Machine Learning assignment 4

Mike Wang (s2538334)

2 May 2021

Question 1: why can the two-dimensional data we stored in X be reconstructed so well using only one principal component?

This is because the principal component stores the relation between the two values belonging to each point in X, and because the matrix was rescaled and recentered before the principal component was calculated.

Question 2: make a scatter plot that contains true and reconstructed values ($k = 1$), as well as the single principal component.

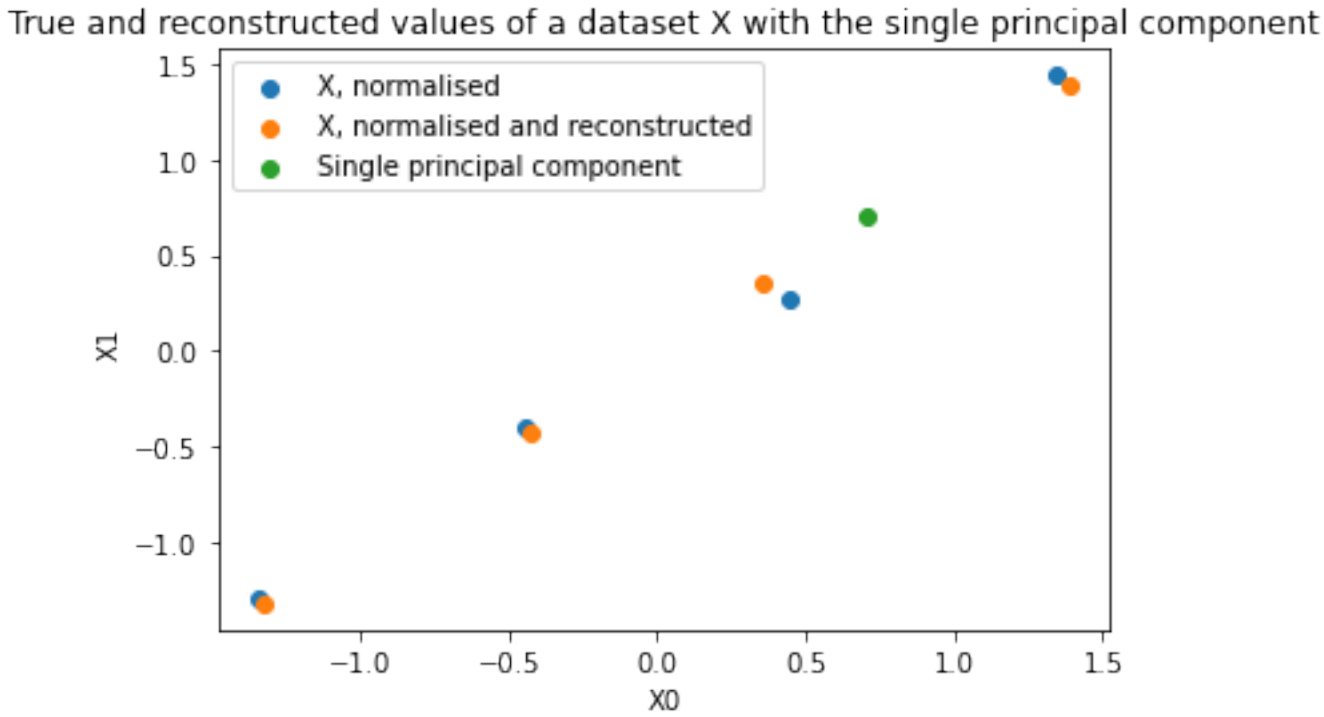

Figure 1: Scatter plot containing the true and reconstructed values of X. The single principal component can be seen. Note that values for X have been normalised.

Question 3: can you come up with a feature transformation (for one or more column(s) of `big_X` by looking at `X_train`) that would make the principal components express more of the data's variance?

There are some columns with a very low variance (value smaller than 0.0001). We could scale those features so that they weigh less than the features with a high variance, by for example scaling and centering them different, so that the PCA express more of the data's variance.

Question 4: what is a good number of principal components to continue with *and why*? (Base your answer only on this training set.)

The amount of variance expressed in each component is given by the r-squared value, with the total of all r-squared adding up to 1. If we want to have the optimum number of principal components, we should plot an elbow plot and look where the elbow makes its sharpest turn.

This can be seen in the above plot to be at k=4, or about 63 percent of the total R squared value. From this we conclude that a good number of principal components to continue with is 4.

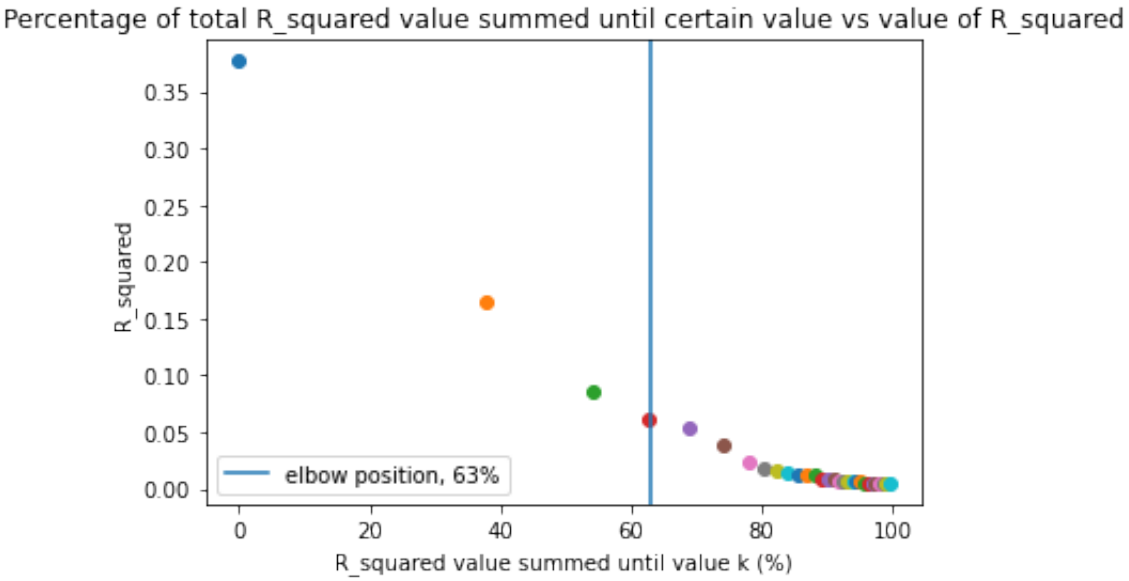

Figure 2: Scatter plot comparing the value of a single R squared value with the summed total of all R squared values up to that point, expressed in a percentage of the total R squared value. We can analyse this plot like an elbow plot, and it can be seen that the sharpest turn in the elbow is at the fourth point.

Question 5 (2 points): how do the computation time and accuracies differ between the repeated evaluations with and without PCA, and how do you explain this? (Specify k if you do not use the value you chose above.)

The calculations with PCA tend to be faster, which is explained by the fact that the PCA simplifices the dataset by reducing the dimension, so that less complex computation is required. The accuracy is lower though, which is the cost of this simplification.

Question 6 (bonus): test your theory.

Question 7: can you describe a situation where (or model for which) PCA would not help?

A situation where PCA would not help is a dataset where one would expect the different features not to be related. As PCA reduces the dimension of a dataset by calculating a relation between the different features, it does not make sense to do so where there are no relations expected to be found, and might even cause the model to lose efficiency.

Question 8: make scatterplots of (either train, test, or all) points expressed in terms of 1: the first two principal components; 2: the first two numerical features.

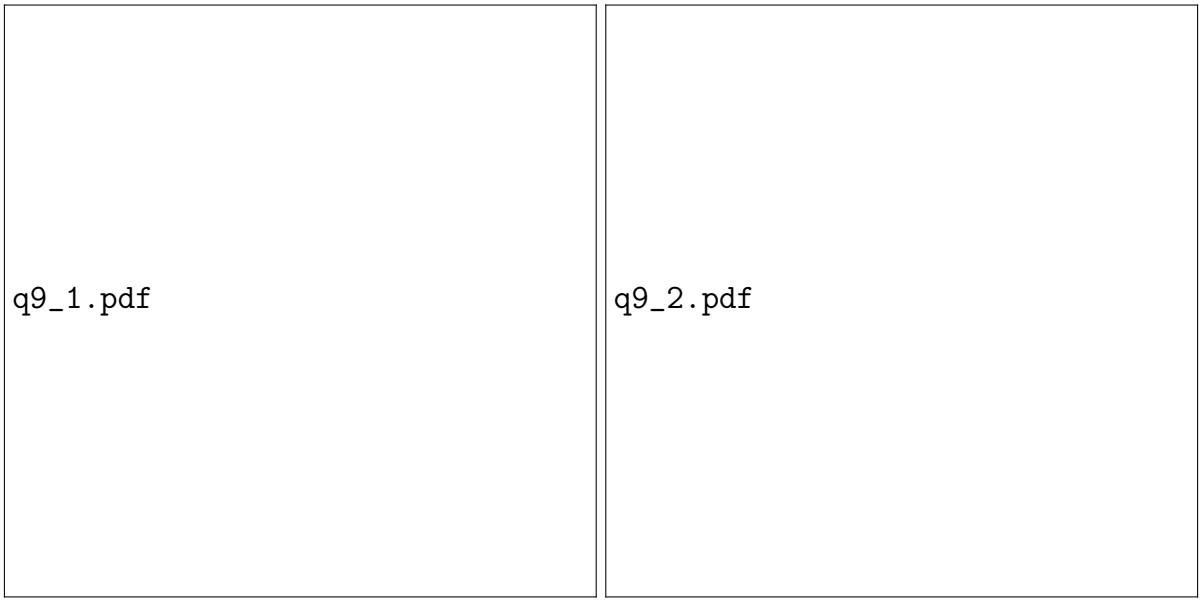

Figure 3

Question 9 (2 points): what do the first two principal components represent, when you think back to what the numerical features are based on?

Those represent the features with the two highest variances. For this dataset, those features are number 6 and 26 as can be seen by using np.nanvar on the columns of the dataset. This corresponds to the mean of the compactness and the mean of the three largest concavities.