CS5920 Machine Learning – Assignment 1

Author: Michael Harrison

Student No: 100885079

KNN Implementation Details

- I have implemented KNN for general K, rather than just 1 & 3.
- If K exceeds the number of training observations, it will simply take the entire training set as each test point's "nearest" neighbours.
- Algorithm can run with a choice of either Euclidean or Tangent distance, although it uses Euclidean by default.
- Observations that are tied as being nearest to the test point are all included in the set of nearest neighbours – unless including all such cases would result in more than K neighbours. In which case, the ties are broken randomly.

  E.g: K=3 and we have 2 joint closest observations and 2 joint second closest observations – the first 2 joint closest would be included as nearest neighbours, but the 3rd nearest neighbour would be chosen at random from the 2 joint second closest.
- Ties in the most common label among the set of nearest neighbours are also broken randomly.
- Tangent distance was implemented using the provided distance.c file.

Iris Data

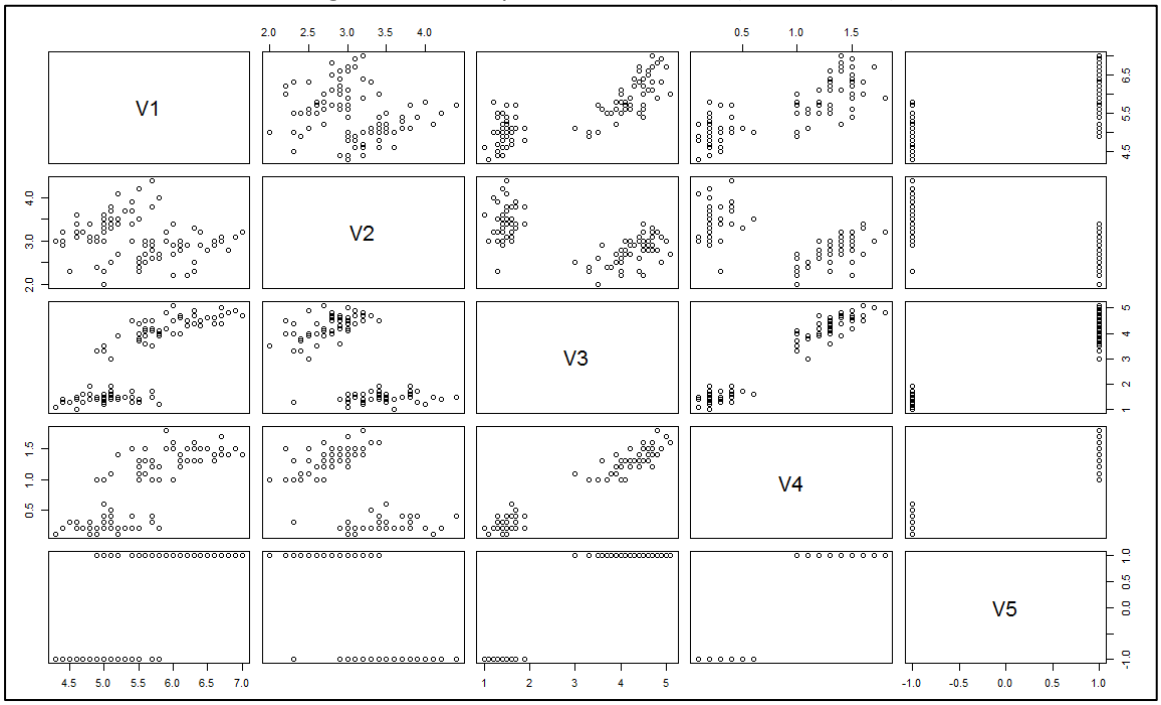
I used the first 70 rows as training data, and the final 30 rows as the test set. Running KNN with Euclidean distance, for K=1 & 3 respectively gave the following results:

K	Total Predictions	Correct Predictions	Wrong Predictions	Error Rate	% Correct
1	30	30	0	0.00%	100%
3	30	30	0	0.00%	100%

That is, KNN predicted the Iris type (Versicolor/Setosa) perfectly for both values of K, across all 30 test points.

This may seem surprising, however looking at the scatterplot pairs of the full Iris data set shows that both the 3rd & 4th variable (petal length & petal width) partition the 5th variable (Iris type) exactly – i.e. with no overlaps.

Figure 1: Scatterplot Paris, Iris Data (All 100 rows)



The other variables only show moderate overlap among the two Iris types. The separation of V5 vs V3 or V4 means that any test point will lie in one of the two clusters, and so its nearest neighbours (for small/moderate K) will also lie in that same cluster – hence the predicted label will be correct.

Note that the ranges of all attributes V1-V4 are broadly similar, so the Euclidean distance isn't distorted by use of different measurement scales.

Ionosphere Data

Results for Ionosphere data, using fist 200 rows as training data and remaining 151 as test data:

K	Total Predictions	Correct Predictions	Wrong Predictions	Error Rate	% Correct
1	151	139	12	7.9%	92.1%
3	151	141	10	6.6%	93.4%

The error rate is fairly low, suggesting a reasonable amount of clustering among the observations. The error rate improves for K=3 vs K=1, suggesting either some isolated points with the opposite label within these clusters acting as the single nearest neighbour for some test points.

The remaining error may be due to points at or near the boundaries of any clusters – i.e. the clusters are unlikely to be entirely separable. This is likely exacerbated by the curse of high dimensionality, since each observation lies in 34-dimensional space, which is large for such a relatively small data set (351 rows).

USPS Data

I have analysed the USPS subset data, partitioning equally (strictly, 233 & 232 rows respectively) into training & test sets at random, with set.seed(5).

The results below show that using tangent distance notably improved the error rate, as might be expected given the unsuitability of the Euclidean metric for the handwriting task.

Distance	K	Total Predictions	Correct Predictions	Wrong Predictions	Error Rate	% Correct
Euclidean	1	232	203	29	12.5%	87.5%
Euclidean	2	232	198	34	14.7%	85.3%
Euclidean	3	232	193	39	16.8%	83.2%
Tangent	1	232	215	17	7.3%	92.7%
Tangent	2	232	210	22	9.5%	90.5%
Tangent	3	232	210	22	9.5%	90.5%

After normalising the USPS data (min-max normalisation, applied rowwise) the error rates under Euclidean distance improve slightly.

Distance	K	Total Predictions	Correct Predictions	Wrong Predictions	Error Rate	% Correct
Euclidean	1	232	205	27	11.6%	88.4%
Euclidean	2	232	202	30	12.9%	87.1%
Euclidean	3	232	200	32	13.8%	86.2%

Note these results were produced on the same sample set as for those above, so are comparable.