模式识别第三次作业报告

2013011570 自31 唐静娴

一、实验内容

1. 数据dataset3.txt和dataset4.txt包括10个特征，用dataset3.txt做训练样本，采用适当的特征选择方法选择1~3个特征，采用适当的分类方法进行分类器设计，考察训练错误率；将设计出的分类器应用到dataset4.txt上，考察测试错误率。结合前两次作业对特征选择和分类进行讨论
2. （选作）用某种K-L变换对dataset3.txt的10维特征进行变换，提取2维新特征进行分类器设计，对dataset4.txt也提取同样的2维特征，测试分类器，与本次和前两次实验的结果进行比较分析。

二、实验方法

1. 本次实验采用的方法列举如下：

    a. **特征的评价准则：**类内类间距离；t统计检验
    b. **特征的选取方法：**穷举法；单独最优特征组合
    c. **分类器设计：**Fisher线性分类器

2. 方法说明

    - **类内类间距离判别准则**
      J_1 ~J_5 描述了类内类间距离之间的关系，其计算公式如下：

$$J_1 = tr(S_w + S_b)$$
$$J_2 = tr(S_w^{-1} + S_b)$$
$$J_3 = ln\frac{|S_b|}{|S_w|}$$
$$J_4 = \frac{trS_b}{trS_w}$$
$$J_5 = \frac{|S_b - S_w|}{|S_w|}$$

      其中S_b描述了类间离散度，S_w描述了类内离散度。J_i的值越大，说明此时选择的特征下两类区分度越大。

    - **t 统计检验判别准则**
      我们不妨假设两类样本都服从正态分布，且方差相同。设两类分别有m个和n个样本，t-检验的统计量为

$$t = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

它服从自由度为n+m-2的t分布。其中 $s_p^2 = \frac{(n-1)S_x^2+(m-1)S_y^2}{m+n-2}$，表示总体的样本方差。我们可以根据每一个特征的t-检验统计量的值来推断“在该特征上两类样本均值是否有明显差异”，t值越大，我们越倾向于接受这一假设。从而根据t值大小便可以完成一组或多组特征的选取。

- **Fisher线性判别法**
  上一次作业中已经使用过，不再赘述。

- **K-L变换**
  这里采用的是用于监督模式识别的K-L变换。方法与教材一致：
  （1）计算总类内离散度矩阵S_w；
  （2）用S_w作为产生矩阵进行K-L变换，求解本征值λ_i和对应的本征向量$\mathbf{u_i}$,得到一组新特征$y_i = \mathbf{u_i^T x}$;
  （3）计算新特征的分类性能指标$J(y_i)$，筛选出得分最高的两个作为新的2维特征

三、结果与讨论

1. **特征选取结果**
   a. 类内类间距离判别准则
   选取的结果为：

选取的特征	J1	J2	J3	J4	J5
1 个	1	5	5	5	8
2 个	1, 10	4, 5	3, 5	1, 5	6, 8
3 个	1, 2, 10	2, 3, 4	1, 2, 3	1, 4, 5	6, 8, 9

   运行结果示例：

```
How many features you want to choose: 2
Please input 1~5 in represent of J1~J5: 5
Recommended feature(s): 6  8
```

   b. t统计检验判别准则
   10个特征按照t值大小递减排序为

5	1	3	2	4	8	7	6	9	10
35.2586	34.8994	31.7094	28.1617	16.3223	0.0539	-0.6537	-0.6773	-1.2504	-2.2725

   因此可选择的前3列特征为第5列，第1列，第3列。根据单独最优特征组合的原理，如果分别使用1~3个特征，将使用第5列；第1，5列；第1，3，5列特征。
   c. K-L变换
   计算得到的变换矩阵为

$$U = \begin{bmatrix} 0.874 & -0.0079 \\ 0.3230 & -0.5893 \\ 0.2395 & 0.6889 \\ 0.0172 & 0.4102 \\ 0.2549 & 0.0992 \\ -0.0214 & -0.0036 \\ 0.0139 & -0.0049 \\ 0.0185 & -0.0026 \\ -0.0115 & 0.0018 \\ 0.0203 & 0.0002 \end{bmatrix}$$

新的2维特征 $Y = U^T X$，其中X为原来的10维特征。

2. **错误率**

根据上一小节的三种特征选择方法，一共得到了15种特征组合，分别考察它们在Fisher线性分类器下的训练错误率和测试错误率，得到的结果如下：

选取特征	训练错误率	测试错误率	选取特征	训练错误率	测试错误率
1	0.13836	0.18293	1，2，3	0.08595	0.09451
5	0.13836	0.18293	1，2，10	0.09644	0.13415
8	0.50629	0.44817	1，3，5	0.09120	0.10366
1，5	0.12055	0.10976	1，4，5	0.09224	0.10366
1，10	0.13103	0.14634	2，3，4	0.08491	0.09756
3，5	0.08491	0.11585	6，8，9	0.48428	0.57012
4，5	0.08700	0.11280	K-L 方法	0.09539	0.10671
6，8	0.48952	0.57622			

3. **结果分析**

首先关注选取1组特征时的错误率。可以看到，第8列特征的训练错误率和测试错误率都达到了50%左右，说明该模型是几乎无效的。再看其它包含了第8列特征的特征组合（[6,8]和[6,8,9]），错误率也相当之高。单独考察第6列特征和第9列特征，发现它们的表现与第8列特征相当，错误率都超出了容忍范围。

```
Please input the feature(s) you want to use:
6
The total training error is 0.48428
The total testing error is 0.57317
```

```
Please input the feature(s) you want to use:
9
The total training error is 0.48323
The total testing error is 0.57012
```

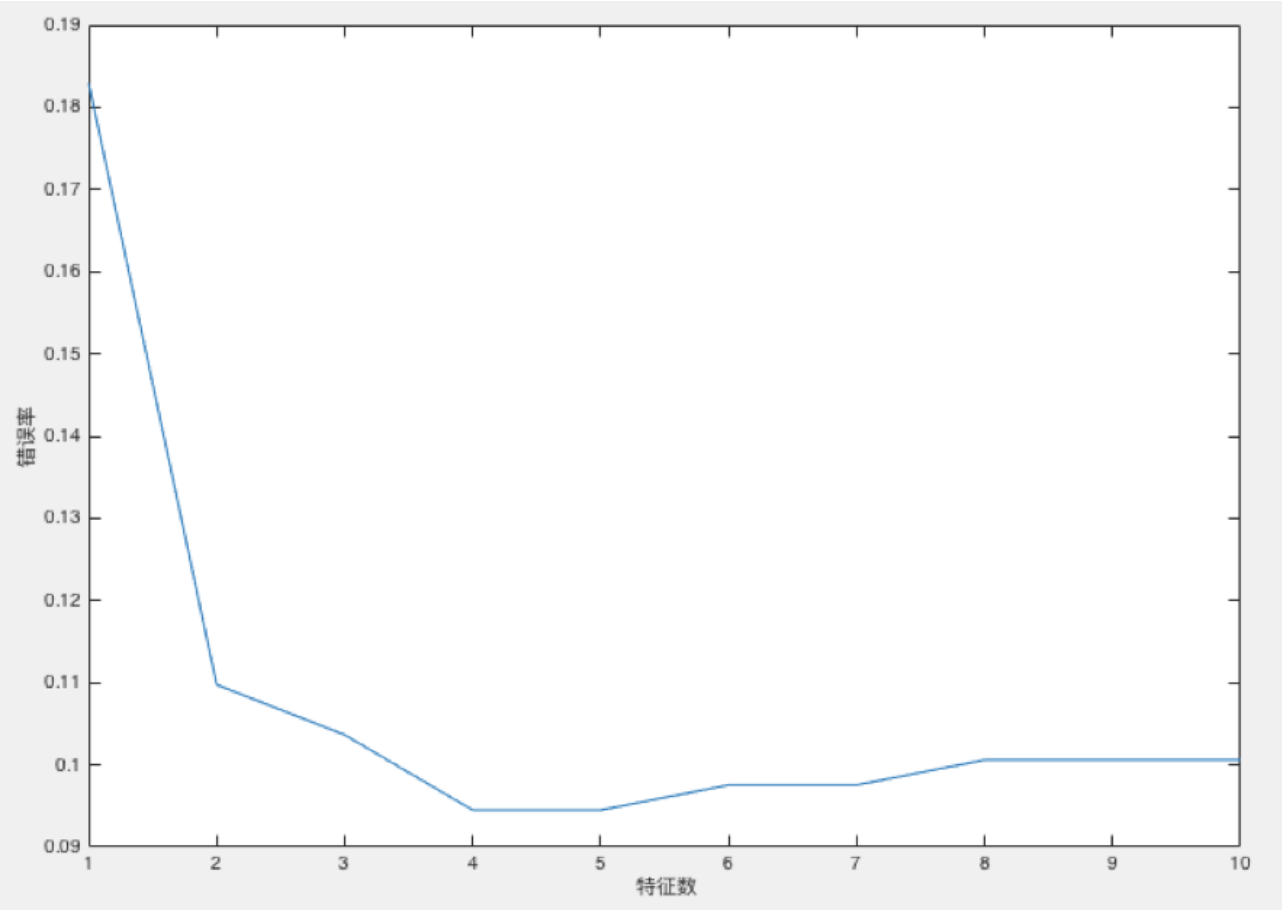
这几个特征组合都是由J_5选取出来的，并且在3次选择中，J_5都选择了第8列特征。这给我们之后的工作提供了一点启示，在使用类内类间距离作为判别准则时，最好将J_1~J_5的选取结果都计算出来，否则可能会得到具有欺骗性的结果。以下的讨论中我们将不考虑J_5选取的特征组合。

在选取1组特征时，J_2, J_3, J_4和t统计检验法都投票给了第5列特征。因此可以断定在此条件下，在

第5列特征上两类样本的差异最明显。

在选取2组特征时，除特征组合(1,10)表现略为逊色以外，其余组合差异并不大；选取3组特征时，所有组合的表现都比较优异。同时我们也可以发现，似乎随着特征数的增加，两类错误率都有所下降。

以t统计检验作为特征选取准则为例，我们可以得到如下的曲线，



当特征数由1个增加到4个时，错误率有着显著的下降，之后随着特征数的增加错误率有所上升，但仍然不比特征数小于4个时的表现差。出于计算成本和分类错误率的综合考虑，选择3组特征可能是一个折衷的办法。

K-L变换可以有效地消除特征之间的相关关系，并且实现了降维，从长远来说节省了数据存储空间。根据两种错误率来看，K-L变换的效果也是非常好的。

四、实验说明

1. 程序来源：全部自行编写
2. 程序代码说明：见程序报告