Intro to Machine Learning - Assignment 1

Maxmillan Ries s3118134, Cristian Rosiu s3742377

February 2020

Contents

1 PCA

1.1 Data Exploration

a The Iris data set provided by contains samples from 3 distinct types of iris plant. The kind of data recorded consists of a list of 50 measurements per plant, for 4 different properties. These properties are the sepal length, the sepal width, the petal length and the petal width.
The data is presented in form of 5 columns, 150 rows, with the first 4 columns matching the previously described properties, and the 5th column being the kind of the plant the results stand for.

Along with the measurements, a table is provided in the "iris.names" file, which contains the min value per category, the max value, the mean, standard deviation and class correlation.

b In broader terms, data variation, is a measurement of the spread of values inside a data set. Specifically, the variance measures the distance of each value in a set from the mean, allowing it to be used to find the distance of each value from each other (**FIND SOURCE**).

As explained in part a of the assignment, the mean values for the data set were given. These values were rounded to the second decimal point. To ensure our results were correct, we computed it after calculating the mean ourselves, comparing it to the numbers calculated using the given mean. In our case, the following variance values were found. Note that the V_ours stands for the values obtained using our mean calculation:

Name	V_ours	V_given
Sepal Length	0.6889	0.6811
Sepal Width	0.1849	0.1868
Peal Length	3.0924	3.0976
Petal Width	0.5785	0.5776

As we can generally observed, the values stand to be similar when calculating, or using the given mean. Specifically to the variance however, it can be seen that the Petal length has a very large variance, showing that it's data stands to be far apart from the mean. The sepal width shows itself to have the lowest variance, implying that most values of this data set are stuck close each other, and the mean.

Our hypothesis stands that an excessively high variance, or high level of spread, would result in the data being scattered more wildly. This would make it more difficult to find correlation patterns.
By this, we are referring to the simple manner of comparing 2 properties together, on a 2D scatter graph. **YO I AINT SURE ABOUT THIS SHITE MAN. Needs more bs**

1.2 Data Analysis: PCA

a As shown in the appendix the PCA algorithm driven from the pseudocode in the slides was implemented. Before evaluation in more detail the procedure, we would like to precise that some functions were taken from the Numpy library, namely the linear algebra functions. The reason for this choice to avoid the hassle of implementing such well established functions as to focus on the core of the assignment and algorithm instead.

The first steps of our procedure was to the take the iris data set and convert it in to a Numpy array as to facilitate it's manipulation. To do this, the as_matrix function was implemented, which reads the lines of the data file and inserts them into a Numpy array. The type of flower is not included in the Numpy array, as it is not needed for the PCA calculations.

The PCA algorithm begins by calculating the mean of each category, something which we chose to do for higher accuracy (instead of taking the mean given, as mentioned in part 1) and immediately centering the data. After some testing with plug-and-play, we found that centering the data was most important, as it ensures that the first principal component (PC) describes the direction of Maximum Variance. If the subtraction was not done, the first PC might have corresponded to the mean of the data instead.

Following this, the algorithm computes the covariance matrix before using it to calculate both the eigenvalues and eigenvectors. As there are no eigenvalues or vectors for a non-square matrix, the covariance matrix is necessary to calculate.

Finally, using f_r (as described in the pseudocode), the number of components for reduction is chosen. Unlike other algorithms, where the number of components is an input, our algorithm uses the degree to which the variance should be preserved as an input, and calculates the minimum number of components required for it. The main idea of the algorithm is to keep the highest amount of variance while using the minimum amount of data, hence why the minimum is found.
Using this minimum number, the subset of eigenvectors is taken and the reduced dimensionality data is calculated. By default, we have chosen to use 0.95 (95%) for the variance, resulting in a number of components equal to 2 for this data set.

As we can see from the graph below, which is colored coded with red being the Iris Setosa, green being Iris Versicolor and blue being Virginia, the data offers the same conclusion as the given information, that one of the clusters is separable from the two.
The Iris Setosa's data is clustered closer to the **XXXX** values, while the other two iris species are clustered more towards the **YYYY** values.

YO LETS PUT A 2D GRAPH HERE AND SAY something like "as you can see, 2 clutters are formed which look dope and clearly show the correlation yada yada. To outlier is X and then in later parts we can refer to this bit."

1.3 Dimensionality Reduction Evaluation

a What we can expect from reducing the dimensionality, is a decrease in the average euclidean distance within a single class. In a perfect, ideal transformation, the average euclidean distance between all points would be unchanged, though in reality, a small reduction is expected, the difference being proportional to the number of dimensions reduced.
As we can see from the data we obtained, this expectation was correct. From 4 PC (original data), to 1 PC, the average euclidean distance is continuously reduced, with a large proportion of the information being lost when reducing down to 1 PC.
insert data here

When comparing the data sets to one and other, we observed that the average euclidean distance between one set an another is larger than the distance between the members of a set. This of course makes sense. As

the sets of data belong to different cluster, the average distance between both clusters is higher.
To calculate this difference, we chose to take the average euclidean distance of each pair of 1 class relative to another, and take the mean of those 50 values, as we felt that was the most concise, yet still effective way of presenting the information.
insert data here

One important note about Dimensionality Reduction, is that the long pair euclidean distance is preserved more than the smaller, closer pairs. The reasons for this is that the outliers in the data remain further apart from one and other, and reducing the dimension of the data does not remove too much of the euclidean distance. For small values which might have large distance along a collapsed axis, the distance would change from significant to otherwise.

For the different level of reduction, the following alpha values were used.

- 1 Principle Component. Alpha = 0.92.
- 2 Principal Components. Alpha = 0.93.
- 3 Principle components. Alpha = 0.98.
- 4 Principle Components. Alpha = 1.0.

We can generally see that a majority of the information can be obtained using a single principle component. However, when comparing the average euclidean distance, the values seem significantly more distant from the initial values (4 PC's).

b As we have discussed before, the Iris Setosa seems to be separable from the other two plant species, as the data is clustered more towards the **XXXX** values. An interesting observation between the data obtained in part e is that the Iris Setosa's average euclidean distance is smaller than that of the other two plants, showing that the data is less spread/scattered. The data seems to be more packed towards one and other, which means that the data is distributed closer to the mean than the data of the other two plants.

Looking at the 3 dimensional graph, the same conclusions can be made.
Overall, the data follows to some extent what we were expecting. Should the data of each plant have varied too much, it would have been difficult to find a pattern between the plants. But overall, the data competently shows the separable plant from the 3.

c Applying the same procedure to g, we can see that the results are nearly identical. Due to the arbitrary choice of +/- for a PC, the graph is flipped along the x and y axis, though the results otherwise show exactly the same as our own procedure.

2 Individual Work

Throughout this assignment, both members of the group did equal amounts of work. Both individuals worked on the code, and on the documents. There were no issues when working together, everything went smoothly.