研究生《模式识别实验》课程总结

研究6个问题，深度挖掘经典算法。

一. Fisher判别平面是最优的吗？

Fisher判别分析（又称线性判别分析，简称LDA）将两类样本投影到一条直线上，使同类样本点尽可能近，异类样本点尽可能远。那么，LDA找到的投影直线一定能获得最优判别效果吗？我们将用一个实例来说说明。

由图1-1可以看出，蓝点和橙点在投影直线上是可分的，只需要在一维的直线上训练一个分类器即可完成分类。

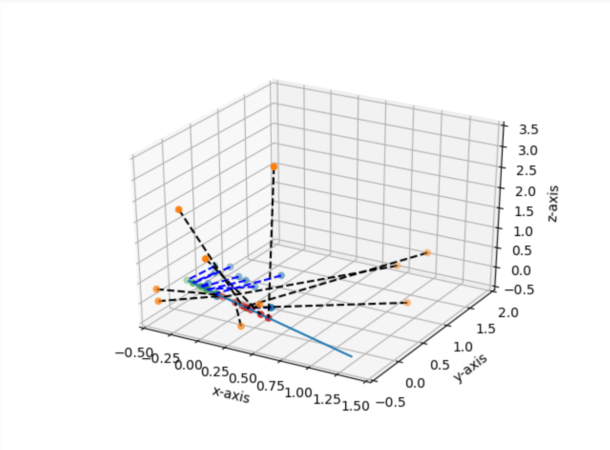

图1-1(1) 正常视角

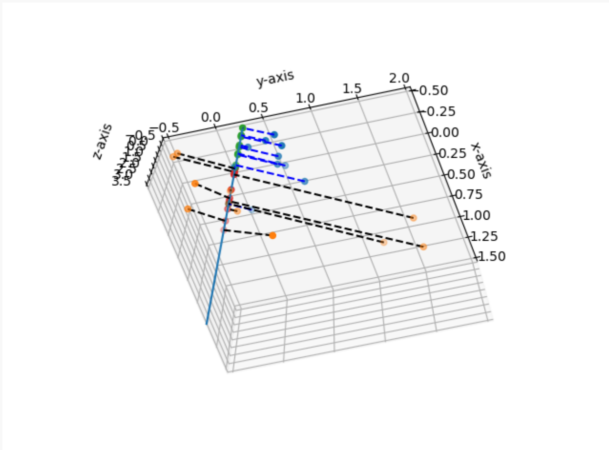

图1-1(2) 更直观的视角

下面来看一下LDA的投影直线是否是最优方向。

图1-2(1)是投影在LDA选取的直线上，用两个高斯分布分别拟合两类数据点，并求出决策面。

图1-2(2)是投影在另一条(巧妙选取的)直线上，也用两个高斯分布分别拟合两类数据点，并求出决策面

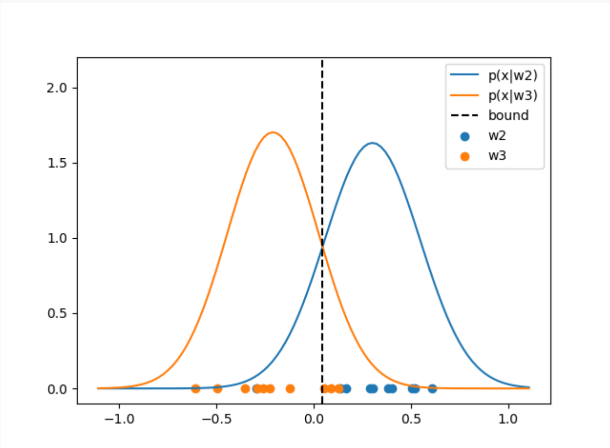

图1-2(1) LDA投影直线及拟合结果

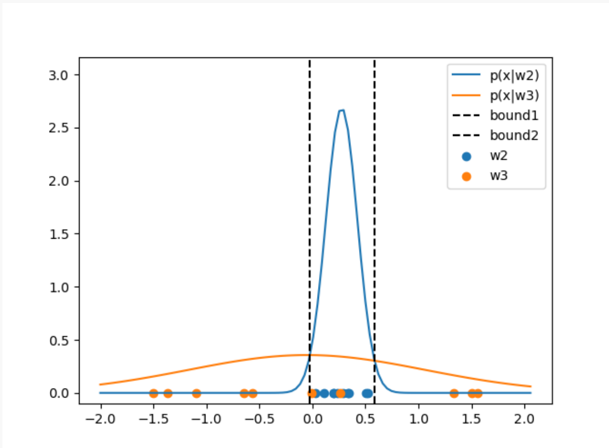

图1-2(2) 另一条投影直线及拟合结果

一个违反直觉的结论出现了：LDA的分类误差0.2，另一条直线0.1，LDA未能获得最优的判别效果。~~为什么会变成这样呢？第一次有子~~ 为什么会产生这一现象呢？

事实上，仔细观察图1-2可以发现，LDA降维后点的分离程度好于另一条投影直线，分类准确率下降是因为用高斯分布拟合了降维后的数据，这就引入了问题的新信息“降维后的数据服从高斯分布”，而LDA施加的偏好“最大判别度”与问题信息不符，故而在此情况下LDA的效果不是最优。

但毋容置疑的是，LDA在“满足两类样本最大可分性”这一意义上是最优的线性降维器，因为其目标函数就是如此。且在两类样本为同协方差的高斯分布时，最优降维器本身就是线性的，此时LDA是“满足两类样本最大可分性的降维器”。

二. 感知器及其变种

感知器有很多变种，其中最重要的一种是带裕量(margin)的感知器。虽然带margin的感知器可以通过增大权向量模长来获得任意大的margin，看似毫无意义，但只要对权向量加入模长约束，我们实际上就得到了SVM。

三. 偏差-方差分解

以回归问题为例，数据真实分布为$F(x)=x^2+\epsilon$，其中$\epsilon\sim N(0,0.1)$。用四种不同的函数来拟合数据：

$$\begin{aligned}
&g_1(x)=0.5\\
&g_2(x)=1\\
&g_3(x)=a_0+a_1x\\
&g_4(x)=a_0+a_1x+a_2x^2+a_3x^3
\end{aligned}$$

产生100个数据集，每个数据集中含有n=10个样本。每次在一个数据集上训练$g_i(x)$的所有参数，并计算$g_i(x)$

在所有数据集上的偏差和方差。当n=100时重复以上过程。计算得到的偏差与方差如表3-1所示，在训练样本变为原来的10倍后，方差缩小到原来的1/10。

表3-1(1) n=10

模型	平方偏差	方差
$g_1(x)$	0.120163	0
$g_2(x)$	0.580837	0
$g_3(x)$	0.040006	0.018861
$g_4(x)$	0.000281	0.037553

表3-1(2) n=100

模型	平方偏差	方差
$g_1(x)$	0.120930	0
$g_2(x)$	0.514928	0
$g_3(x)$	0.100161	0.002025
$g_4(x)$	0.000032	0.004245

图3-1是偏差-方差分解图，图像离y轴的距离表示平方偏差，图像在x轴上的投影长度表示方差，可以明显看出模型越复杂，偏差越低，方差越高；且无训练参数的常数函数方差为0。

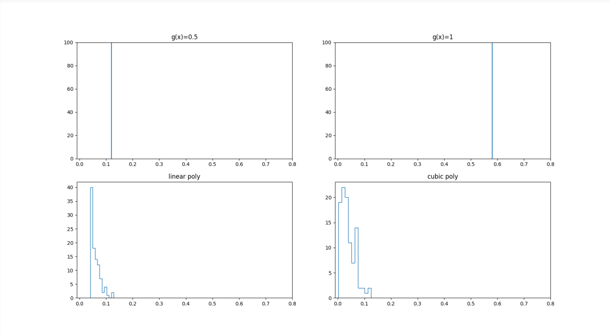

图3-1(1) n=10时

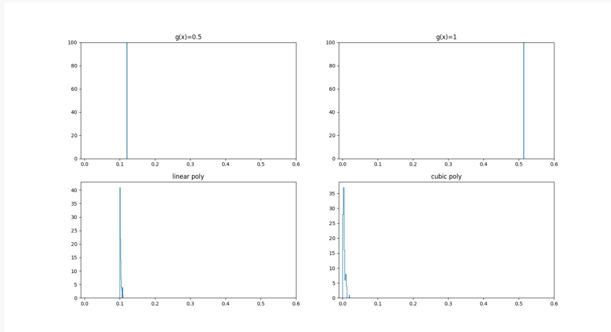

图3-1(2) n=100时

四. 没有免费的午餐

机器学习中有个著名的“没有免费的午餐(NFL)定理”：不可能存在这样的算法A和B，使得算法A在所有问题上都比算法B好。我们只能说在某个问题P上，算法A施加的偏好符合更加问题Q的先验信息，使得“在问题Q上，算法A好于算法B”。

不妨设想这样一个例子：从概率分布中采样120个点，其中60个点作为第一类C_1，另外60个点作为第二类C_2，按照9 : 1 : 2划分训练 : 验证 : 测试集。在验证集上为K近邻算法选择参数K，选择出一个在验证集上表现最好的K，一个在验证集上表现最差的K，并在测试集上测试K近邻分类结果，实验重复5次。

表4-1 K近邻分类器在测试集上的误差

	实验1	实验2	实验3	实验4	实验5
验证集最优K	0.55	0.65	0.6	0.45	0.5
验证集最差K	0.5	0.65	0.65	0.45	0.4

由表4-1可以看出，验证集最优K和最差K在训练集上的表现并无区别，误差都在0.5左右，等同于随机猜测。

从没有免费的午餐定理理解，KNN施加了“样本的标记等于该样本邻居中出现最多的标记”的偏好，但是在此问题中样本点的标记和邻居的标记之间没有任何联系。施加的偏好与问题不匹配， 所以无法获得好的效果。 在此情况下，交叉验证也无法挽救分类器的效果。

五. K-means和Fuzzy K-means的行为分析

表5-1 K-Means和FCM在不同初始化条件和不同距离度量下的迭代次数

		实验a	实验b	实验c	实验4
欧氏距离	K-Means	2	3	2	4
	FCM	10	12	21	27
$\beta = 0.001$	K-Means	2	3	2	4
	FCM	10	12	22	28
$\beta = 0.01$	K-Means	2	3	2	4
	FCM	14	17	28	33
$\beta = 0.1$	K-Means	2	3	2	4
	FCM	5	5	5	4
$\beta = 1$	K-Means	4	6	4	4
	FCM	2	2	2	2
$\beta = 10$	K-Means	2	2	2	2
	FCM	2	2	2	2
$\beta = 100$	K-Means	2	2	2	2
	FCM	2	2	2	2

在使用欧氏距离和β较小的距离时，K-Means 的迭代次数远小于 FCM，一种直观但不太严谨的理解方式是：K-Means 的收敛条件是样本的标记不再发生变化，标记是{0,1}的离散值；而 FCM 的收敛条件是样本的隶属度不再发生变化，隶属度是[0,1]上的连续值；[0,1]连续值比{0,1}离散值更难收敛。但是当使用β较大的距离时，两个算法的迭代次数都很小，这是因为在此距离度量下，样本到聚类中心的距离几乎都为 1，算法在一开始就完成了聚类分配，导致无法继续更新。这说明说明聚类算法对距离度量敏感，一个坏的距离度量（例如β很大时的距离）会使得聚类算法无法正常收敛到较好的值。

图5-1给出了一个聚类结果示意图。

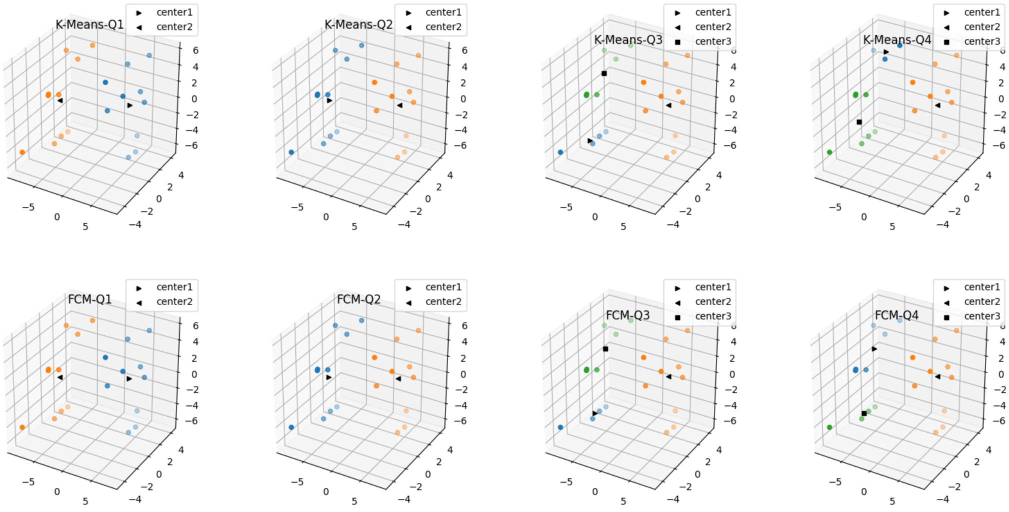

图5-1 在欧式距离下K-Means与FCM聚类结果示意图

六. PCA、MDA降维对比

实验设置：PCA降到80维，MDA降到20维，DPDR降到200维。

表6-1 降维的运行时间和降维后使用KNN分类的测试误差

	正确率(%)	运行时间(s)
PCA(一般方法)	89.5	514.636
PCA(使用技巧)	89.5	0.064
MDA	88.5	627.107
DPDR	90	0.126

注：DPDR来自Hyunsoo Kim, Haesun Park, Hongyuan Zha, Distance Preserving Dimension Reduction Using the QR Factorization or the Cholesky Factorization, available by google (scholar).

结果分析：

使用技巧的PCA的速度大约是一般PCA的10000倍，这是因为只需要对一个200×200的矩阵进行特征值分解，而不需要分解$(112 \times 92) \times (112 \times 92)$的协方差矩阵。使用技巧的PCA得到了与一般PCA相同的分类正确率，这说明当样本维度大于样本数量时，选择技巧型PCA总是可以受益。PCA的目标是求中心化样本矩阵$X \in R^{N \times D}$的右奇异向量，在$D > N$时，使用技巧的PCA先求较易求得X的左奇异向量，再利用奇异向量之间的关系得到右奇异向量，即投影矩阵。

表6-2 PCA和DPDR的重建误差

	仅训练集	训练集和测试集
PCA	133.09	81.24
DPDR	53.14	1.34

结果分析：

在训练集和测试集上得到的DPDR可以精确重建测试集，说明DPDR“记住”了所见过的所有样本。在训练集上得到的DPDR对测试集的重建误差小于PCA，说明DPDR在此问题上的外推能力强于PCA。