Contents

This report includes 7037 characters, including content-overview and printed results from the algorithms. It is 6275 actual text.

Introduction

This report is documenting a micro-scale data mining project that was conducted within the course Data Mining. The foundation of this report is a dataset that was produced by the students of this course in the first lecture, where they were presented with a survey containing X questions. The resulting dataset was then handed over to the students to formulate research questions and conduct research using algorithms that were taught within the course. The work that was done during the lab-sessions will be used to carry out the remaining work.

Dataset: Getting to know your data

The dataset contains all possible attribute types and has not been pre-processed before it was handed over. For this report it was not mandatory to process the entire dataset, but only a chosen set of attributes.

For later calculations I created methods for the following statistical descriptions in the Student Class:

- getSum
- getMean
- getMedian
- getStandardDeviation

Research Questions

1. For a given game X, how likely is it that another game Y is played by the same person?
2. Based on age, height and shoesize, how good are predictions towards gender?
3. Is there a pattern in the distrubution of age and shoesize?

Based on these research questions the following attributes were chosen:

Age – Gender -Shoesize – Height - Which of these games have you played?

Consequently only this set of attributes were preprocessed and used within the algorithms.

Pre-Processing

For the data cleaning I created a class called “PrettyMaker”. Data cleaning that has been conducted:

- Removing double quotes
- Replacing commas with dots (mainly for shoe-size)
- Parsing and saving the data correctly
- Age:
    - 0 is considered INVALID_NUMBER
    - Age below 15 and above 70 is considered INVALID_NUMBER
    - Both are replaced with the median for age
- Gender:
    - Takes the first letter of the answer, checks if it is “m” or “f” and if not it replaces with “m”.
    - If gender is netiher “m” nor “f” it is repalced with “m” (knowing there is more “m” than “f”)
    - Normalisation: “m” is replaced with 1 and “f” is replaced with 0
- Shoesize:
    - Shoesize below 35 and over 50 is replace by media shoesize
- Height
    - Height below 100 and above 230 cm is replaced with the median height
    - Cleaning “172cm” by removing everything that is not a number
- PlayedGames:
    - If someone answered “I have not played any of these games”, the answer is not added to the games-array that is used for apriori.
    - Splitting at every semicolon, using StringEnumerator to convert the nominal into a numeric attribute (Fifa 2017 = 0 etc.)
    - Saving numeric representation of a game in an array to allow processing in Apriori
- The method *removeFishyStudents* removes students that have age, shoesize and height replaced with the median (which means they entered invalid input in all those three questions); thus they not relevant for my algortihms and are excluded from the studentlist.

Algorithms

Frequent Pattern Mining: Apriori

To answer question one,

For a given game X, how likely is it that another game Y is played by the same person?

A frequent pattern mining technique called Apriori was implemented. The implementation of Apriori was quite time-consuming. The generation of candidate-1-itemset C1 and frequent-1-itemset L1 had to be done differently from the following reoccurring steps. Especially the joining was difficult because it required a lot of if-statements and nested for-loops.

The support-threshold was increased step-by-step, starting with 7 and increasing it up to 20 to actually cut down results to noteworthy and easily observable trends in the dataset.

The result is showing 5 frequent-2-itemsets and no frequent-3-itemsets. Only results with a confidence of more than 60% are included in the output:

72.41% of the students who played Wordfeud, also played Angry Birds.

60.61% of the students who played Counter Strike Go, also played Minecraft.

69.70% of the students who played Counter Strike Go also played Angry Birds.

62,50% of the students who played Minecraft, also played Counter Strike Go.

65.62% of the students who played Minecraft, also played Angry Birds.

77.42% of the students who played Candy Crush, also Angry Birds.

Classification: K-Nearest-Neighbour

In order to provide an answer to question two,

Based on age, height and shoesize, how good are predictions towards gender?

I decided to implement k-nearest-neighbor, a lazy-learner classification model. For this purpose, the student-list was split into test- and trainingset. Based on age, height and shoesize, the Euclidean distance between the so called test-student and the student-for-comparison was calculated and saved. If more than half of the close-students are male, we assume the student is male. Then it is checked, if the student is actually male or not.

As a result I got a 100% success rate, meaning that only True-Positives were printed out. Especially considering that there were less than 10% females in the dataset that is a very good success rate.

Compared to Apriori, k-nearest-neighbour is less time-consuming to implement because it requires basically only two methods for calculating the distance and making assumptions.

Clustering: K-Means

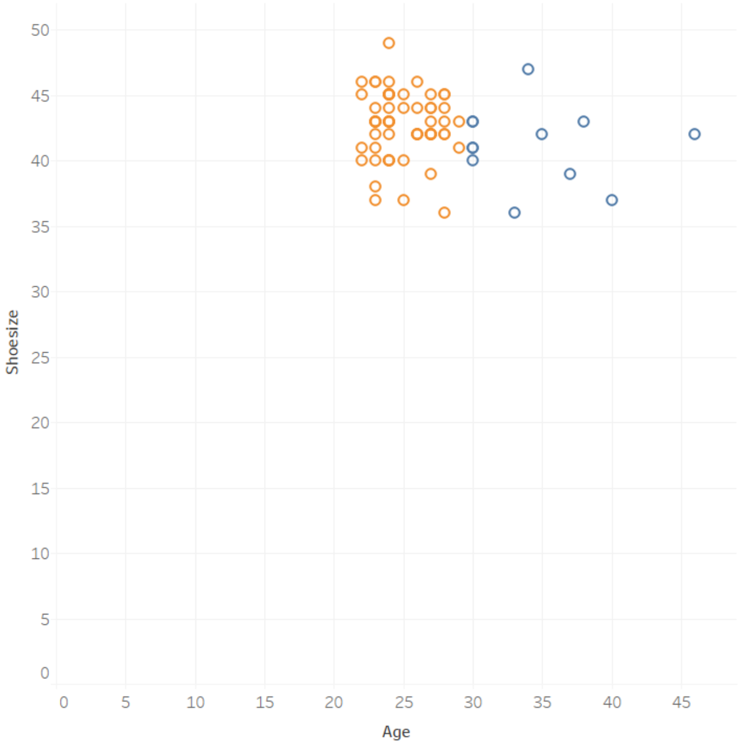
For the third research question,

Is there a pattern in the distrubution of age and shoesize?

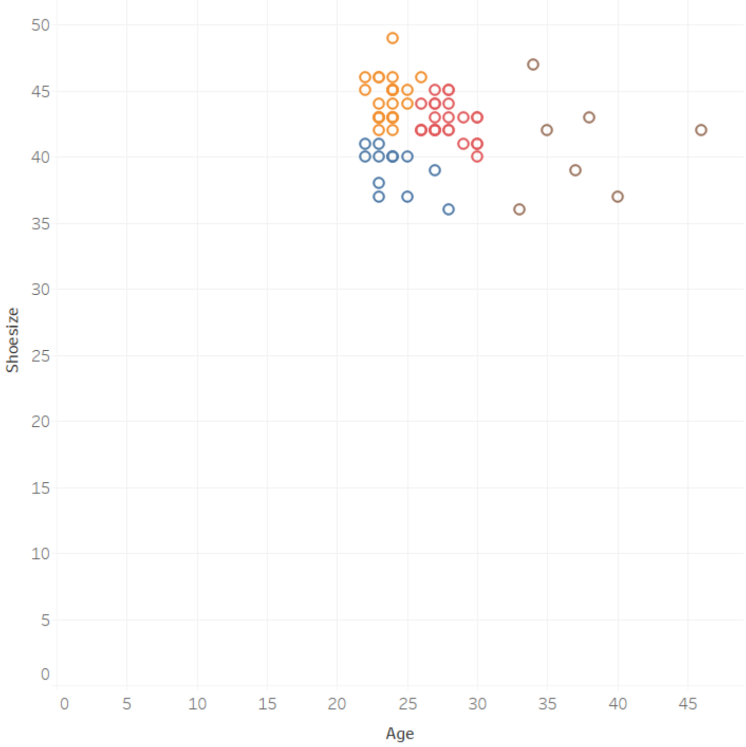
I decided to use the clustering algortithms K-Means. K-Means creates a k number of clusters from at least two attributes. Random students are placed in k initial clusters and their centroids are calculated. Every time a student is about to be added, the distance to both existing clusters and their centroids is calculated; the student will then be added to the cluster which is the closest in terms of euclidean distance of chosen attributes. The process continues until the adding of new students does not cause a shift of the centroid any longer.

K-Means was the easiest to implement because it is based on only two calculations and relatively easy to comprehend. The most difficult part within was the calculcation of (euclidean) distance between cluster one and two.

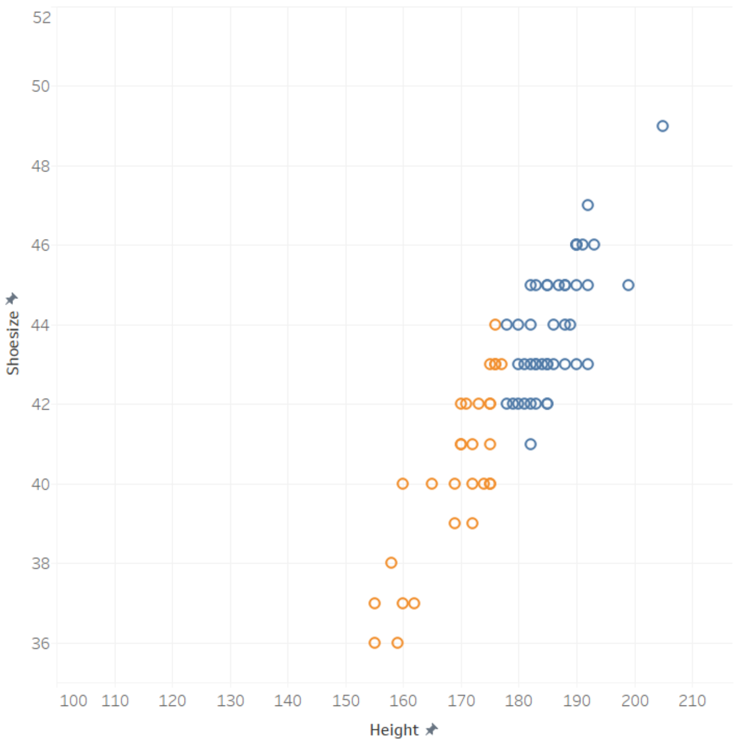
The result can be seen below. The two clusters seem to be mainly separated across age-distribution (students below 30 and students above 30), however not so much along the shoe-sizes. Apart from that this cluster has not helped me to make new discoveries.



I also tried 4-Clusters to compare with, but the result is not very useful from my perspective. It might be more useful with a very large dataset that includes a broader range of ages so that patterns of shoesize across different generations could be found (e.g. 60-year-olds having bigger feet than 20-year-olds). Also a possible cluster between women and men would probably be more likely to emerge if there would be more data for women.



Another execution using height and shoesize.


Shoesize
Height