25%作为测试集，给出Softmax的测试集正确率。
- 请**随机**取出5个人的图像，75%作为训练集，25%作为测试集，给出Softmax的测试集正确率。
- 请**随机**取出7个人的图像，75%作为训练集，25%作为测试集，给出Softmax的测试集正确率。
- 请使用所有人的图像，75%作为训练集，25%作为测试集，给出Softmax的测试集正确率。
- 以上的测试中你的正确率是如何变化的，总结变化并给出合理解释。

提示：

本题目提供的图片是彩色图片，请大家在进行分类前**将图片转化为灰度图片**，即大家在分类问题中处理的图片是48像素×48像素的灰度图片。如果你读入程序的图片是3×48×48，请按照以下的提示对图片重新进行处理。

在Python中，你可以使用下面的方法来进行转换。Matlab下请参考rgb2gray函数。你也可以手动编程实现，原理请参考该问题。其它语言请自行查询。

```
# 直接调用函数，在读取图片时将图片读取为灰度图片
from skimage import io
img = io.imread('image.png', as_grey=True)
```

在python中，softmax的实现函数是sklearn中的LogisticRegression。设置multi_class参数为"multinomial"即可以使用softmax函数对每一类的概率进行预测。