Homework 1 Machine Learning

Marco Treglia

1 Principal Component Analisys

Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. PCA has many applications;

First, compression representing $x(i)$ with lower dimension $y(i)$ were, $x(i)$ is the data set given $x(i); i = 1, ..., m.$ and $y(i)$ is the data set after the PCA.

Another standard application is to preprocess a dataset to reduce its dimension before running a supervised learning algorithm with the $x(i)$s as inputs. Apart from computational benefits, reducing the datas dimension can also reduce the complexity of the hypothesis class considered and help avoid overfitting.

Lastly, we can also view PCA as a noise reduction algorithm.

2 Mathematical step

2.1 Pre-processing data

Prior to running PCA per se, typically we first pre-process the data to normalize its mean and variance, as follows:

1. Let $\mu = \frac{1}{m} \sum_{i=1}^{m} x^{(i)}$.
2. Replace each $x(i)$ with $x(i) - \mu$.
3. Let $\sigma_j^2 = \frac{1}{m} \sum_i (x_j^{(i)})^2$
4. Replace each $x_j^{(i)}$ with $\frac{x_j^{(i)}}{\sigma_j^2}$.

2.2 Covariance Matrix

First compute the covariance matrix of the data, which is a d x d matrix where each element represents the covariance between two features. The covariance between two features is calculated as follows: :

$$\sigma_{jk} \ = \ \frac{1}{n \ - \ 1} \ \sum_{i=1}^{N} \left(\left(x_{ij} \ - \ \overline{x}_j \right) \left(x_{ik} \ - \ \overline{x}_k \right) \right)$$

where the $\overline{x}$ is the mean vector :

$$\overline{x} = \sum_{k=1}^{n} x_i$$

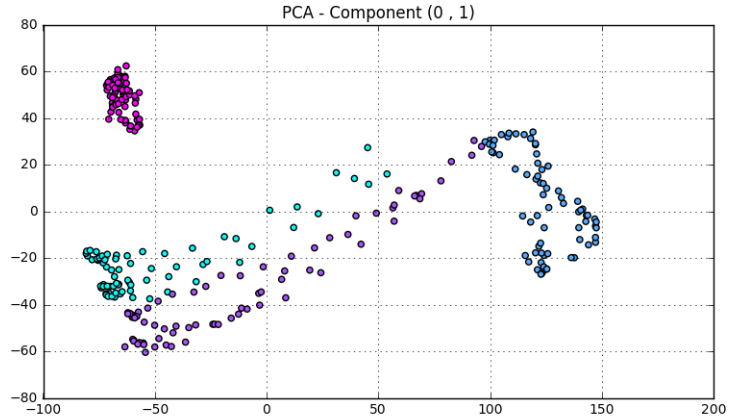
2.3 Eigenvector and eigenvalue

From the covariance matrix compute the eigenvectors and eigenvalues. This represent the core of a PCA: The eigenvectors (principal components) determine the directions of the new feature space, and the eigenvalues determine their magnitude. In other words, the eigenvalues explain the variance of the data along the new feature axes.

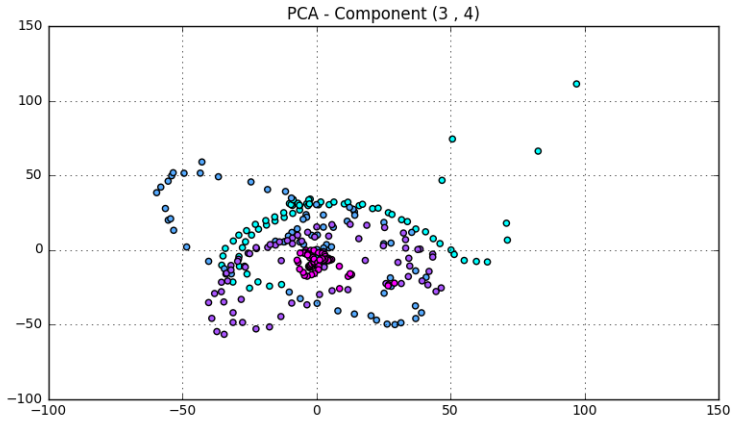
3 Principal Component Visualization

For this section i chosed the first four image, loaded with PIL.image, prepocessed for standardize the data then compute the principale component analisys.

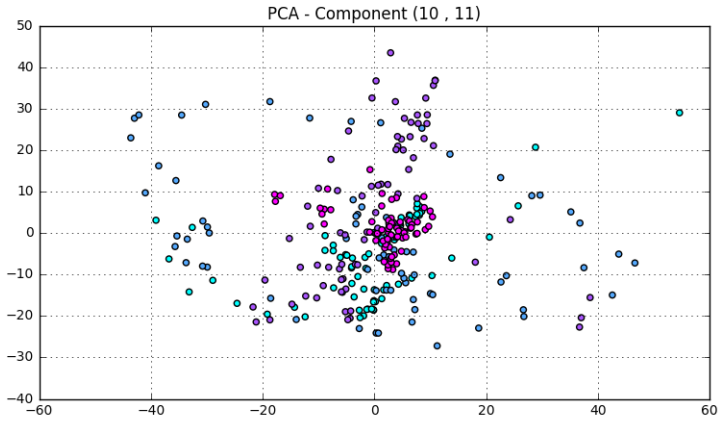
3.1 First and Second Componet



3.2 Third and Four Componet



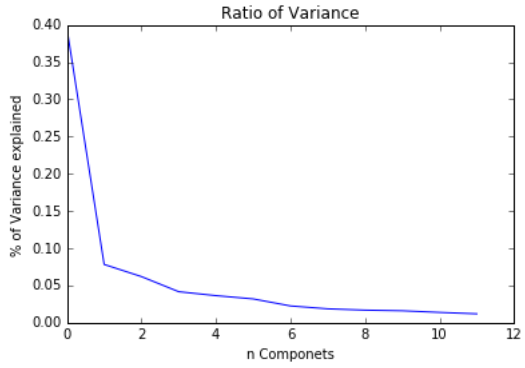
3.3 Thenth and Eleventh Componet



3.4 Principale component needed

For determinade the principale component needed for getting the best evaluation of the data , it can be used the explained variance. This tells us how much information (variance) can be attributed to each of the principal components.

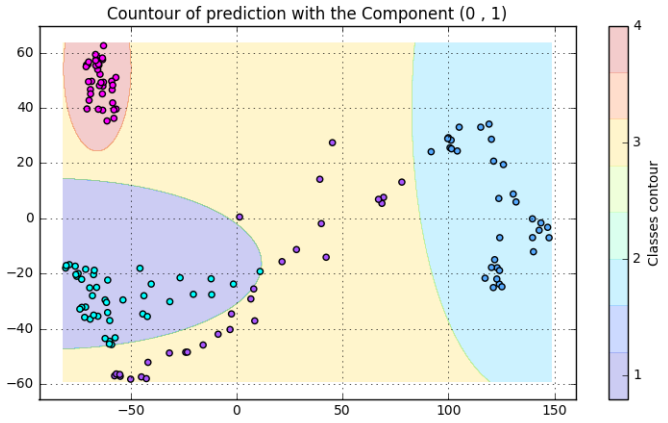
Component	Unitary Variance	Cumulative Variance
1	64,77%	64,77 %
2	12,81%	77,58 %
3	10,15%	87,73 %
4	6,81%	94,54 %
5	5,46%	100 %



4 Classification with Naive Bayes

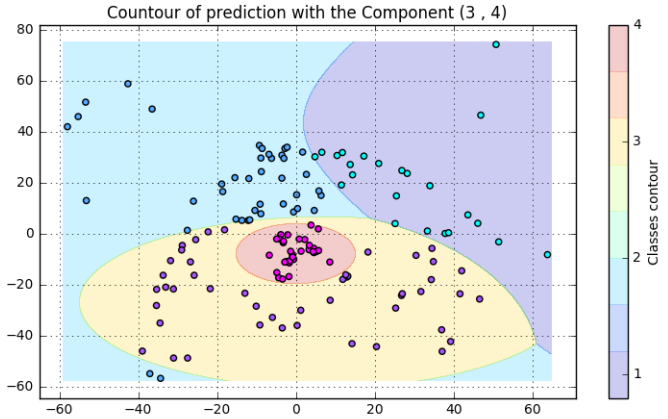
The Bayesian Classification represents a supervised learning method as well as a statistical method for classification.provides practical learning algorithms and prior knowledge and observed data can be combined.

4.1 First and Second Component prediction



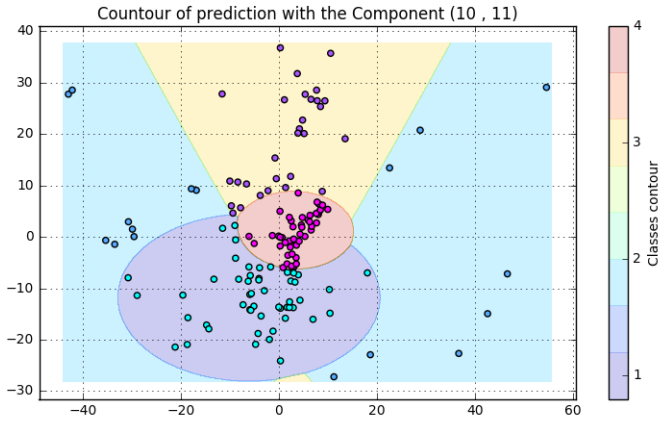
Accuracy of 90.277 %

4.2 Third and Four Component prediction



Accuracy of 59.722 %

4.3 Thenth and Eleventh Component prediction



Accuracy of 50.694 %