**Pat Kujawa, CSCI 548-01 Pattern Rec, Spring 2013**

LDA Project

Motivation

This project will acquaint you with the LDA analysis tools available in R. Additionally, you will begin the process of becoming familiar with how to visualize data once transformed by the LDA process. You will also compare some of the plots to PCA and MDS results.

Big Picture

You will read in several datasets, perform any necessary manipulation on the data to strip out unnecessary data and or labels, perform MDS upon the data, and plot the results.

Project Instructions

Iris

1. Create a working directory for your project. Create a text file to be used as an R script in this directory and start a submission document in the directory.

2. Perform LDA on the iris data (in the script). Let’s create a random sample from the set of 150 instances of size 125 on which to train. If you don’t remember how to do this take a look at the slides.

3. Test the accuracy of your classifier on remaining (test) data. Remember the trick with the minus sign to generate the test set (again, see the slides).

When you acquire the predictions you can compare them to the actual with something like “which(predictions != actual)”. This will give you a vector of missed predictions. You can then determine the length of this vector to get a count of those missed. Have the script print the accuracy to the screen. Paste or enter this value along with the appropriate heading into the submission document (mine achieved 98.7% accuracy; yours may vary due to random sampling).

[1] "iris"

[1] "Accuracy was 100%" # This varies a bit on different runs

4. Next you are to create a scatter plot of the data in reduced dimensional space (two dimensions represented by the first two Linear Discriminants). **Remember, the lda function does not project the data for you. You must do this yourself**(using the %\*% operator; see the slides if you don’t remember). Actually, you will create three scatter plots: one for LDA, one for MDS, and one for PCA. They will be placed in a three panel figure and will be colorized by species. Make the plot a 6.6X2.2 window and tighten up the margins like we did in the previous project. The points should be filled-in (solid) points and the plot should have an appropriate title and X and Y labels. Also, don’t forget to include a legend.  To get a three panel plot you will use the layout command.

windows(6.6,2.2)

layout(matrix(1:3, 1, 3), widths=c(1, 1, 1)  );

layout.show(3)

The layout.show command shows the three panel plot (see below). If you were to perform the plot command three times, each would go in an associated panel. Try running matrix(1:3, 1, 3) at the command line. This should give you an idea of how it is used in the layout command. Put these commands in the script.

5. There will be no title for these plots. Instead there will be an (A), (B), and (C) and they will be inside the bounds of the plots.  Since there will be no titles, make sure you make the upper margin zero. We will also not be putting a legend in each (not enough room). Mine looks like the following. I did this with the mtext command. I placed the following command right after the first plot:

mtext("(A)",line = -2)

6. Place code in the script to display the LDA, MDS, and PCA plots. Feel free to copy portions of your scripts from previous projects. In journal and conference papers, figures should stand alone. That is, someone should be able to look at a figure and its caption (figures tend to have captions below while tables above) and tell enough about it to interpret the underlying purpose and meaning of the figure. Copy and paste the three panel figure into your document and create a caption that has a short, descriptive title (like “Iris Data” or “Iris Scatter Plots”) followed by descriptions for the three panels and the colors. **Note: grad students will be expected to use this kind of format in their reports**.

Machine generated alternative text: .
.. (A) — . (B) .
(C)
.
0
.
. cN_ .: •. ELfl . .
.. . C
. . ®
.
¿‘e.
. .4• 3
E 0•
. 0
0c
..
„t t n.
.
t •
_______ .— C
.
. t
:. a0 • a ‘t
SetS! _ . .
arsotr ‘ :
, . .
I I I I I I I I I I I
-10 -5 0 5 -2 0 2 4 -3-2-101234
Linear Discriminant I Principal Coordinate I Principal Component 1

Figure 1: Iris data projected onto first two linear discriminants (A), principal coordinates (B), and principal components (C).

7. Which if any shows the best separation? Place this question and its answer in the submission document. Don’t forget to put the appropriate header, title, and section heading information in the document for the plots and the answers.

MDS and PCA produce essentially the same results, just mirrored; LDA, on the other hand, produces a different plot of points. I think LDA presents the best separation by class because there is less overlap between versicolor and virginica, while setosa is still off in its own cluster.

Fruit, Mouse, Tumor

1. Perform the same operations for the fruit, mouse, and tumor data (you should have a section for each with a three panel figure in each complete with heading and caption).

[1] "fruit"

[1] "Accuracy was 90%"

Machine generated alternative text: .
0
CN
C
C
o
o
co.- -
Q-.
o
C
D
.
.
.
C
CN
.-
E
o
(-)
o
C
C
Q_o.
• •?•.•
o
Linear Discriminant I
.
. .
.
.
.
.
I I I I I
-100 -50 0 50 100
Principal Component 1
.
.
I I I I
-100 -50 0 50 100
Principal Coordinate I

Figure 2: Fruit data projected onto first two linear discriminants (A), principal coordinates (B), and principal components (C). It is clear that LDA (A) provides the best separation of classes.

2. Note: for the mouse data classify by proximal and distal (not by mouse type). Colorize that way as well. There are only 20 data points and close to a hundred dimensions. For this reason, let’s use the entire data set as both the training data and testing data.

3. The data is quite sparse (lots of zeros) and 38 of the dimensions appear to be constant within groups (this is an error you get when you try to run LDA). The help screen indicates that the “function tries hard to detect if the within-class covariance matrix is singular.” It checks whether any variable has within-group variance less than tol^2 and if so it will stop and report the variable as constant. Just for the heck of it, let’s force it to run anyway (set tol=0 in the lda command). You’ll get a warning but it should run (it did for me, though removing certain data points caused singular matrices and failure to compute).

4. You should get a fairly strange looking plot (mine was two vertical columns of points with the distribution along LD2 very small. This led to 100% accuracy for my classifier (not trustworthy because the test data was also the training data).

[1] "Mice"

[1] "Used all data to train, so accuracy is irrelevant."

Machine generated alternative text: .
(A) O (B) g (C)
c-iC c-Ic-i
C
C
CC • C •
. 0..
E
O 0c
(-)
L-
5 - ci
C
t - t
a .
Cfl
IazaI o
cv)
à
— I I I I I I I I I
9.Oe+14 1.le+15 -0.3 -0.2 -0.1 0.0 0.1 0.2 -0.3 -0.2 -0.1 0.0 0.1 0.2
Linear Discriminant I Principal Coordinate I Principal Component 1

Figure 3: Mouse data projected onto first two linear discriminants (A), principal coordinates (B), and principal components (C). Again, it is clear that LDA (A) provides the best separation of classes.

5. When performing LDA on the tumor data, I could not get the function to provide two linear discriminants (possibly it causes the accuracy to go down). So I plotted the distribution values on the first linear discriminant. This is what I want you to do as well. Also, LDA did not seem to like the “?’s” in the data.

[1] "Tumor"

[1] "Accuracy was 100%"

Machine generated alternative text: (A) (C)
.
Cl
C
t. .- •
4 _ ____
• .
0- .
•• •. o
• a
. , — . .
r t
. t n
.
. .• _________
j
• eelkgn
MaI9fl o .
I I I • I I I
2 4 6 8 -15 -10 -5 0 5
Linear Discriminant I Principal Coordinate I Principal Component 1
.

Figure 4: Wisconsin breast cancer (tumor) data projected onto first two linear discriminants (A), principal coordinates (B), and principal components (C). It seems that LDA (A) has just as many benign and malignant interspersed values as MDS (B) and PCA (C), but both MDS and PCA have a tighter cluster for benign values, so either would be my choice for a separability measure.

(NOTE: I had no problem getting two linear discriminants for the tumor data, so I did a scatter plot.)

10. Don’t forget to include an answer to the question about which provides the best separation. Also don’t forget to include the accuracy for each dataset. It may be difficult for the tumor data since they it isn’t a scatter plot.

11. Finish-up the write-up with your opinion as to why the best separation is found in the plots you identified. Also include some speculation as to whether the mouse data LDA plot indicates there is hope for a classifier to be able to identify proximal vs. distal, even across different mice.

My opinions on separation are included in the titles for each figure. As for the mouse data, the LDA plot would strongly suggest that a classifier could discriminate between proximal and distal. However, it is important to keep in mind that the mouse data has only twenty experiments and over ninety dimensions, so any classifier would need significantly more data to preform accurately on new experiments.