模式识别第三次作业实验报告

2013012245 基应 31 白可

实验目的：这回的实验目的是让我们学会在若干组特征中选择比较好的特征或特征的组合，从而达到较好的训练和分类的效果

数据集：954 个样本，每个样本有 10 个特征
训练集：328 个样本，每个样本有 10 个特征

实验设计：我的实验主要分为四个主要部分

一、对单个特征进行“基于类内间距的可分性判据”。

从而获得不同特征的类间离散度矩阵S_B和类内离散度矩阵S_w。计算每个特征的$\frac{S_b}{S_w}$，并对他们进行排序，从而衡量出每个特征的优越性。

分别计算：

$$S_b = \sum_{i=1}^{c} P_i(\overline{m}_i - \overline{m})(\overline{m}_i - \overline{m})^T \qquad S_w = \sum_{i=1}^{c} P_i E_i[(X - \overline{m}_i)(X - \overline{m}_i)^T]$$

并利用

$$J = \frac{tr\, S_w}{tr\, S_b},$$

衡量该特征的特性。

二、对单个特征进行“基于概率分布的可分性判据”。

显然在一般情况下由于概率分布本身的复杂形式，以上这些基于概率分布的距离相当复杂。这些判据在概率分布具有某种参数形式，尤其是正态分布时可以得到进一步简化。下面讨论两类别正态分布时散度判据的表达式。

设两类别分别表示为 $\omega_i \sim N(\mu_i, \Sigma_i), \omega_j \sim N(\mu_j, \Sigma_j)$ 则

$$p(X \mid \omega_i) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_i \mid^{1/2}} \exp[-\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1}(X - \mu_i)]$$

$$p(X \mid \omega_j) = \frac{1}{(2\pi)^{d/2} \mid \Sigma_j \mid^{1/2}} \exp[-\frac{1}{2}(X - \mu_j)^T \Sigma_j^{-1}(X - \mu_j)]$$

对数似然比

$$l_{ij} = \frac{1}{2}\ln \mid \frac{\Sigma_j}{\Sigma_i} \mid -\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1}(X - \mu_i) + \frac{1}{2}(X - \mu_j)^T \Sigma_j^{-1}(X - \mu_j)$$

利用矩阵迹的性质 $A^TB=tr(BA^T)$，其中 A、B 表示向量，上式可改写成

$$l_{ij}=\frac{1}{2}\ln|\frac{\sum_j}{\sum_i}|-\frac{1}{2}tr[\sum_i^{-1}(X-\mu_i)(X-\mu_i)^T]+$$

$$\frac{1}{2}tr[\sum_j^{-1}(X-\mu_j)(X-\mu_j)^T]$$

$$I_{ij}=\frac{1}{2}\ln|\frac{\sum_j}{\sum_i}|+\frac{1}{2}tr[\sum_i(\sum_j^{-1}-\sum_i^{-1})]$$

$$+\frac{1}{2}tr[\sum_j^{-1}(\mu_i-\mu_j)(\mu_i-\mu_j)^T]$$

由 J_D 的定义 $J_D=I_{ij}+I_{ji}$ 得

$$J_D=\frac{1}{2}tr[\sum_i^{-1}\sum_j+\sum_j^{-1}\sum_i-2I]$$

$$+\frac{1}{2}(\mu_i-\mu_j)^T(\sum_i^{-1}+\sum_j^{-1})(\mu_i-\mu_j)$$

三、通过调用 matlab 库函数使用“ttest 以统计检验作为可分性判据”，获得“单独最优特征的组合”。

各特征按单独使用计算其判据值，然后取其前 d 个判据值最大的特征作为最优特征组合。这种做法的问题在于即使各特征是独立统计的，也不一定得到最优结果。但如果可分性判据可写成如下形式

$$J(X)=\sum_{i=1}^{D}J(x_i)$$

$$J(X)=\prod_{i=1}^{D}J(x_j)$$

则用这种方法可以选出一组最优的特征来。

四、运用特征选择的次优算法之“顺序前进法”(SFS)

首先计算每个特征单独进行分类的判据值，并选择其中判据值最大的特性，作为入选特征。然后每次从未入选的特征中选择一个特征，使得它与已入选的特征组合在一起时所得的 J 值为最大，直到特征数增至 d 个为止。

我为什么要有这样的选择？

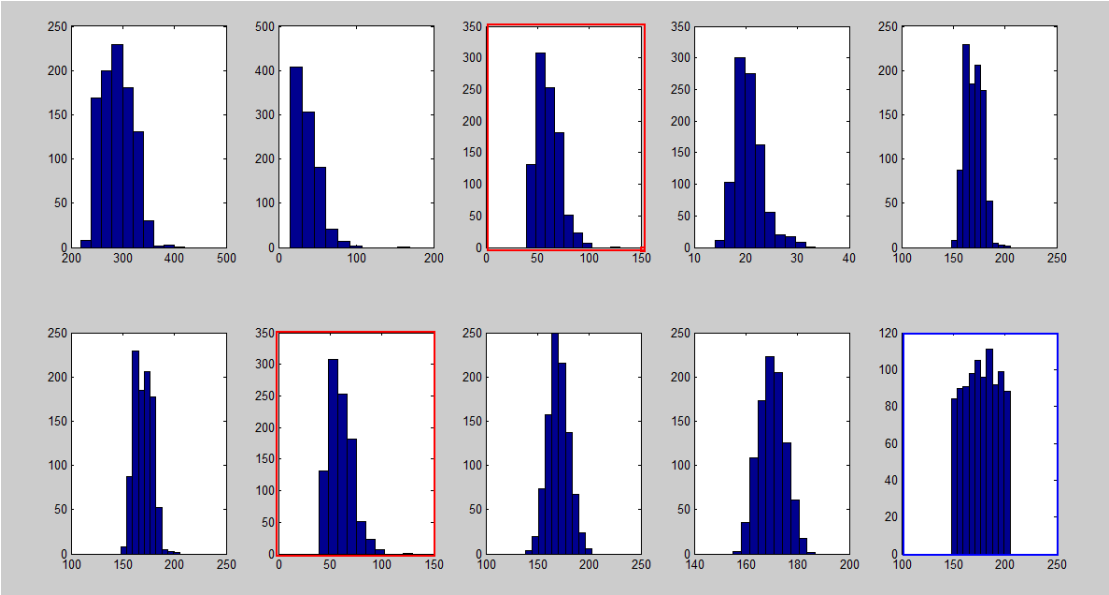
课本上介绍了很多种判据，我希望了解在不同的判据下，得到的对于特征筛选的结果是什么样的。因此我选择了基于“概率”和“距离”，这两个不同角度出发，计算得到的判据。希望比较这两种方式的差异。

后面使用顺序前进法。这种方法更加节省时间，更加快速。由于要求 1_3 个特征，因此，比起顺序后退法，它更快地找到最终的答案。

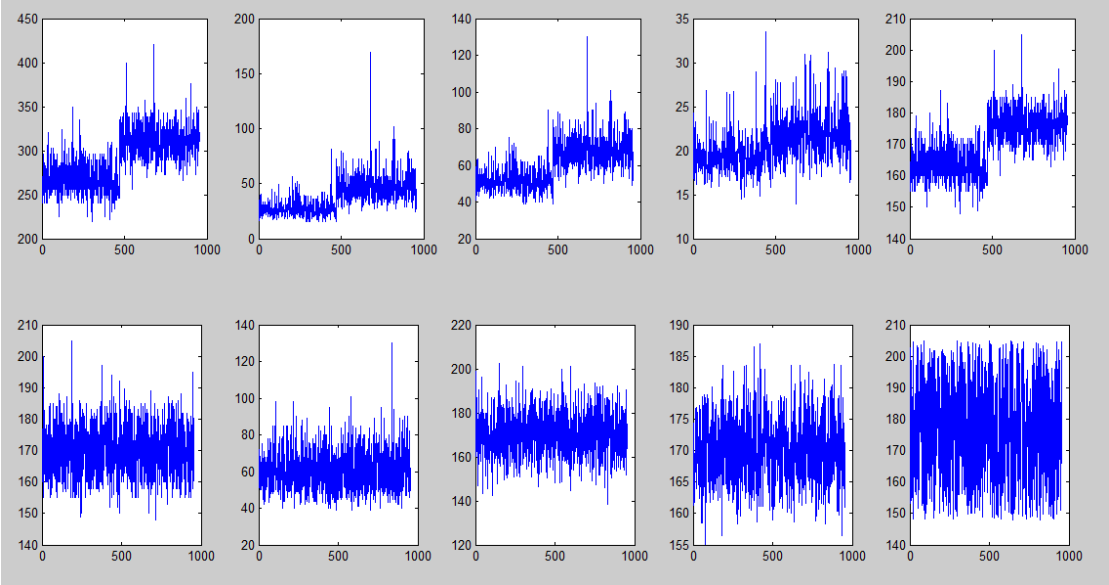
分类方法：贝叶斯判别法。在第二次实验中，它的分类正确率是最低的。因此，我想试图通过调整训练的特征数目提升它的正确率，看一看它的正确率不高到底是因为自身的其他因素还是因为特征选择的因素。而且贝叶斯判别法，它的概率密度函数对应着神经科学中的

tuning curve，也值得我更进一步地探索。

我观察了一下各个特征的直方图。



同时观察了一下每组的数据



将这张图片从左至右，从上到下分别记为 1~10，方便之后讨论

虽然这张图已经让我们观察到 1——5 的特征十分明显，但是，我们不能仅仅将这些特征组合在一起就断言这是最好的组合，还需要进一步假设验证。

特征选择本质上是一个组合优化问题，求解组合优化问题最直接的方法就是搜索，理论上可以通过穷举法来搜索所有可能的特征组合，选择使得评价标准最优的特征子集作为最后的输出，但是 n 个特征的搜索空间为 2n，穷举法的运算量随着特征维数的增加呈指数递增，实际应用中经常碰到几百甚至成千上万个特征，因此穷举法虽然简单却难以实际应用。

实验结果：

一、利用“基于类内间距的可分性判据”，将判据选择为 J = $\frac{tr\,S_w}{tr\,S_b}$，

	1	2	3	4	5	6	7	8	9	10
1	5	3	1	2	7	4	6	9	10	8
2	0.5494	0.5950	0.5980	0.8460	0.8734	1.2699	2.0863	7.0851	188.0566	209.3032

第三行计算的是从大到小排列的 J 值，第二行为他们所对应的特征号。我们可以选择前 3 个特征（尽管理论上这样并不是最优的选择，但我们希望可以进行实验和尝试）

计算得到：

特征	训练错误率-bayesian	测试错误率_bayesian	训练错误率-fisher	测试错误率_fisher
5	12.16%	10.98%	13.8%	18.29%
5:3	9.2%	12.5%	12.05%	10.9%
1:3:5	9.54%	13.4%	9.12%	13.4%

讨论：根据选择的特征以及他们对应的图，可以看到，编号为 6~10 的特征，类间特征不明显，类内也较为分散，因此最后计算得出的值较大。Fisher 的效果与 bayesian 近似，主要由他们不同的分类方式决定，但是它在三种特征的训练错误率最小。这种途径筛选出的特征，在筛选特征的时候就考虑到了类间距，这与 fisher 的初衷一致。

二、对单个特征进行“基于概率分布的可分性判据”。

	1	2	3	4	5	6	7	8	9	10
1	2	5	3	1	4	10	7	9	8	6
2	6.2074	5.2278	5.2277	5.2119	1.2869	0.0305	0.0154	0.0081	0.0050	0.0025

第三行计算的是从大到小排列的 J 值，第二行为他们所对应的特征号。

特征	训练错误率-bayesian	测试错误率_bayesian	训练错误率-fisher	测试错误率_fisher
2	15.6%	20.42%	15.6%	20.4%
2：5	9.75%	13.11%	9.43%	13.41%
2:3:5	11.4%	16.7%	8.60%	9.45%

讨论：同上，第三行计算的是从大到小排列的 J 值，第二行为他们所对应的特征号。这是后，选择出来的三个特征分别是 2,5,3.考虑到我们这里运用的是基于概率分布分可分性判据，“散度”，目的是衡量两类密度是否有交叠，交叠少的得出的 J 值更大，交叠较多的得出的 J 较小。

选择 2,5,3 根据图片可以看出，实际上，这三个特征也是交叠最少的三个特征。

但是看到，这种情况下，特征数目较少时，似乎最优的“单独特征”的组合并没有达到最好的效果，因此，我们需要尝试其他方法，寻找出“组合最优特征”。

并且，这种方式是假设分布式一个正态分布。对于不是正态分布的样本，不能很好地适应，但是在这道题中，根据我们先前的观察，

三、通过调用 matlab 库函数使用“ttest 以统计检验作为可分性判据”，获得“单独最优特

征的组合”。

	1	2	3	4	5	6	7	8	9	10
1	5	1	3	2	4	10	9	6	7	8
2	35.2812	34.9663	31.9065	28.4416	16.3839	2.2750	1.2498	0.6771	0.6546	0.0539

特征	训练错误率	测试错误率
5	12.16%	10.98%
1,5	13.8%	18.3%
1:3:5	9.54%	13.4%
1:3：5:2	11.4%	16.7%

我们可以看到，这种情况下，对比试验二，筛选出的前五组特征相似，后面五种略有不同，这种基于统计的方法实际上也是检验的整个样本的总体样本方差和分布。

四、运用特征选择的次优算法之“顺序前进法”(SFS)

在这里我使用了两种不同的途径

1） “基于类内间距的可分性判据”

2） “基于概率分布的可分性判据”

		特征编号	训练错误 bayesian	测试错误 bayesian	训练错误率-fisher	测试错误率_fisher
1）	1	5	12.16%	10.97%	13.83%	18.29%
	2	5|4	9.44%	12.50%	12.05%	10.97%
	3	5|4|3	9.33%	13.40%	9.12%	10.37%
2）	1	2	15.60%	20.40%	13.83%	18.29%
	2	2|3	17.40%	23.70%	12.05%	10.97%
	3	2|3|4	11.80%	16.77%	9.1%	10.36%

讨论：这里我们使用了顺序前进法。对数据无要求，比单独最优特征组合要宽泛，但是可能会选择出不太好的特征的组合。

其他的讨论

一、关于归一化。在实验的最开始，我一直在考虑是否需要使用“归一化”。后来我发现不一定每种方法都要使用归一化。在计算基于“类内间距的可分性判据”时，如果选择的平方距离判据是 $J = \mathrm{tr}(S_W + S_b)$。在某些情况下就需要考虑归一，否则因为特征本身的值较大可能造成比较大的影响。但本实验中，我使用的两种方法无需归一，前者基于距离的是相除关系，分子分母相除，归一化实际没有什么用。后者是算概率分布判据，实际操作的对象是概率，已经归一好，因此，本题中不需要再归一。

二、关于需要满足的条件。看书的时候发现，这里运用的很多方法都是有先决条件的，比如需要满足原函数是正态分布的假设、需要“特征增多时判据值不会减少”等等。具体已经在每一个案例的讨论中说明了就不再详细阐述，这也是未来的做相关实验需要考虑的因素。

三、根据本实验得到的结果来看，使用关于“类内和类间距离”的判据得到了较好的效果。其中一个重要因素是：基于概率的判据首先假设原分布式一个正态分布，但是观察样本我们能够发现，部分特征其实并不是“正态分布”，这可以解释为什么它的错误率相较于前者更高。

四、与原来实验的比较

训练样本数	特征数	1
469+485	10	0.095(训)/0.159
	2	0.092(训)/0.125

由于我的分类方法选择的是“bayes 决策”，因此，我只列出了上次实验的 bayes 决策部分。根据这次的实验结果可以看到，选择 3：5 特征时，正确率达到了最高，错误率最小，且不仅训练集这样，测试集这样。

这次实验交教给我一个很重要的启示：特征并非越多越好。上次实验时，我还一直在纳闷儿，为什么特征数减少了反而正确率上去了。我当时只是“粗略”地将其归因于“误差”，现在才真正弄明白了背后的道理。拿到一组数据，我们最好首先对它进行一个粗略的预处理，观察一下数据，在有直观感受之后再进行实验，可能效果会更好。

Fisher 的效果目前看来稍微优于 Bayesian，但是我在网上找了很多资料，也没有弄清楚到底背后的原因，除了我在之前探索过的，未来还应该继续更进一步探索

【参考资料】

《模式识别》张学工（第二版）

程序来源：

自己