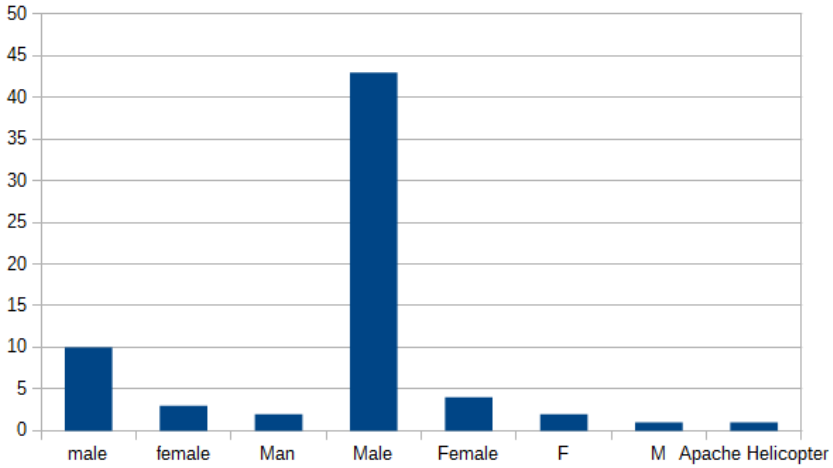
Thomas Roberts

Pre-processing

I created a set of functions that can be used to examine and clean attributes before each data mining method can process them.

1. A histogram function counts the frequency of nominal values in an attribute (see histogram – right) to be able to detect any discrepancies or inconsistencies in labelling. This information is used to transform the data to be more consistent (using domain knowledge)
2. For numerical data, min-max normalisation is used to give attributes equal weighting for distance measurements
3. When calculating the mean of data (such as in k-means), a 10% trimmed mean is used instead of the arithmetic mean to account for outliers in the data



Frequent Pattern Mining – Apriori

Question: Which groups of programming languages do people know?

Firstly, I needed to extract each individual programming language from a string of languages. As this was free-text input, the data was quite unclean. I created a function that splits a string along a given list of delimiters using a regular expression. Each language is then stripped of any trailing/leading whitespace and converted to lowercase. I created a custom function to do this because some languages include symbols in their name (which the standard string conversion to lowercase can't seem to handle).

Results: With 20% minimum support (~13 responses), the results are:
L1 = [{c}, {c#}, {c++}, {f#}, {java}, {javascript}, {python}]
L2 = [{c, java}, {c#, c++}, {c#, java}, {c#, python}, {c++, java}, {f#, java}, {java, javascript}, {java, python}]
L3 = [{c#, c++, java}, {c#, java, python}]

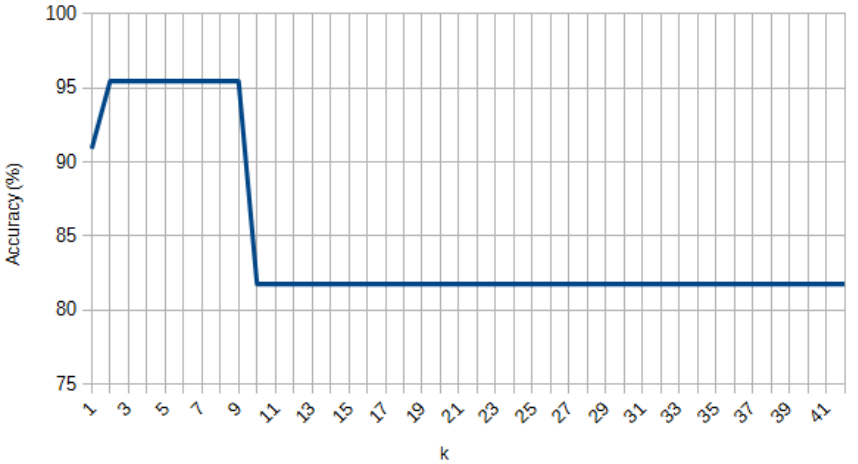
The results show that people most commonly know C#, C++, and Java, or C#, Java, and Python.

Thomas Roberts

Supervised Learning – k-Nearest Neighbours

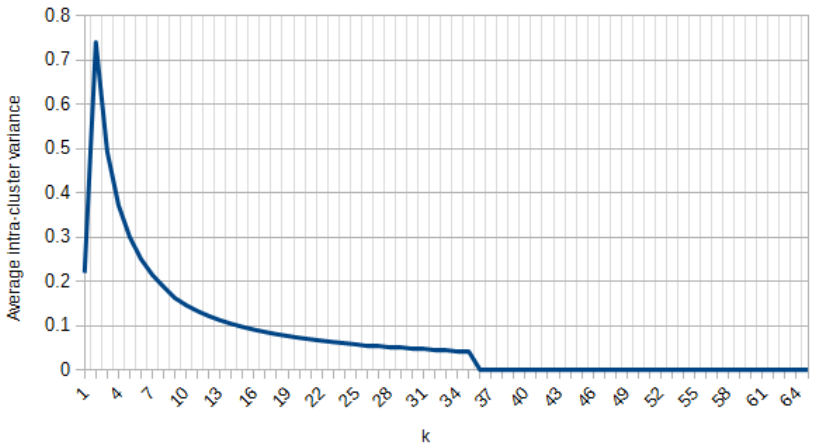
Question: What is a person's gender based on their shoe size and height?

I extracted/cleaned the class label (gender) and the predictor attributes (shoe size and height). For the gender, I used discrepancy detection to create ad-hoc cleaning rules. These rules transform inconsistent gender labels to 'male' or 'female', and skip any other values (i.e. the entire vector is not included in the cleaned data). The shoe size and height were normalised using min-max normalisation. I then split the data into 2/3 training data and 1/3 testing data to be able to test the accuracy of the algorithm. In order to find the value of k giving the highest prediction accuracy, I tried all possible values {1..number of training data points} as this data set is quite small.



Results: The line chart (see right) shows that k = {2..9} gives 95.5% accuracy. I suspect that the 81.8% accuracy as the value of k increases may correlate to the ratio of male to female respondents (since there are only two class labels).

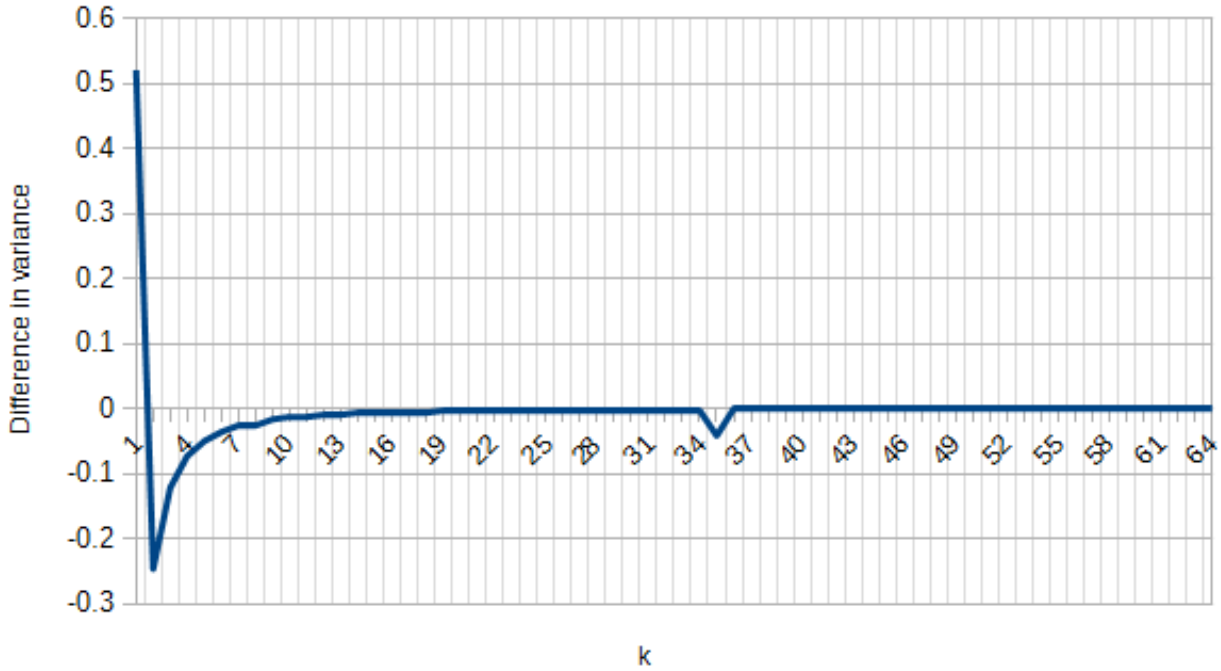
Clustering – k-means



Question: What are the clusters of age, shoe size, and height?

Each numerical attribute was normalised using min-max normalisation. Again, I tried k = {1..number of data points} to try to discover a good value for k.

Results: The line chart (see left) shows the variance of data points in the clusters as k increases. The second line chart (see next page) shows the difference in the variance as k increases. There begins to be a negligible difference in cluster quality around k = 7, suggesting that there are approximately 7 quality clusters of age, shoe size, and height.


Difference in variance
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
1
4
7
10
13
16
19
22
25
28
31
34
37
40
43
46
49
52
55
58
61
64
k