Mandatory assignment 3 - IN3050

Alice Monceyron Jonassen, alicemj

April 2022

PCA

Initial questions

What is the variance?
The variance is defined mathematically as

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2,$$

and it describes the spread of the data.

What is the covariance?
The covariance indicate how the variance of two stochastic variables depend on each other.

How do we compute the covariance matrix?
We can compute the covariance matrix by

$$\frac{1}{N}(XX^T),$$

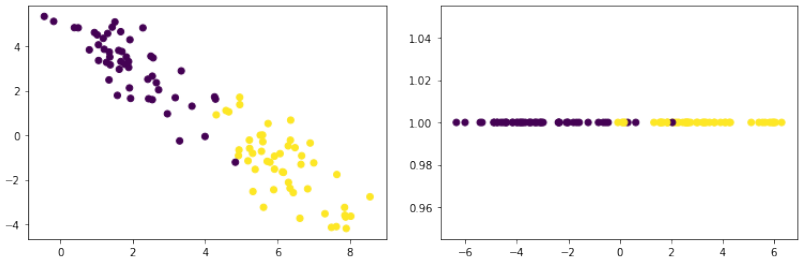
where N is the number of rows in the data set X.

What is the meaning of the principle of maximum variance?
The principle of maximum variance is finding the direction of most spread on the projection.

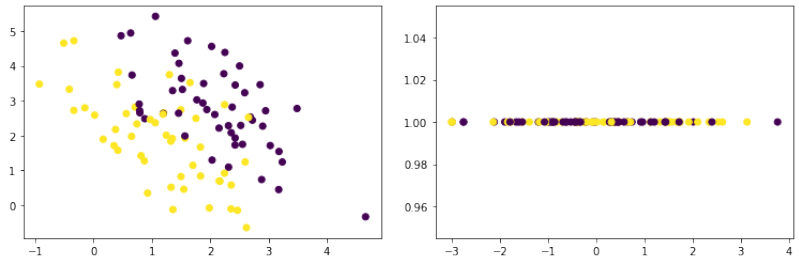
Why do we need this principle?
We need this principle to find a good lower dimensional representation of the data.

Running PCA - 2D-data

We run the PCA for two different data sets. For the first data set we plot the data and then plot the PCA transformation, to get the following two plots:



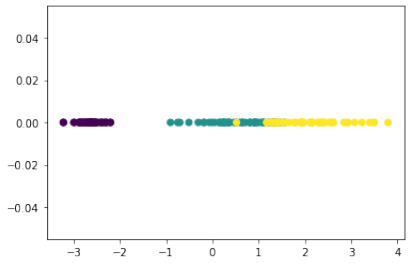
We can clearly distinguish the two group that are present in the the original data set.
For the second data set the two groups are overlapping on both the x- and y-axis, which means when we try to project the data to a one-dimensional space we have a harder time distinguishing the boundaries of the two groups, as we see when we plot the data and the PCA transformation for the data set:



How would the result change if you were to consider the second eigenvector? Or if you were to consider both eigenvectors?
Since the covariance matrix is symmetrical, the eigenvectors will be orthogonal, which means that if the data is nicely distributed on the first eigenvector, like in the first plotting we did, the data will overlap for the second eigenvector, and vice versa. It becomes obvious when drawing the two eigenvectors that the data would projects differently on to each of them.

Visualization

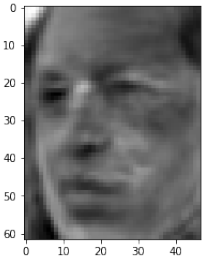
We have a data set with four features giving information of three types of iris flower. By plotting two and two features, we see how the different types of flower differ from each other in various ways, and it is not a clear cut way of representing the different groupings.
When we plot the results from the PCA we get the following:



By using PCA on the the data we simplify it and we get a single plot that easier to relate to. One of the flower types is clearly separate from the two others and we can assume that it has more unique features, while the two other types might share some features with each other.

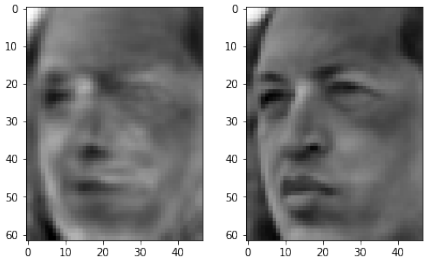
PCA for compression

We compress an image to 200 features and reconstruct it and we get the following result:



We see that we loose a lot of detail in the image, but the features of the face are still recognisable.

When we try other levels of compression we see how much information we loose or retain based on the number of features we keep. The images below are the compressed reconstructions of 100 features and 1000 features:

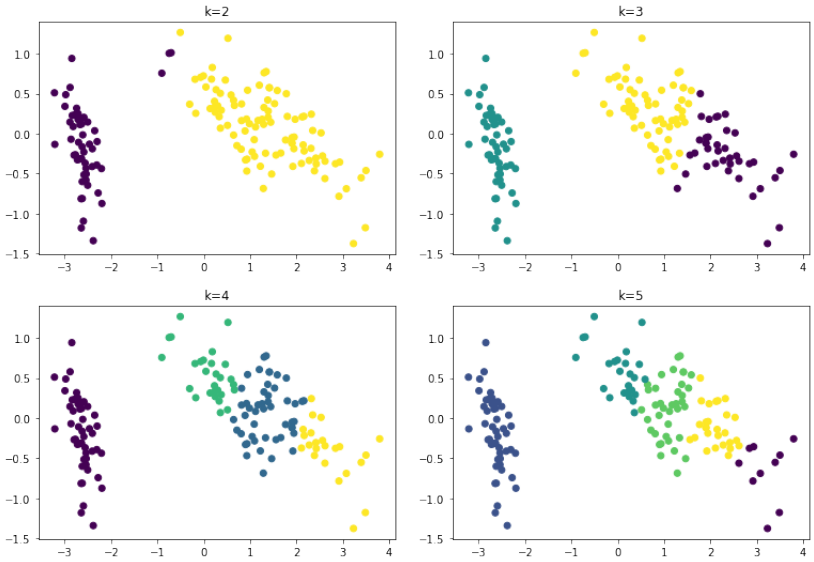


We see that there is a considerable difference in the level of detail retained in

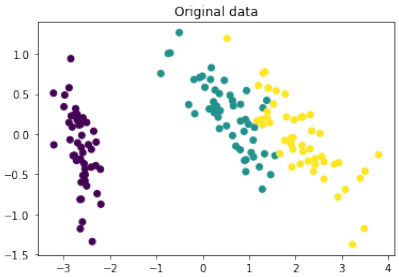
the two reconstructions. For the image with only 100 features, we get the vague outlines of features and completely lose an eye, while with the image with 1000 features we retain much of the detail from the original picture, but the image now has a slightly darker coloration.

K-mean clustering

We return to the iris data set to cluster the data with the k-mean method. The number of clusters we choose, will impact how well the data is grouped. Here we have the the k-mean clusters from $k = 2$ to $k = 5$:



We also have the original data with the three labels, corresponding to each species of Iris flower:



We see that the different numbers of clusters, give different descriptions of the data. For $k = 3$, which is we see that the k-mean algorithm does a fairly

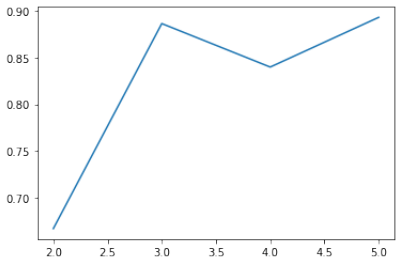
good job at finding the true clusters. For $k = 5$ there are too many clusters, but by combining the two green clusters and the yellow and purple clusters, we again get a fairly good estimate. For $k = 4$ we have that the blue cluster doesn't easily combine with either neighbour, because it eats away at both of the two original clusters and so it falls between two stools in terms of finding the original clusters. This will be evident when we study the accuracy.

Quantitative Assessment of K-Means

We now want to find the accuracy of the k-mean clustering for the different values of k. There was some problems with using the suggested function `accuracy_score()` because of dimensional errors. We have therefor used the `score()` function that is built in to the `LogisticRegression`. We get the following results:

```
Training set score: 0.96667
K: 2,  score: 0.66667
K: 3,  score: 0.88667
K: 4,  score: 0.84000
K: 5,  score: 0.89333
```

and when plotting the values for $k = 2 \ldots 5$ we get the following plot:



We see from these results that the accuracy goes up until $k = 3$, and then for $k = 4$ we see a drop, before it goes back up again for $k = 5$. This is caused by reasons discussed earlier when analysing the plots. The best accuracy is approximately 0.89 for both $k = 3$ and $k = 5$.