作业 2-Logistic 回归和 Fisher 线性判别

1. 推导[40%]

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

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

类别间方差矩阵和类别内方差矩阵为：

$$S_B = (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T, S_w = \sum_{i\in C_1}(\mathbf{x}_i - \mathbf{m}_1)(\mathbf{x}_i - \mathbf{m}_1)^T + \sum_{i\in C_2}(\mathbf{x}_i - \mathbf{m}_2)(\mathbf{x}_i - \mathbf{m}_2)^T \tag{2}$$

我们将n/n_1、$-n/n_2$ 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：

$$E = \frac{1}{2}\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i + w_0 - t_i)^2 \tag{3}$$

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

(1)证明最优 $w_0 = -\mathbf{w}^T\mathbf{m}$；

(2)推导最优$\mathbf{w}$ 服从下方等式：

$$(S_w + \frac{n_1 n_2}{n}S_B)\mathbf{w} = n(\mathbf{m}_1 - \mathbf{m}_2) \tag{4}$$

(3)通过公式(4)推导出$\mathbf{w} \propto S_W^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$，这意味着我们得到了与 Fisher 准则相同的形式。

2. 编程题[60%](请在作业中提供程序源代码)

请使用 logistic 回归和 Fisher 线性判别两种方法设计分类器实现对威斯康辛州乳腺癌诊断数据集的分类。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: (2 for benign, 4 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 70%的数据作为训练集,剩余的 30%作为测试集，给出两种方法的测试集准确率。

提示：

1. Matlab 中没有自带的 logistic 回归函数，在作业附件给出了参考的函数，供参考使用；

2. 请自己编写实现 Fisher 线性判别的程序。

3.数据集网站，供感兴趣的同学了解。

http://mlr.cs.umass.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)