Data Mining - Assignment 1

Simon Min Olafsson and Nadia Günther

October 2023

1 Data Mining Questions

- Can we identify distinct groups of students by analyzing heights in inches and shoe size?
- Can we predict a student's study programme by counting how many words they used for their reasoning on why taking the course, their shoe size and height?

2 Pre-processing

2.1 Normalization

Our initial thoughts on the preprocessing part was to normalize our data as there were a lot of inconsistencies present in multiple columns. We transformed the tab seperated file into a CSV file for better visualization and manipulation with the pandas framework.

The next step was to transform the string columns of the students study programme and the "why" column. As our data mining questions was to predict a students study programme, we decided to make it a target column. Therefore, we converted the categorical string arrays into numerical values ranging from 1-4 depending on the students programme. We modified the *Why* feature to get the word count by tokenizing the sentences and replaced contractions with the actual words.

We also normalized the numerical values of height, shoe size and number of written words to be between 0 - 1 values to make the distance between the features more comparable. At last, we computed the z-score for each column to remove possible outliers and evaluated if it made sense to remove it, since our dataset is very small.

2.2 Missing values

We saw that the "seattle" column was missing over 50 percent of its values. As the dataset is very small, we calculated the mode (most frequent value) of the existing values and used `DataFrame.fillna()` to fill out the missing values. Even though the column is filled by the end of the preprocessing, the column was not gonna be utilised to answer our exploratory questions as over 50 percent of the values contain automatically filled values, which is not very valid nor representative.

3 Clustering: K-means

3.1 Feature selection

For convenience reason, we have limited the feature selection to numerical features:

- Shoe size
- Height in inches

3.2 Implementation

We chose k-means clustering as we chose to limit our feature selection to numerical data.

The k-means algorithm assign random centroids inside the dimensional matrix of our datapoints and as the iterations continue, the assigned datapoints will reassign the position of the centroids according to a newly calculated mean of the datapoints in each iteration. The elbow method is used determine the optimal number of clusters. We found that 3 cluster was most optimal. The visualizations can be seen in results.

It seemed reasonable to use the Euclidian distance as it resembles a straight line between two datapoints / shortest path between two datapoints. Manhattan distance were also a viable option, though it measures the distance differently as it's the sum of the absolute differences of their cartesian coordinates.

3.3 Result

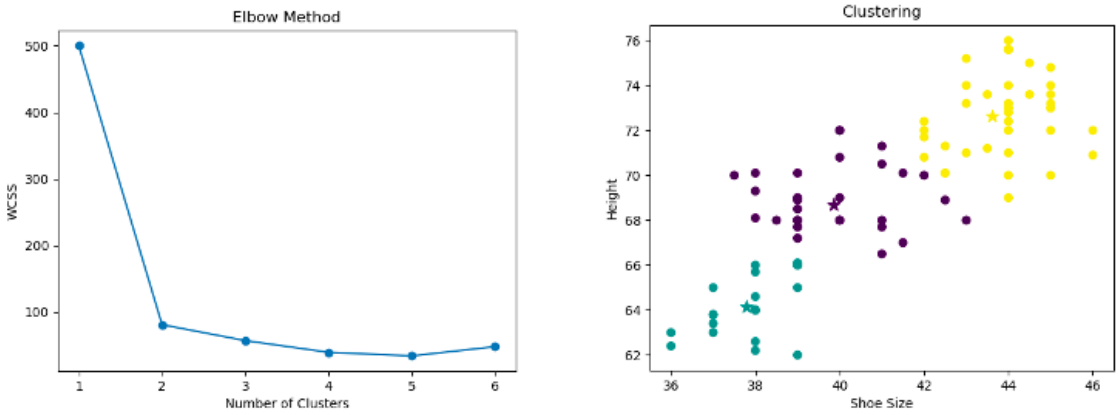

Figure 1: Elbow method plot and k-means clustering with $k = 3$

The elbow method plot on the left and the k-mean with $k = 3$ on the right scatterplot.

We then specified a new column to be used in our predictive model with the cluster labels for each datapoint.

4 Supervised learning: Naive bayes

4.1 Feature selection

For this part we also used a limited amount of features:

- Programme (target)
- Shoe size
- Height in inches
- Number of words used in the why column

4.2 Implementation

For this supervised learning part, we decided to use naive bayes, using the *Programme* feature as target.

We implemented it by using its probabilistic principles. By calculating conditional probabilities, the model effectively assigns predicted class values to the datapoints.

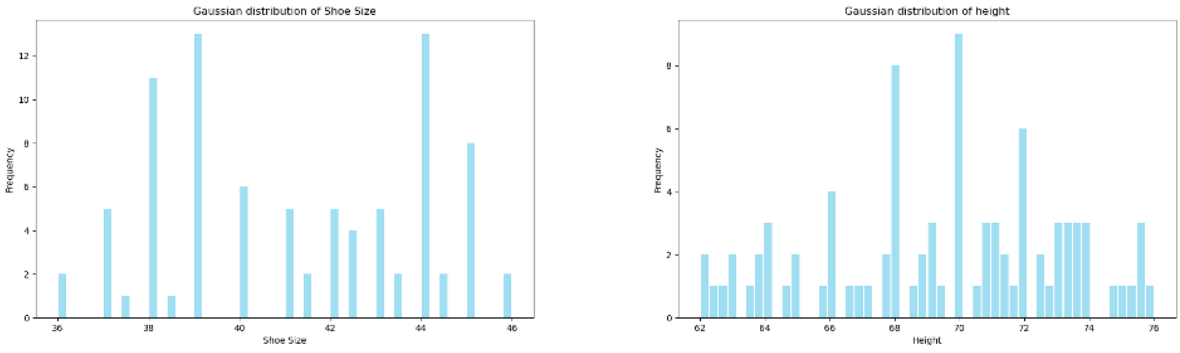

Figure 2: Gaussian distribution of shoe size and height

There are different kind of naïve bayes classifiers, and we chose the gaussian classifier, as it assumes gaussian distributed data (normal distribution) as shown below. We are aware that the data are not perfect normal distributed data, as the shoe data is bimodal, but it was the most suitable classifier dependent on our data.

In our division of the dataset, we defined the training data to contain 80% of the samples to prioritize the training, as it is a very small dataset. As a result, our predicted accuracy is smaller on the test set than the training set but should reach a "breakeven" point with the calculated training accuracy, if the test set was increased as much.

4.3 Results

Test accuracy	Train accuracy
70.59%	77.14%

Table 1: Test and train Accuracies

5 Conclusion

The preprocessing was a vital step into normalize and prepare out data. We went with different kind of solutions and decided to drop the timestamp and "seattle" columns, as our questions didn't relate to those. Also we tried different solutions in both the supervised and unsupervised methods (k-means and NB) with different numerical distances between the datapoints with no significant difference.

For our first data mining question, the k-means clustering algorithm was be very suitable for our continues numerical data in our height and shoesize columns. By using the plot library from matplotlib, we could see that our data was naturally segregated into three clusters, visualizing three distinct groups of students based on their physical attributes.

For the second datamining question, our gaussian Naive Bayes classifier achieved a test accuracy of approximately 70.59% which is reasonable. We still believe that we could have achieved higher accuracy with a bigger dataset as we prioritized the training data over our test set, as the whole sample set only contained 90 rows.

6 Note:

NB: As the report exceeds the specified two pages, we declare the number of used units:

Total units in rapport without space: - 4664

Total units in rapport with space: - 5580