Dennis Thinh Tan Nguyen
dttn@itu.dk
Character count : 5282 | 2.2 pages

Individual Assignment

In this individual assignment, we are to analyze a dataset based on a questionnaire the students had filled during the first lecture.

Thus the following questions have been created which we will try to anwswer:

- Based on the dataset is it possible to predict which phone OS a given student use?
- Based on the dataset is it possible to find patterns based on programming languages such that one may find how they may relate to each other?
- Based on the dataset is it possible to find a relation based on the students' height, shoe size and age?

Preprocessing

Based on the given the dataset, it consists of data entries from many people. Thus the dataset may not be entirely consistent and error-free. By example, all attributes entered manually are mostly inconsistent and contains errors. The non-numeric attribute **Gender** consists of many different variations such as *male, man, m* which makes the attribute inconsistent. Also, the input may be incorrect such as *King Gizzar…, make, fluid* or even left empty.

For numeric attributes such as **height**, some input the data as decimal numbers if it is in meters, *1.75* while others input them as integers if it is in centimeters, *175*. Some also input the data in ranges such as 40-42 for **shoe size**.

For this assignment, I have selected a subset of attributes that may be interesting to use: **age, height, shoesize, gender, degree, phoneOS, playedGames, WhyCourse, Programming Languages** and **Commute**

Missing Value

For each non-numeric attribute, any missing values will be assigned a default constant **null,** and for numeric attributes, 0 will be used as default.

Since the dataset is rather small, some attributes were also filled manually if the error was obvious. For example, the attribute **age** had a case where 23 was written as twenty-three. These errors may be considered as rare and was corrected manually. However, If the dataset was huge this solution may not be applicable.

Normalization

For normalization, all non-numeric attributes were set to lower case and trimmed. Also, inputs were replaced with a consistent value. By example, all values of male gender must be "male," thus if the string contains man, male or anything related to male, we transform it to male or else use null. For numeric values, we take average of ranges, remove all non-digits, except for "," or ".". The floating point is transformed to ".". Finally, extreme cases of age, shoesize, and height is set to 0.

Dennis Thinh Tan Nguyen
dttn@itu.dk
Character count : 5282 | 2.2 pages

Supervised learning with K-nearest-neighbor

To answer the first question on how to predict phoneOS, we use the K-Nearest neighbor algorithm. To predicting phone OS, a student is likely to use based on his data and the trained model. Distance is calculated based on the attributes: **Why course, Gender, Degree**

Data Size: Total size: 84 | Training size: 74 |Test size: 10

Results when predicting Phone-OS: Android on unclassified students

Actual Label	Expected Prediction	Actual Prediction
Android	$\mathrm{Android} * 6$	$\mathrm{Android} * 6$
iOS	$\mathrm{Null} * 3$	$\mathrm{Null} * 3$
Windows	$\mathrm{Null} * 1$	$\mathrm{Null} * 1$

TP	TN	FP	FN
6	4	0	0

Accuracy %	Error Rate %	Precision %	Specificity %	Sensitivity %
60.0	40.0	100.0	100.0	100.0

Based on the results it can be seen that all the test data has been predicted correctly.

Frequent pattern mining Apriori

To answer the second question to find any patterns in **programming languages**, we use the Apriori algorithm. Below is the results of all item sets with a confidence greater than 90 percent

Total Size: 84 | Support Threshold: 14 or total transactions/6

Item set	Confidence %
$\{c\#\ c++\} => \{java\}$	90
$\{c++\ java\} => \{c\#\}$	90
$\{c\#\ javascript\} => \{java\}$	91
$\{f\#\} => \{c\#\}$	100
$\{f\#\} => \{java\}$	100
$\{c\} => \{java\}\vert$	100
$\{c\ c\#\} => \{java\}$	100
$\{c\ javascript\} => \{java\}$	100
$\{c\#\ f\#\} => \{java\}$	100
$\{f\#\ java\} => \{c\#\}$	100
$\{f\#\} => \{c\#\ java\}$	100

What can be seen that almost all item sets lead to Java. Thus, it may be expected based on the data that students know java if they know f# or C#. This is especially true if the students have done their undergrad at ITU since this is the first language they would learn, while F# and C# are other languages you will learn during your undergrad at ITU.

Dennis Thinh Tan Nguyen
dttn@itu.dk
Character count : 5282 | 2.2 pages

Unsupervised learning with k-means

To classify the students based on the relation of the students' **height, shoe size, and age** we use the k-means unsupervised learning algorithm to group the data into clusters. For this assignment, 6 clusters are to be generated, and the results are as below:

Cluster	Age	Height	Shoe	Class / commonality
1	21-28	160-173	169-173	Young small (f + m)
2	22-28	178-187	178-187	Young medium (m)
3	22-29	188-195	188-195	Young large (m)
4	30-38	181-186	42-47	Old large (m)
5	23,31	166-173	0	Outlier with some data
6	0,0,27	0	0,41-44	Outlier with few data

The algorithm has created 3 clusters of young students between 21-29 where each cluster is based on their height and shoe size. The first cluster consists of "small" people and usually a mix of male and females. The second cluster consists of people of "medium" size and the third consists of "large" people. The fourth cluster consists of all students above 29 years old. Since there is only a few of them only a cluster is created here. Finally, cluster 5 and 6 contains all the data entries that were missing data and are thus considered as outliers.

Based on these clusters one can further apply other data mining methods on each cluster to extract further information.