Data Mining and Machine Learning

F21DL: Coursework 1

Andrew Beveridge
H00013703

How I Did It and What Tools I Used

I wrote one PHP script, running on my webserver. I read the data directly into arrays using loops and regular expressions, built-in functions like "getcsv", and performed z-normalization using your awk script from within PHP. For my histograms I used the pChart library. URLs below may be of interest.

PHP Source Code: http://j.mp/dmml1-source

Histograms generator: http://j.mp/dmml1-hist

1-Nearest-Neighbour Classification

Data Set	Correct	Accuracy	Run Time	Link
Communities	1688 / 1994	84.65396 %	472 sec	→
- Z-Normalized	1692 / 1994	84.85456 %	517 sec	→
Pima	522 / 768	67.96875 %	6.7 sec	→
- Max-Min Normalized	542 / 768	70.57291 %	5.8 sec	→
Yeast	1036 / 1484	69.81132 %	24 sec	→
- Max-Min Normalized	1045 / 1484	70.41778 %	21 sec	→

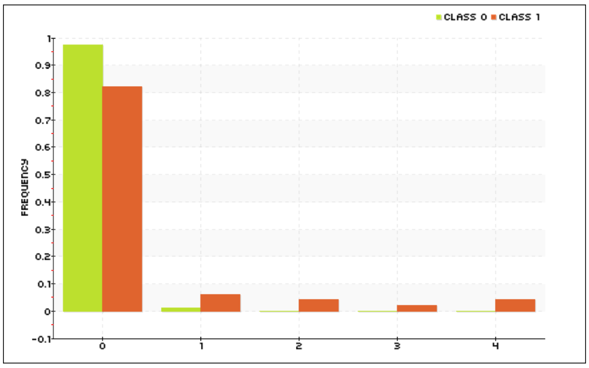
The performance of 1-Nearest-Neighbour Classification on the original Communities data set is very similar to the performance on the z-normalized version, because the original data is normalized already.

For the most part the same can be said about the Yeast data set - the original data is not completely normalized, but as all the values are within a similar range between 0 and 1, nearest neighbour does a pretty good job already without additional normalization, hence there only being a 0.6% difference in performance.

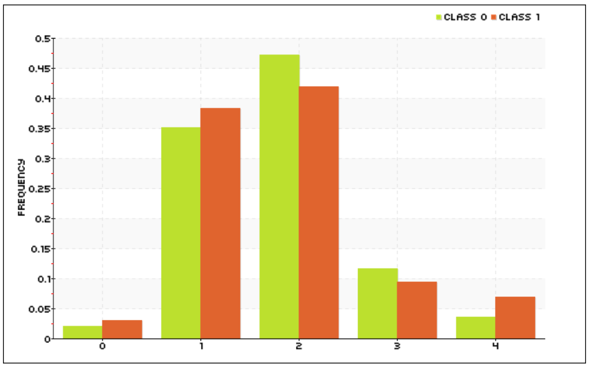
The larger gap in performance between original and normalized data from the Pima data set is caused by the wide variance between fields in the original data - one field having a value in the hundreds could cause the distance formula to be biased towards that field, giving an unhelpful nearest neighbour for classification. As such, performing simple min-max normalization on the data allows better classification.

Communities and Crime Data Set Histograms

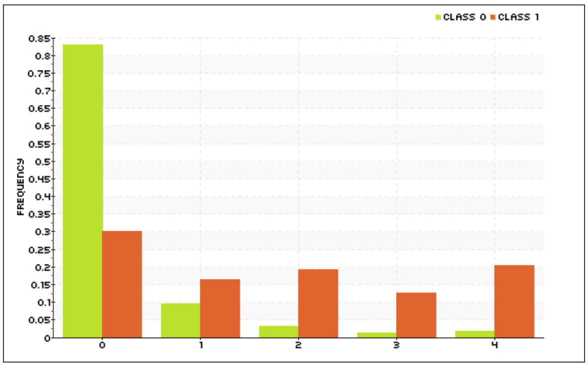
Field 1:



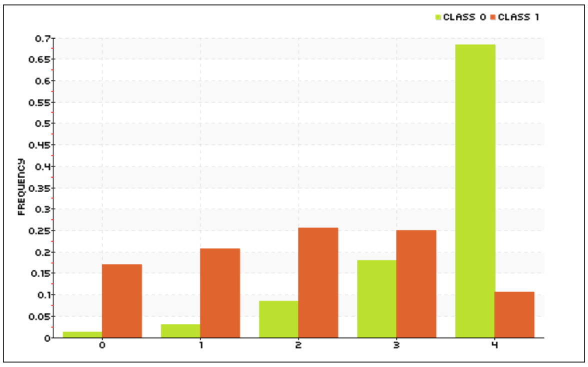
Field 2:



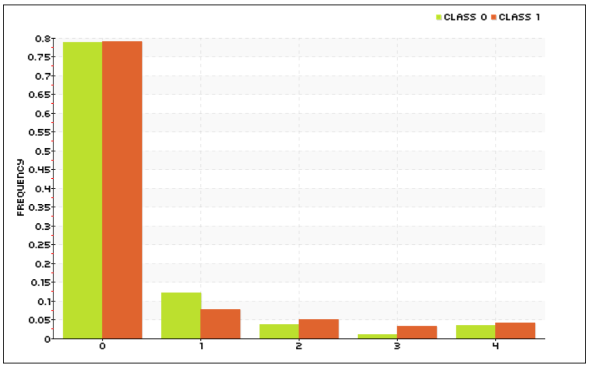
Field 3:



Field 4:



Field 5:



Notes

Most of the data in field 1 is in the first bucket, and there isn't a large difference between the two classes in that bucket. However, as class 0 seems to have no values at all in bucket 3, 4, or 5, yet class 1 does, it could still be useful in classification.

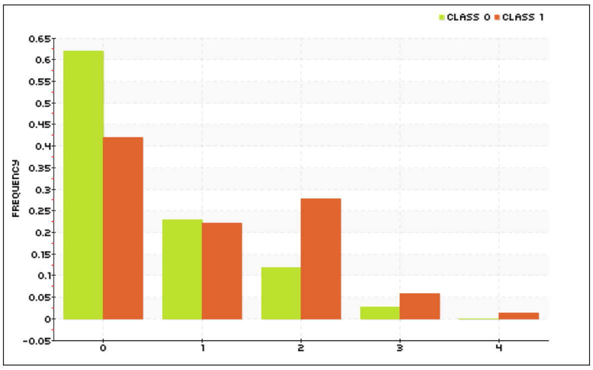
Field 2 has a fairly normal distribution for both classes, and is not very helpful.

Field 3 and 4 have fairly even distribution for class 1, and wide variance in values for class 0 - as such I feel these two fields are the most important for classification.

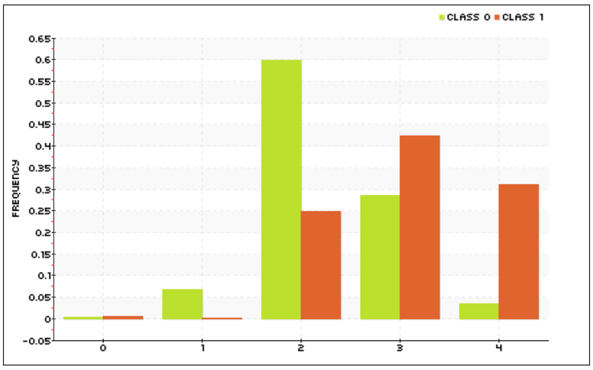
Field 5 has most data in bucket 1, with little class variance, so is not helpful.

Pima Indians Diabetes Data Set Histograms

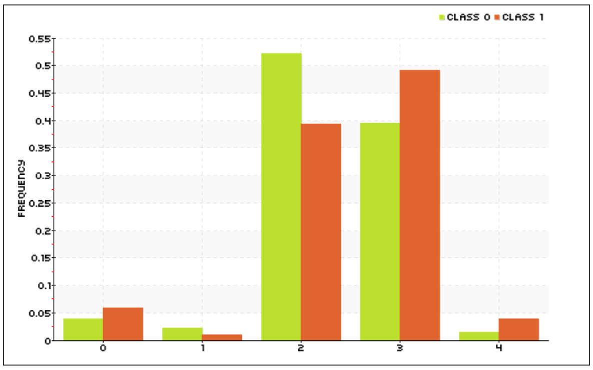
Field 1:



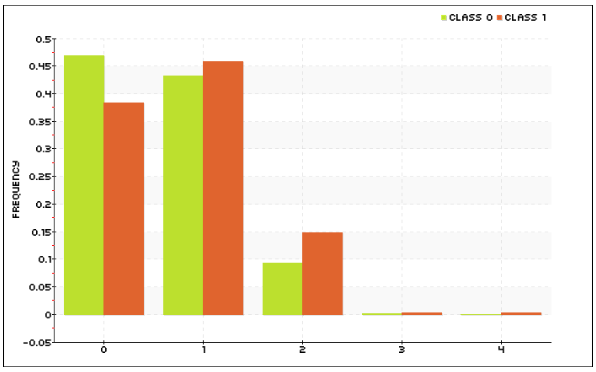
Field 2:



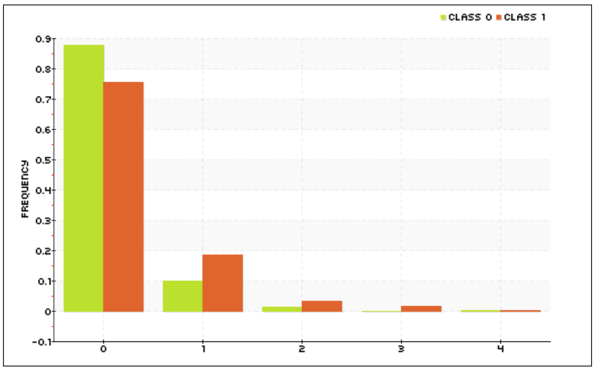
Field 3:



Field 4:



Field 5:

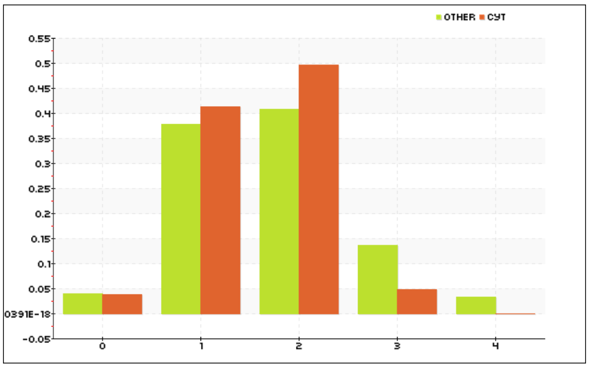


Notes

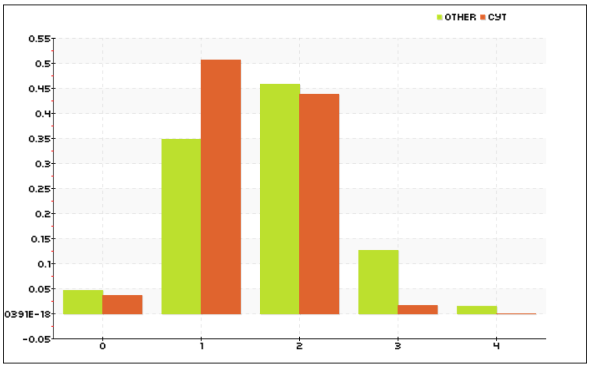
Field 1 has a higher frequency of class 0 in bucket 1 then class 1, and more than double the frequency of class 1 than class 0 in bucket 3. Bucket 4 has no class 0 data at all, and some class 1, making field 1 quite useful in classification.
Field 2 has a much larger frequency of class 0 in bucket 3, and almost 10 times the frequency of class 1 than 0 in bucket 5, with notable differences in bucket 2 and 4 also, therefore field 2 is also very useful in classification.
Field 3 has more class 0 in bucket 3 and more class 1 in bucket 4, so is somewhat useful but not substantially. Fields 4 and 5 have fairly similar frequencies for both classes, so are not particularly useful.

Yeast Data Set Histograms

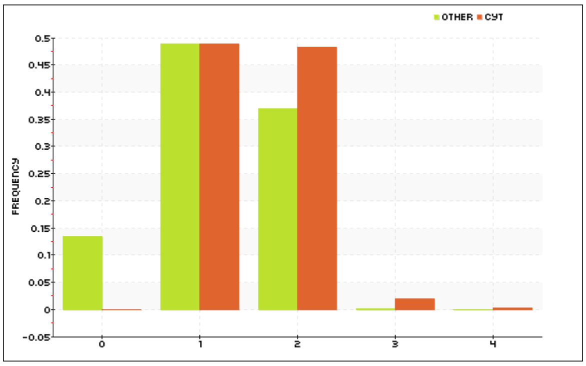
Field 1:



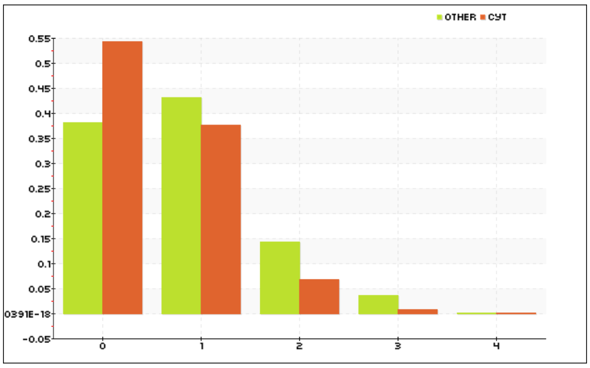
Field 2:



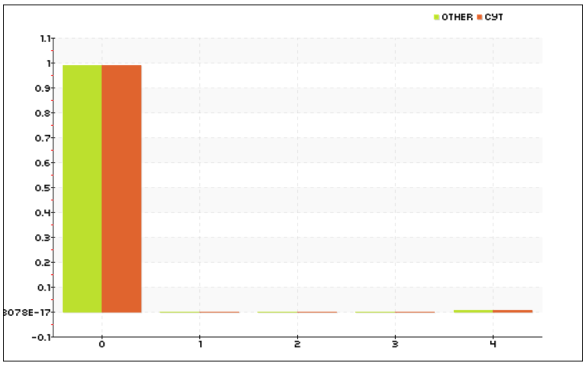
Field 3:



Field 4:



Field 5:



Notes

Field 1 has most data in buckets 2 and 3, with small differences in frequency; bucket 3 has a little more CYT, bucket 4 has about twice as much OTHER as CYT and bucket 5 has no CYT at all and some OTHER - as such, field 1 may be somewhat helpful in classification.
Field 2 has a big difference in bucket 4, which has far more OTHER, and bucket 2 which has about 1/3rd more CYT.
Field 3 has entirely OTHER in bucket 1, and entirely CYT in bucket 4, making it useful in classification.
Field 4 has more CYT in bucket 1 but is mostly fairly similar between classes.
Field 5 is identical for both classes, therefore useless.
I would suggest **Field 2 and Field 3** as the most useful fields for classification.

1-Nearest-Neighbour With Reduced Data Sets

Data Set	Correct	Accuracy	Run Time	Link
Communities	1550 / 1994	77.73319 %	14.7 sec	→
- Z-Normalized	1547 / 1994	77.58274 %	14.9 sec	→
Pima	498 / 768	64.84375 %	1.9 sec	→
- Max-Min Normalized	491 / 768	63.93229 %	1.8 sec	→
Yeast	943 / 1484	63.54447 %	7.5 sec	→
- Max-Min Normalized	945 / 1484	63.67924 %	6.6 sec	→