Ayon Biswas

15110037

Probability assignment 1

Principal component analysis

Introduction

Principal component analysis is a statistical tool used to reduce the dimensionality of a data set. The idea is to convert a set of correlated variables into a set of uncorrelated variables. The transformation takes place in such a way such that the variance along the principal components decreases with the first component showing the maximum variance.

Approach Followed

1. Construct the data matrix X_{mxn}, where m is the number of features and n is the number of data instances.

$$X = \begin{pmatrix} x11 & x12 & \dots & \dots & x1n \\ x21 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ xm1 & xm2 & \cdots & \cdots & xmn \end{pmatrix}$$

   So for the iris dataset m = 4, n = 150

2. Find the mean column vector μ by averaging all the column vectors and subtract it from each of the column vectors to from the mean shifted data matrix D

$$\mu = \sum_{i=1}^{n} \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ \vdots \\ x_{3i} \end{pmatrix}$$

$$D_i = X_i - \mu_i$$

3. If the number of dimensions is less than the number of data points,
   - compute the covariance matrix, $cov = (\frac{1}{n})DD'$. Then find the eigenvectors, eigenvalues of the covariance matrix by invoking the eig function in MATLAB.

$$cov = \begin{pmatrix} var(d1) & Cov(d1,d2) & \dots & \dots & Cov(dm,d1) \\ Cov(d1,d2) & var(d2) & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ Cov(dm,d1) & Cov(dm,2) & \dots & \cdots & Var(dm) \end{pmatrix}$$

   Eigenvectors of matrix A are computed using the following equation:

$$VA = \lambda V$$

where V and λ represent the eigenvector and eigenvalue of matrix A

4. If the number of dimensions is more than the number of data points as in case of images then computing the covariance matrix would be computationally expensive.so the following trick is used:
    - Compute the eigenvector, eigenvalues of $\left(\frac{1}{n}\right) D'D$. The required eigenvector matrix is obtained by $E' = DE$ and E' is then normalised
5. But since in case of iris there are only 4 features therefore covariance matrix is computed directly as in 3.
6. Eigenvalues are sorted in descending order and the corresponding eigenvectors are arranged in the same order. It is done because the eigenvectors with large eigenvalues give high variation of data along its direction. So we need to omit those eigenvectors along which variation is insignificant to consider.
7. So we take first k eigenvectors which we call the principal components and take the projection of the mean shifted data matrix D along those components.

$$Y = P^T X = \sum_{i=1}^{n} P^T (X_i - \mu)$$

Where Y is the projected matrix and P is the principal component matrix

8. Then we plot the projected data when the number of principal components is 1,2 and 3. We observe maximum variance along the first principal component. This variation decreases along subsequent principal components.
9. We then reconstruct data matrix, D_r using the projected data matrix P by using the following relation:

$$\tilde{X} = PY + \mu$$

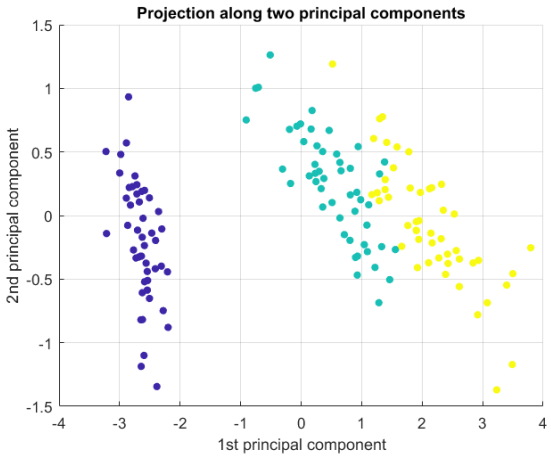
10. Thereafter we calculate the error E by taking the 2-norm of the data matrices, D and D_r and we also calculate robustness as defined by

$$error = \sum_{i=1}^{n} \left\| \tilde{X}_i - X_i \right\|^2$$

Robustness is a measure of the information that is retained by the given number of principal components.it is defined by

$$Robustnes = \frac{Total\ variance\ of\ P}{total\ variance} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

Observations and Inference



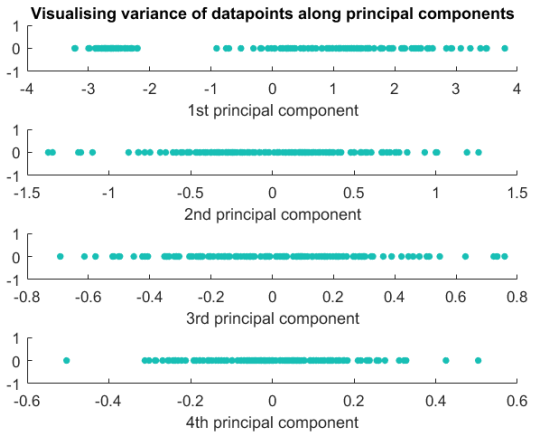
Labels:

purple – setosa

blue – versicolor

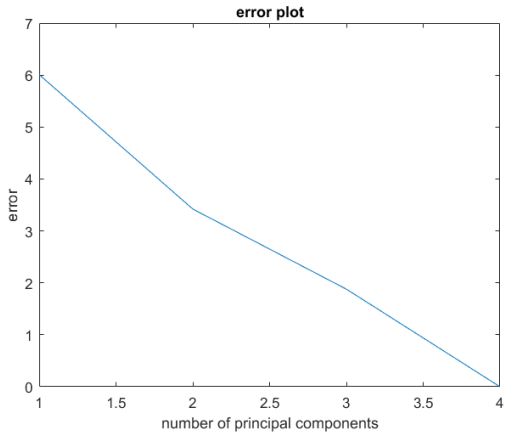
green - virginica

We observe maximum variation along first principal components,



We observe variation of data when projected along one principal component at a time

The variation of data points decreases as we go from first principal to subsequent components



The reconstruction error increases for as we take less principal components.

Error when we take three components = 1.875, for two components = 3.425, for one component = 6

Robustness obtained for three principal components = 0.9948, for two principal components = 0.9776,

For one principal component = 0.9246

We can see that maximum information is contained along the first principal component as it alone accounts for 0.9246 robustness.

References

- www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf
- A Tutorial on Principal Component Analysis *Jonathon Shlens*
- https://archive.ics.uci.edu/ml/datasets/iris