用身高体重进行性别分类的实验二

2013011570 自 31 唐静娴

一、实验内容

用 dataset3 作为训练数据，用 dataset4 作为测试数据，进行性别分类实验。

◆ **采用特征组合：**

a) 10 个特征都用 b) 只使用其中的第 3 列和第 5 列特征

◆ **试验的训练样本：**

a) 从 dataset3 中任选 20 个训练样本（男女各 10 例） b) dataset3 中全部训练样本

◆ **试验的分类器方法：**

a) 最小错误率贝叶斯分类器 b) Fisher 线性判别 c) 线性 SVM d) 采用 BP 算法的 MLP 神经网络

二、实验结果

2.1 测试错误率汇总表

训练样本数	特征数	Bayes	FLD	Linear SVM	MLP
10+10	10	0.1433	0.2164	0.1771	0.3232
	2	0.1128	0.1264	0.1235	0.2134
469+485	10	0.1585	0.1006	0.1220	0.0823
	2	0.1250	0.1159	0.1220	0.1006

注：随机挑选 20 个训练样本的所得测试错误率均为 30 次测试后的**平均错误率**

2.2 部分结果截图

- 2 个特征 and 全部训练样本 and 最小错误率贝叶斯分类器

```
err     0.1250
err_f   2
err_m   39
```

- 2 个特征 and 全部训练样本 and FLD 分类器

```
err     0.1159
err_f   3
err_m   35
```

- 10 个特征 and 20 个训练样本 and FLD 分类器（单次运行结果）

```
err     0.1799
err_f   5
err_m   54
```

- 10 个特征 and 20 个训练样本 and 线性 SVM 分类器（单次运行结果）

err	0.1585
err_f	22
err_m	30

- 10 个特征 and 全部训练样本 and 线性 SVM 分类器

err	0.1220
err_f	3
err_m	37

- 2 个特征 and 全部训练样本 and MLP 神经网络

err	0.1067
err_f	4
err_m	31

三、程序说明

- **function [male,female] = preprocess(file_name,lessFeature,lessNum)**

	参数名称	参数说明	参数类型
输入	file_name	即将进行预处理的文件名称	string
	lessFeature	是否只取 2 组特征	bool
	lessNum	是否只挑选 20 个训练样本	bool
输出	male	分离出的男性数据	M*2(10) double
	female	分离出的女性数据	N*2(10) double

【随机挑选方法说明】

1° 从 dataset3 中先分离出男性数据组 male 和女性数据组 female，大小分别为 N 和 M；

2° 利用 MATLAB 的 randperm 函数生成 1~M 的随机排列 index1,以及 1~N 的随机排列 index2；

3° female(index1(1:10),:)即为随机挑选的 10 个女性训练样本，male(index2(1:10),:)即为随机挑选的男性样本

【挑选出的样本示例】

1° 10 组特征

男生组：

	1	2	3	4	5	6	7	8	9	10
1	282.2400	43.5600	66	23.3800	168	168	52	173.8000	168.8600	191.4900
2	306.2500	36	60	19.5900	175	177	52	172.2200	165.2200	164.1200
3	309.7600	54.7600	74	23.8900	176	160	69	167.9900	161.7700	197.0900
4	272.2500	36	60	22.0400	165	175	70	167.3000	167.9600	175.9600
5	306.2500	56.2500	75	24.4900	175	176	67	168.9600	163.7300	165.2400
6	331.2400	49	70	21.1300	182	168	75	179.9400	167.1900	153.3500
7	306.2500	42.2500	65	21.2200	175	171	84	155.4600	171.2400	171.6600
8	313.2900	46.2400	68	21.7100	177	160	53	175.0900	170.2600	183.2700
9	282.2400	32.4900	57	20.2000	168	158	67	167.3800	169.1200	177.0800
10	338.5600	49	70	20.6800	184	158	66	159.4500	168.1400	149.8800

（注：每一列表示一组特征，每一行表示一个样本）

女生组：

1	2	3	4	5	6	7	8	9	10
256	30.2500	55	21.4800	160	156	40	168.4000	173.5800	189.7000
275.5600	25	50	18.1400	166	166	61	172.4800	177.6400	173.6700
292.4100	30.2500	55	18.8100	171	173	65	180.1800	166.3700	202.4500
272.2500	42.2500	65	23.8800	165	160	70	186.8500	172.6200	152.6300
282.2400	31.3600	56	19.8400	168	172	55	171.4300	168	194.0100
256	20.2500	45	17.5800	160	171	65	155.9900	183.5500	194.2800
246.4900	25	50	20.2800	157	175	63	146.6500	169.7700	170.7400
256	29.1600	54	21.0900	160	158	45	167.5200	168.1300	172.1200
262.4400	49	70	26.6700	162	149	49	175.4100	176.0600	185.9600
289	30.2500	55	19.0300	170	171	65	153.4500	166.6300	173.4900

（注：每一列表示一组特征，每一行表示一个样本）

2° 2 组特征

男生组：

1	2	3	4	5	6	7	8	9	10
70	66	62	68	81.7000	61	60	70	76	75
185	176	173	178	200	170	178	178	184	179

（注：每一行表示一组特征，每一列表示一个样本）

女生组：

1	2	3	4	5	6	7	8	9	10
49	50	47	50	50	46	49	49	45	55
160	163	155	163	159	160	158	158	161	160

（注：每一行表示一组特征，每一列表示一个样本）

➢Bys_detection.m

【方法描述】

根据所需要的不同特征组数、不同样本数，调用 preprocess 函数并更改其输入参数，从 dataset3 中得到训练样本；

假设所有特征均服从正态分布，先验概率各为 0.5。与作业 1 采用相同的方法，若

$$\frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_1)}{P(w_2)} = \frac{0.5}{0.5}$$

则x ∈ w_1，否则x ∈ w_2。其中$p(x|w_1)$为多元正态分布的类条件概率密度。

➢ Fisher.m

【方法描述】

根据所需要的不同特征组数、不同样本数，调用 preprocess 函数并更改其输入参数，从 dataset3 中得到训练样本；

设计 Fisher 线性判别分类器$g(X) = w^T X + w_0$，其准则为：选择投影方向，使得投影后类间离散度尽可能大，而类内离散度尽可能小。采用严格的数学描述为：

$$J_F(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

$$\mathbf{w}^*: \ \max_{\mathbf{w}} J_F(\mathbf{w})$$

其中，S_b为样本类间离散度矩阵，S_w为总类内离散度矩阵。最优投影方向为：

$$\mathbf{w}^* = S_w^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

阈值选择为：

$$w_0 = \frac{1}{2}(\tilde{m}_1 + \tilde{m}_2) + \frac{1}{N_1 + N_2 - 2}\ln\frac{P(\omega_1)}{P(\omega_2)}$$

根据以下规则对测试集中数据进行分类：

$$\begin{cases} g(X) > 0, & \text{决策} X \in w_1 \\ g(X) < 0, & \text{决策} X \in w_2 \\ g(X) = 0, & \text{可任意分类} \end{cases}$$

- **Svm.m**

【方法描述】

根据所需要的不同特征组数、不同样本数，调用 preprocess 函数并更改其输入参数，从 dataset3 中得到训练样本；

线性 SVM 法在于求样本集的最优分类超平面，构造不等式约束下的优化问题，利用拉格朗日法求解。实验中直接使用 Matlab 自带的 svmtrain 和 svmclassify 函数求取。得到的最优超平面的权值向量为支持向量的加权和。

- **BP.m**

【方法描述】

根据所需要的不同特征组数、不同样本数，调用 preprocess 函数并更改其输入参数，从 dataset3 中得到训练样本；

使用 Matlab 自带的神经网络工具包，自行选择参数。**本次实验所使用的参数参考了自 21 朱正达（2012011539）的报告**，最终使用的网络结构为 2 层感知器组成的神经网络，第一层传递函数为 logsig，第二层为 purelin。允许最大训练步数为 500 步，训练目标最小误差 0.01，学习速率 0.05。

- **程序来源与软件版本说明**

除前文中特别说明的地方以外，所有程序均为自行编写，使用的软件为 MATLAB_2014a。其中 Svm.m 使用了 MATLAB 的支持向量机工具包，BP.m 使用了 MATLAB 的神经网络工具包。

四、分析与体会

从分类结果汇总可以看出，基本上训练样本数量越多，分类错误率相应越小，这与我们的直觉是一致的。错误率最大的情况是 “10 组特征 and 20 个训练样本”，因为此时训练集所能提供的信息太少，仅靠这些信息就得到 10 组特征的正确分类显然是不切实际的。

从 MLP 的多次测试结果来看，当只挑选 20 个训练样本时，得到的测试错误率变化范围很大。这说明当训练样本数目较少时，BP 神经网络对于训练样本的依赖性是比较高的，选择好的训练集对于得到较为理想的分类结果来说很有必要。