Data Mining - Mandatory Individual Assignment

Irina Alina Gabriela Luca (irlu@itu.dk)

This report contains a total of 4840 characters without table results and 6365 characters with them.

April 16, 2017

1 Data pre-processing

1.1 Attribute extraction and selection

The dataset contains several attributes, out of which I selected and pre-processed the following: age, height, shoe size, gender, study program, phone OS, known programming languages, played games, the picked number and the set of 4 picked numbers in range [0,15] and the reason of picking this course. Additionally, two new attributes have been inferred: the number of played games and the number of known programming languages.

1.2 Data cleaning and transformation

Data cleaning and transformation are briefed only for some of the pre-processed attributes, as such:

- age: values were sorted and 2% from both ends were removed (e.g.: the outlier age '999' got excluded)
- height: both outliers and string attachments (e.g.: '179cm', '55') were removed
- shoe size: values were mapped to float after I made sure the decimal delimiter used was '.' and each value was bigger than a chosen minimum value
- gender: values were lowercased and turned into 'f' or 'm'
- study program: whitespaces were stripped and program names were reduced into more compact categories
- programming languages: values were lowercased and languages were split, sorted, stripped and list-mapped
- course reason: values were mapped to a specific letter for easiness (e.g.: 'I am interested in the subject' => 'i')
- numbers picked in range [0,15]: values were split and only those consisting of 4 numbers were selected, according to the original question

1.3 Data normalization

In order to assign equal significance to each attribute, data has been standardized with min-max normalization, such that all values are mapped to values in the range [0,1]. Therefore, the euclidean distances calculated between the numerical values used in clustering, for instance, will no longer be dominated by one specific attribute (all attributes will equally contribute to the computation).

2 Applied Methods

2.1 Classification (with k-Nearest Neighbours)

'Can the study program of a course attendant be concluded if taking into account his/ her gender, the reason why he/ she picked the course, how many programming languages he/ she knows and the number he/ she picked?'

After the data columns for the above mentioned attributes are merged, a dataset containing 55 tuples is given as input to the k-Nearest Neighbours Algorithm, out of which I use 60% for the training set, 30% for the test set and 10% for the validation set. Before splitting, data is shuffled, so that attributes are equally distributed in the different partitions and the model is more reliable. Since the data selected for answering this question implies both numerical and nominal attributes, I treated the distance between them accordingly (e.g.: for nominal attributes, I added 1 to the deltas' exponent of 2 if they are far from each other). The validation set is afterwards used in order to find the best 'k', where k is tried out for the range [5, 15), with the following results for two distinct runs:

Results for one run	Results for the other run
1. Validation phase: (k, [accuracy percentage, correct predictions, total made predictions]) => (5, [100.0, 6, 6]), (6, [100.0, 6, 6]), (7, [100.0, 6, 6]), (8, [100.0, 6, 6]), (9, [100.0, 6, 6]), (10, [83.33, 5, 6]), (11, [83.33, 5, 6]), (12, [83.33, 5, 6]), (13, [66.67, 4, 6]), (14, [66.67, 4, 6])	1. Validation phase: (k, [accuracy percentage, correct predictions, total made predictions]) => (5, [83.33, 5, 6]), (6, [100.0, 6, 6]), (7, [83.33, 5, 6]), (8, [100.0, 6, 6]), (9, [100.0, 6, 6]), (10, [100.0, 6, 6]), (11, [100.0, 6, 6]), (12, [83.33, 5, 6]), (13, [83.33, 5, 6]), (14, [66.67, 4, 6])
2. best k => 5	2. best k => 6
3. Test phase: [accuracy percentage, correct predictions, total made predictions] => [93.75, 15, 16]	3. Test phase: [accuracy percentage, correct predictions, total made predictions] => [87.5, 14, 16]

2.2 Pattern Mining (with Apriori)

'What is the probability that a course attendant knows particular programming language(s) if taking into account what language(s) he/ she already knows?'

'What is the probability that a course attendant picks specific random number(s) in the range [0, 15]' given some other random number(s) he/ she picks as well?'

Apriori Algorithm was applied and the association rules below were revealed. The attributes used are the known programming languages in the first case, and the 4 picked numbers in the second case. The assumption for the first case was that there must exist patterns in terms of students' programming knowledge, while the second case is more based on the fact that the numbers were picked from a limited range. The tables below show results retrieved for both questions, where the support chosen for programming languages is s=10, with P(B|A) bigger than 0.5, while the support for the numbers picked is s=6 with P(B|A) bigger than 0.3 .

A => B, P(B|A) for programming languages	A => B, P(B|A) for numbers picked
1. [...] ['c++'] => ['c#', 'java'], 82.60% ['f#'] => ['c#', 'java'], 84.61% ['javascript'] => ['java'], 94.73% ['c#'] => ['java'], 96.67% ['python'] => ['java'], 90.90% ['c'] => ['java'], 100.0% ['c#', 'c++'] => ['java'], 95.0% ['c#', 'javascript'] => ['java'], 100.0% ['c', 'java'] => ['c#'], 66.67% ['c', 'c#'] => ['java'], 100.0% ['c++', 'java'] => ['c#'], 86.36% [...]	1. [...] [4] => [1], 43.75% [1] => [4], 31.81% [1, 3] => [2], 69.23% [1, 2] => [3], 90.0% [2, 3] => [1], 100.0% [3] => [1, 2], 42.85% [2] => [1, 3], 45.0% [1] => [2, 3], 40.90% [3] => [2], 42.85% [...]

It seems plausible that 95% of the students who know C# and C++ also know Java, just as well as it seems plausible, by looking at the second result set, that all the students who picked 2 and 3 in the 4 chosen numbers, also picked 1.

2.3 Clustering (with k-Means)

'Can a course attendant's gender be implied if taking into account his/ her shoe size, age and height?'

The starting point for this question relies on the fact that a correlation between students' shoe size and their height might exist. However, these two were merged with the data for age, as well, still having the expectation of concluding with two clusters, one representative for men and the other one representative for women. The input given to the k-Means Algorithm consisted of 58 data tuples, which were clustered in two clusters, whose representation is shown by the figures below.

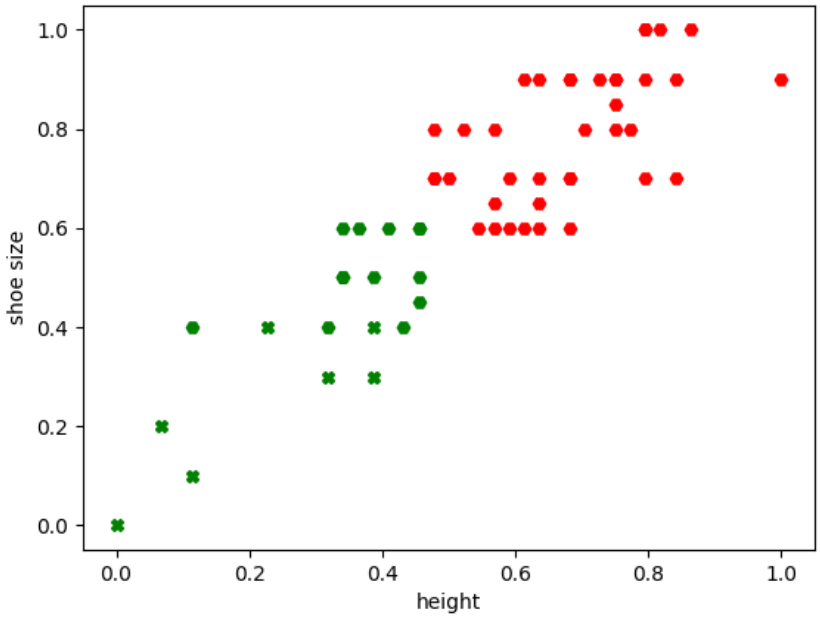

Figure 1: 2D representation of the two clusters, where the red hexagons are all males collected in one cluster, while the green points mix both males (hexagons) having lower to average shoe size and height and females (crosses); it seems plausible that the green cluster mixes both genders and also that a female has the smallest proportion between shoe size and height of all students.

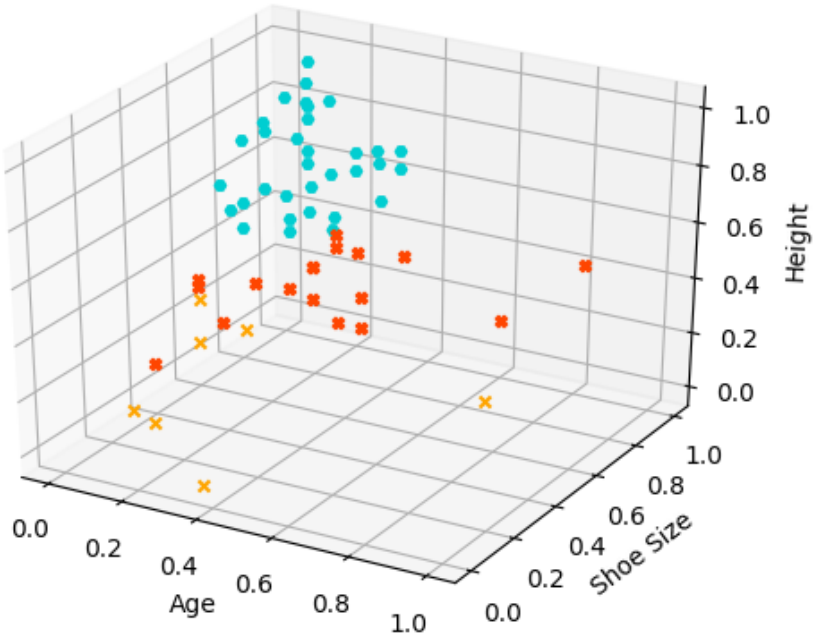

Figure 2: 3D representation of the two clusters, where the turquoise points above are all males collected in one cluster, while the orange-red and orange points mix both males (orange-red) and females (orange); the positive correlation between shoe size and height is visible in the 2D representation, but the 3D representation shows that the attribute age has a smaller influence than the other two attributes have in the whole process.