Homework #2

劉淑雯, Rachel Lau

0540083, CS Dept., NCTU

Nov. 2016

1 Dimension Reduction

1.1 Question (a)

For a K-class problem, the between-class scatter matrix is defined by:

$$S_B = \sum_{k=1}^{K} N_k\,(m_k - m)(m_k - m)^T$$

where $m_k = \frac{1}{N_k}\sum^{n\in C_k} x_n$ and $m = \frac{1}{N}\sum_{i=1}^{K} N_i\,m_i$. Show that the maximum rank of S_B is $K-1$.

Proof:

$$S_B = \sum_{k=1}^{K} N_k\,(m_k - m)(m_k - m)^T = [N_1(m_1 - m) \quad \dots \quad N_k(m_k - m)]\begin{bmatrix}(m_1 - m)^T \\ \vdots \\ (m_k - m)^T\end{bmatrix}$$

∵ Gaussian Elimination on $[N_1(m_1 - m) \quad \dots \quad N_k(m_k - m)]$:

$$\Rightarrow \left[\sum\nolimits_{k=1}^{K} N_k(m_k - m) \quad N_2(m_2 - m) \quad \dots \quad N_k(m_k - m)\right]$$
$$\Rightarrow [0 \quad N_2(m_2 - m) \quad \dots \quad N_k(m_k - m)]$$

∴ det(A)=0

∴ $\det(S_B) = \det(\mathrm{A})\det(\mathrm{B}) = 0$

∴ $\mathrm{rank}(S_B) \le \mathrm{k} - 1$

1.2 Question (b)

1.2.1 Question b_1

The classification accuracy on training set is 97.50%.

Table 1 Classification chart of Training set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET39	39	0	0	**39**
VIR	0	40	1	**41**
VER	0	2	38	**40**
Total	**39**	**42**	**39**	**120**

The classification accuracy on test set is 100.00%.

Table 2 Classification chart of Test set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET	11	0	0	**11**
VIR	0	9	0	**9**
VER	0	0	10	**10**
Total	**11**	**9**	**10**	**30**

1.2.2 Question b_2

(1) PCA – 3 dimension

The classification accuracy on training set is 98.33%.

Table 3 Classification Chart of Training set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET39	39	0	0	**39**
VIR	0	41	0	**41**
VER	0	2	38	**40**
Total	**39**	**43**	**38**	**120**

The classification accuracy on test set is 100.00%.

Table 4 Classification Chart of Test set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET	11	0	0	**11**
VIR	0	9	0	**9**
VER	0	0	10	**10**
Total	**11**	**9**	**10**	**30**

(2) PCA - 2 dimension

The classification accuracy on training set is 96.67%.

Table 5 Classification Chart of Training set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET39	39	0	0	**39**
VIR	0	39	2	**41**
VER	0	2	38	**40**
Total	**39**	**41**	**40**	**120**

The classification accuracy on test set is 100.00%.

Table 6 Classification Chart of Test set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET	11	0	0	**11**
VIR	0	9	0	**9**
VER	0	0	10	**10**
Total	**11**	**9**	**10**	**30**

(3) PCA - 1 dimension

The classification accuracy on training set is 84.17%.

Table 7 Classification Chart of Training set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET39	39	0	0	**39**
VIR	0	30	11	**41**
VER	3	5	32	**40**
Total	**42**	**35**	**43**	**120**

The classification accuracy on test set is 76.67%.

Table 8 Classification Chart of Test set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET	11	0	0	**11**
VIR	0	7	2	**9**

VER	1	4	5	**10**
Total	**12**	**11**	**7**	**30**

1.2.3 Question b_3

(1) LDA- 3 dimension

The classification accuracy on training set is 97.50%.

Table 9 Classification Chart of Training set

Truth\Predict	SET	VIR	VER	Total
SET39	39	0	0	**39**
VIR	0	40	1	**41**
VER	0	2	38	**40**
Total	**39**	**42**	**39**	**120**

The classification accuracy on test set is 100.00%.

Table 10 Classification Chart of Test set

Truth\Predict	SET	VIR	VER	Total
SET	11	0	0	**11**
VIR	0	9	0	**9**
VER	0	0	10	**10**
Total	**11**	**9**	**10**	**30**

(2) LDA- 2 dimension

The classification accuracy on training set is 97.50%.

Table 11 Classification Chart of Training set

Truth\Predict	SET	VIR	VER	Total
SET39	39	0	0	**39**
VIR	0	40	1	**41**
VER	0	2	38	**40**
Total	**39**	**42**	**39**	**120**

The classification accuracy on test set is 100.00%.

Table 12 Classification Chart of Test set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET	11	0	0	**11**
VIR	0	9	0	**9**
VER	0	0	10	**10**
Total	**11**	**9**	**10**	**30**

(3) LDA - 1 dimension

The classification accuracy on training set is 98.33%.

Table 13 Classification Chart of Training set

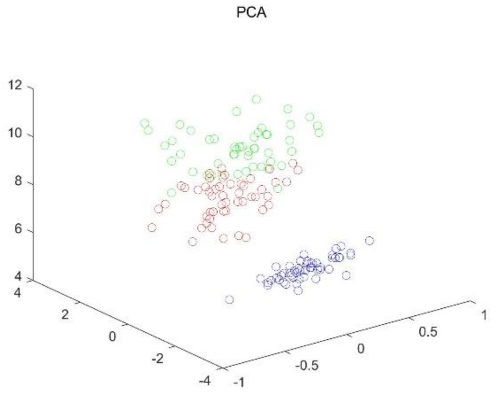
Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET39	39	0	0	**39**
VIR	0	41	0	**41**
VER	0	2	38	**40**
Total	**39**	**43**	**38**	**120**

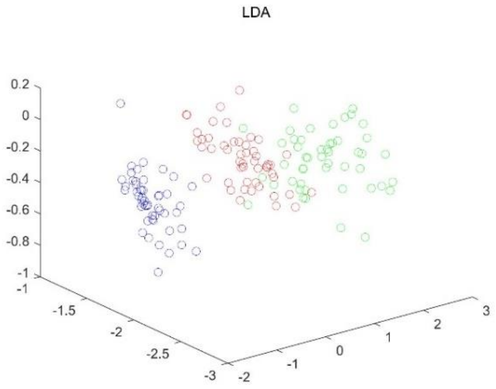
The classification accuracy on test set is 100.00%.

Table 14 Classification Chart of Test set

Truth\Predict	**SET**	**VIR**	**VER**	**Total**
SET	11	0	0	**11**
VIR	0	9	0	**9**
VER	0	0	10	**10**
Total	**11**	**9**	**10**	**30**

1.2.4 Question b_4



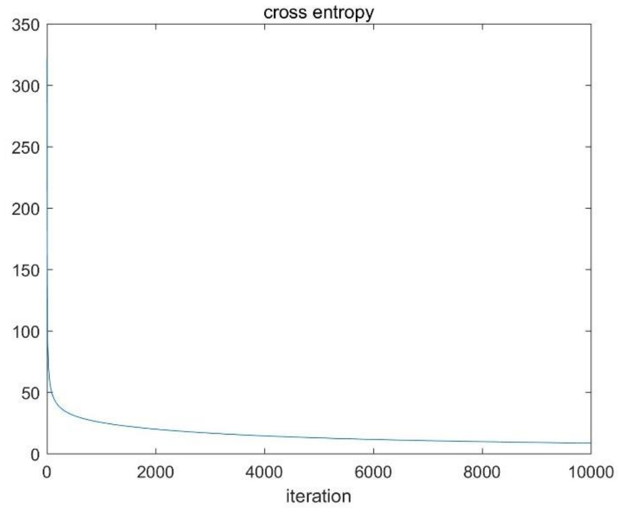


Discussion:

PCA 追求的是在降维之后能够最大化保持数据的内在信息，并通过衡量在投影方向上的数据方差的大小来衡量该方向的重要性。但是这样投影以后对数据的区分作用并不大，反而可能使得数据点揉杂在一起无法区分。这也是 PCA 存在的最大一个问题，这导致使用 PCA 在很多情况下的分类效果并不好。与 PCA 保持数据信息不同，LDA 是为了使得降维后的数据点尽可能地容易被区分，即同类的数据点尽可能的接近（within class）、不同类的数据点尽可能的分开（between class）。

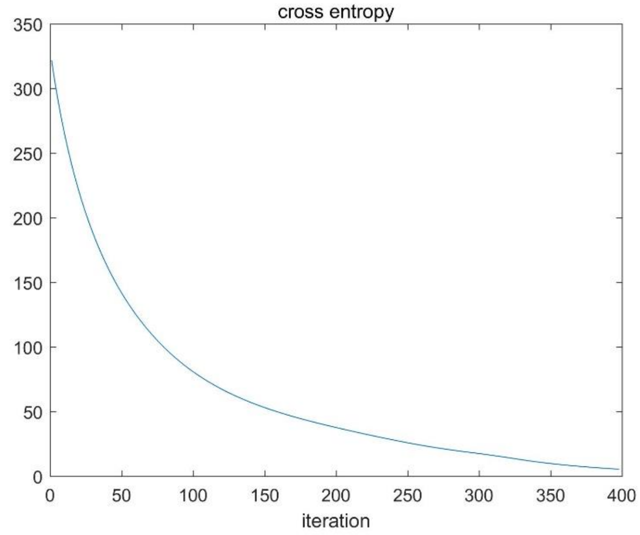
2 Logistic Regression

2.1 Question (a)



The Test Miss Classification Rate is 0.0800.

2.2 Question (b)



The Test Miss Classification Rate is 0.0800.