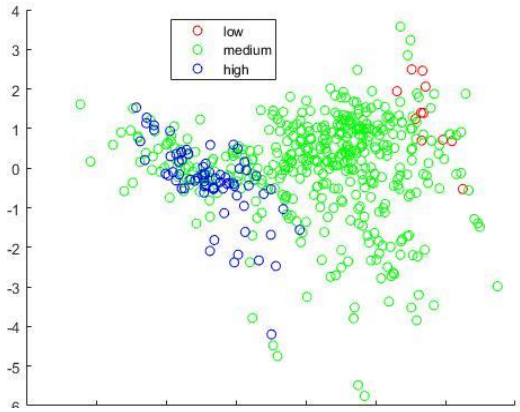
Intelligent Data Analysis

Martin Russell
School of Computer Science
Thursday, 20 February 2020

Solution sheet – week 4 – Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA)

1. Complete the application of PCA to the investigation of the 'nox' parameter in the Boston data. Check that the projection onto the first two principal components is the same as on the slides

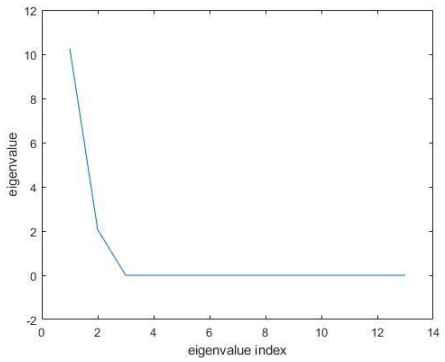
   Solution: Check your solution against the figure in the "PCA Example" slides. Your figure should look like this:

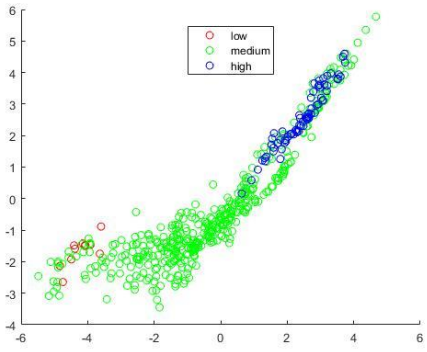


2. Apply LDA to investigate the 'nox' parameter in the Boston data. Project the data onto the first two LDA eigenvectors.

   Solution: I used the MATLAB script "bostonLDA.m" which I will put on the Canvas page in the "MATLAB scripts" box.

   The plot of singular values and scatter plot of the data projected onto the 2-dimensional space spanned by the two LDA principal vectors should look like this:





3. Let $X = \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ be the set of two-dimensional vectors defined by:

$$x_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x_2 = \begin{bmatrix} 5 \\ -1 \end{bmatrix}, x_3 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, x_4 = \begin{bmatrix} 4 \\ -3 \end{bmatrix},$$
$$x_5 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}, x_6 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, x_7 = \begin{bmatrix} 7 \\ -4 \end{bmatrix}, x_8 = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$$

(a) Calculate the covariance matrix of X. (I suggest you use a calculator!)

Solution: First calculate the average m of $\{x_1, \dots, x_8\}$: $m = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$.

Next let X be the 8 x 2 matrix whose i^{th} row is $x_i - m$. Then the covariance matrix is

$$C = \frac{1}{7} X^T X = \begin{bmatrix} 13.143 & -7.429 \\ -7.429 & 4.857 \end{bmatrix}.$$

(b) Explain the sequence of steps involved in applying PCA (Principal Components Analysis) to a set of data, and how the result should be interpreted.

Solution:

1. First calculate the covariance matrix C: Suppose there are N data points $\{x_1, \dots, x_N\}$ each of dimension d. Let m be the average of $x_1, \dots, x_N$ and X be the $N \times d$ matrix whose n^{th} row is $x_n - m$. The covariance matrix C can then be calculated as

$$C = \frac{1}{N-1} X^T X$$

2. Calculate the eigenvalue decomposition of C. In other words find a $d \times d$ orthogonal matrix U and $d \times d$ diagonal matrix D such that

$$C = UDU^T$$

3. The eigenvalues are the columns of U. The eigenvector u_i corresponding to the biggest eigenvalue d_{ii}, is the direction of maximum variance of the data set and the variance of the data in that direction is d_{ii}.

4. If u_j is the eigenvector corresponding to the second biggest eigenvalue d_{jj}, then u_j is the direction of maximum variance of the data in the subspace obtained by discarding u_i (i.e. the $N-1$ dimensional subspace spanned by the eigenvalues except u_i.

5. For dimension reduction to $M < N$, project the data onto the M dimensional subspace spanned by the M eigenvectors corresponding to the M biggest eigenvectors. For any data point x_n this is the M dimensional subspace for which the mean-squared-error between the projection of x_n onto that space and the original point x_n is minimized on average.

For visualization the data points should be projected onto the subspace spanned by the two eigenvalues corresponding to the two biggest eigenvalues.

(c) Apply PCA to the data set X. Write down the two Principal Components and the variance of X in the directions of each of the Principal Components. *(If you know how to calculate eigenvalues and eigenvectors then do this by hand. If you don't know*

how to calculate eigenvalues and eigenvectors but you've done it in the past then revise it and do it by hand. If you don't know how to calculate eigenvalues and eigenvectors and you've not done it in the past then use MATLAB (or similar).

Solution:

The eigenvalue decomposition of the covariance matric C is

$C = \begin{bmatrix} 13.143 & -7.429 \\ -7.429 & 4.857 \end{bmatrix} = UDU^T$ where $U = \begin{bmatrix} -0.506 & -0.862 \\ -0.862 & 0.506 \end{bmatrix}, D = \begin{bmatrix} 0.494 & 0 \\ 0 & 17.506 \end{bmatrix}$.

Hence the two principal components and variances in their directions are:

$e_1 = \begin{bmatrix} -0.862 \\ 0.506 \end{bmatrix}$ with variance $v_1 = 17.506$ and $e_2 = \begin{bmatrix} -0.506 \\ -0.862 \end{bmatrix}$ with variance $v_1 = 0.494$.

4. A 6 dimensional data set has sample covariance matrix C, which has eigenvalue decomposition $C = UDU^T$

(i) What are the properties of the matrices D and U?

Solution:
D and U are both 6×6 marices. D is a diagonal matrix whose diagonal entries are all real and greater than or equal to zero. The diagonal elements of D are the eigenvalues of C. U is an orthogonal matrix. Its values are all real numbers and it satisfies

$$UU^T = I = U^T U \qquad (2.1)$$

where I is the 6×6 identity matrix. If u_i is the i^{th} column of U, then from equation (2.1) each u_i is a unit vector ($u_i \cdot u_i = \|u_i\|^2 = 1$) and $u_i \cdot u_j = 0$ if $i \neq j$. The unit vector u_i is an eigenvector of C with eigenvalue d_{ii}, the i^{th} diagonal element of D.

(ii) Given a 6-dimensional vector v, explain how you would calculate the projection of v onto the two most significant principal components of the data set.

Solution:
Suppose that d_{ii} and d_{jj} are two the biggest elements of D. Then the most significant two principal components are u_i and u_j. The projection of v onto the subspace spanned by u_i and u_j is $\begin{bmatrix} v \cdot u_i \\ v \cdot u_j \end{bmatrix}$.