## Project: 2 CS 170. Introduction to Artificial Intelligence

Due Date: Thursday of finals week at 11:59pm

## **Feature Selection with Nearest Neighbor**

Important: I am leaving the country for two weeks shortly after this deadline. I will be offline the whole time. Therefore, I will not offer any extensions. If you are late, you will have to take a low/failing grade, or ask for an incomplete (which is a huge pain for both of us).

As we have seen in class this quarter, the nearest neighbor algorithm is a very simple, yet very competitive classification algorithm. It does have one major drawback however; it is very sensitive to irrelevant features. With this in mind you will code up the nearest neighbor classifier, and then use it inside a "wrapper" which does two kinds of searches (listed below)

- 1) Forward Selection
- 2) Backward Elimination

Don't be scared by the phrase "search algorithm", in this case it is really simply nested loops, nothing else.

To make life simple, you can assume the following. I will only give you datasets that have two classes. I will only give you datasets that have continuous features.

Think carefully before you start coding this. Students in the past seem to have made this more complicated than it need be. In particular, in Matlab I was able to write the nearest neighbor algorithm in 8 lines of code, and the two search algorithms in another 17 lines of code.

C++ and Java programs tend to be longer, but even so, I would be surprised if this took more than 100 lines of code (although I will not penalize you for this).

**Very important**: Make sure your nearest neighbor algorithm is working correctly before you attempt the search algorithms. We will provide some test datasets for this purpose.

You may use some predefined utility routines, for example sorting routines. However I expect all the major code to be original. You must document any book, webpage, person or other resources you consult in doing this project (see the first day's handout).

You may consult colleagues at a high level, discussing ways to implement the tree data structure for example. But you may **not** share code. At most, you might illustrate something to a colleague with pseudocode.

You will hand in a report, see Project\_2\_sample\_report.pdf

You must keep the evolving versions of your code, so that, if necessary you can demonstrate to the course staff how you went about solving this problem (in other words, we may ask you to prove that you did the work, rather than copy it from somewhere).

You can use a simple text line interface or a more sophisticated GUI (but don't waste time making it pretty unless you are sure it works and you have lots free time). However, your program should have a trace like the one below, so that it can be tested.

The data files will be in the following format. ASCII Text, IEEE standard for 8 place floating numbers. This is a common format; you should be able to find some code to load the data into your program, rather that writing it from scratch (as always, document borrowed code). The first column is the class, these values will always be either "1"s or "2"s. The other columns contain the features, which are **not** normalized. There may be an arbitrary number of features (for simplicity I will cap the maximum number at 64). There may an arbitrary number of instances (rows), for simplicity I will cap the maximum number at 2,048. Below is a trivial sample dataset. The first record is class "2", the second is class "1" etc. This example has just two features.

```
2.0000000e+000 1.2340000e+010 2.3440000e+000
1.0000000e+000 6.0668000e+000 5.0770000e+000
2.0000000e+000 2.3400000e+010 3.6460000e+000
1.0000000e+000 4.5645400e+010 3.0045000e+000
```

Welcome to Bertie <u>Woosters</u> Feature Selection Algorithm.

Type in the name of the file to test: **eamonns\_test\_2.txt** 

Type the number of the algorithm you want to run.

- 1) Forward Selection
- 2) Backward Elimination

1

This dataset has 4 features (not including the class attribute), with 345 instances.

Running nearest neighbor with all 4 features, using "leaving-one-out" evaluation, I get an accuracy of 75.4%

Beginning search.

```
Using feature(s) {1} accuracy is 45.4% Using feature(s) {2} accuracy is 63.7% Using feature(s) {3} accuracy is 71.4% Using feature(s) {4} accuracy is 48.1%
```

Feature set {3} was best, accuracy is 71.4%

```
Using feature(s) {1,3} accuracy is 48.9% Using feature(s) {2,3} accuracy is 70.4% Using feature(s) {4,3} accuracy is 78.1%
```

Feature set {4,3} was best, accuracy is 78.1%

```
Using feature(s) {1,4,3} accuracy is 56.9% Using feature(s) {2,4,3} accuracy is 73.4%
```

(Warning, Accuracy has decreased! Continuing search in case of local maxima) Feature set {2,4,3} was best, accuracy is 73.4%

```
Using feature(s) {1,2,4,3} accuracy is 75.4%
```

Finished search!! The best feature subset is {4,3}, which has an accuracy of 78.1%

| SID | LAST NAME          | FIRST NAME   | Small Large     | 66  |                                    | 7814 | M  | Ar                               | 32  | 90  |
|-----|--------------------|--|-----------------|-----|------------------------------------|------|----|----------------------------------|-----|-----|
|     | 5249 AI            | Cł   | 81 118          | 67  |                                    | 7689 | M  | In                               | 99  | 32  |
|     | 0030 Ar            | CF   | 105 81          | 68  |                                    | 4812 | M  | Ra                               | 64  | 99  |
|     | 7833 Ar            | Lu   | 53 105          | 69  |                                    | 9022 | Na | Ja                               | 109 | 64  |
|     | 7884 Ar            | Li   | 79 53           | 70  |                                    | 7874 | Ng | Ra                               | 5   | 109 |
|     | 0805 Ba            | H:   | 94 79           | 71  |                                    | 6909 | No | AI                               | 16  | 5   |
|     | 0322 Ba            | Cr   | 104 94          | 72  |                                    | 6910 | No | Za                               | 10  | 16  |
| -   | 5321 BH            | Ka<br>Al   | 48 104          | 73  |                                    | 2778 | Od | TH                               | 106 | 10  |
|     | 6336 Bd<br>4471 Bd | Ju   | 38 48<br>11 38  | 74  |                                    | 4043 | Pa | Zi                               | 33  | 106 |
|     | 3453 Bd            | Ar   | 39 11           | 75  |                                    | 7380 | Pa | Br                               | 61  | 33  |
|     | 1115 Bu            | Fr   | 67 39           | 76  |                                    | 1854 | Pa | Ju                               | 78  | 61  |
|     | 1761 Ca            | Jo   | 24 67           | 77  |                                    | 3965 | PH | Hi                               | 46  | 78  |
|     | 1945 CH            | Ja   | 74 24           | 78  |                                    | 8256 | Ph | Ar                               | 4   | 46  |
|     | 2333 CH            | Ni   | 103 74          | 79  |                                    | 2843 | Ph | Et                               | 108 | 4   |
|     | 0396 CH            | Br   | 91 103          | 80  |                                    | 7794 | Pc | Ez                               | 8   | 108 |
|     | 5214 Ch            | Cr   | 23 91           | 81  |                                    | 1940 | Pr | Kε                               | 122 | 8   |
|     | 6340 CH            | Ar   | 28 23           | 82  |                                    | 5850 | Ra | Cr                               | 113 | 122 |
|     | 3516 Cc            | Al   | 97 28           | 83  |                                    | 9641 | Ra | Ac<br>Al                         | 65  | 113 |
|     | 8465 Cd            | Di   | 36 97           | 84  |                                    | 7355 | Ra | AI                               | 56  | 65  |
|     | 8616 Da            | Za   | 70 36           | 85  | 1                                  | 7957 |    | Di                               | 85  | 56  |
|     | 5551 De            | Ja<br>Er   | 116 70          | 86  |                                    | 2100 |    | N:                               | 100 | 85  |
|     | 8565 De<br>2286 De | Cē   | 15 116<br>93 15 | 87  |                                    | 3417 |    | Zε                               | 114 | 100 |
|     | 4795 De            | M  | 57 93           | 88  |                                    | 8714 |    | Br                               | 27  | 114 |
|     | 3921 Do            | H  | 77 57           | 89  |                                    | 5098 |    | Sk                               | 110 | 27  |
|     | 1078 Fu            | M  | 73 77           | 90  |                                    | 3625 |    | Is                               | 121 | 110 |
|     | 6117 Ga            | Ar   | 49 73           | 91  |                                    | 7096 |    | Ni                               | 84  | 121 |
|     | 5609 Ga            | Jo   | 112 49          | 92  |                                    | 7285 |    | Br                               | 75  | 84  |
|     | 6679 Gi            | Ju   | 2 112           | 93  |                                    | 9033 |    | Al                               | 45  | 75  |
|     | 2487 G             | D:   | 82 <b>2</b>     | 94  |                                    | 8895 |    | Ra                               | 102 | 45  |
|     | 9874 Ha            | Ki   | 87 82           | 95  |                                    | 8900 |    | Di                               | 66  | 102 |
|     | 4078 Ha            | Dy   | 76 87           | 96  |                                    | 6109 |    | H                                | 50  | 66  |
|     | 6392 H€            | Ju   | 72 76           | 97  |                                    | 0103 |    | Cc                               | 86  | 50  |
|     | 1318 Hd            | Ci   | 1 72            | 98  |                                    | 8533 |    | M                                | 68  | 86  |
|     | 1603 Hs            | Al<br>Jir  | 98 1<br>117 98  | 99  |                                    | 3701 |    | Sa                               | 123 | 68  |
|     | 0164 Hu<br>2398 Hu | Aā   | 117 98          | 100 |                                    | 3492 |    | Ja                               | 18  | 123 |
|     | 7512 Hu            | Ac   | 69 119          | 101 |                                    | 5893 |    | Al                               | 115 | 18  |
|     | 6343 Ici           | Ke   | 89 69           | 102 |                                    | 5936 |    | Je                               | 3   | 115 |
|     | 5931 In            | Ju   | 83 89           | 103 |                                    | 9395 | 1  | Ri                               | 19  | 3   |
|     | 3588 Is            | Sh<br>M  | 62 83           | 104 |                                    | 7455 |    | Va                               | 7   | 19  |
|     | 2815 Ja            | М  | 40 62           | 105 |                                    | 9565 |    | lu                               | 63  | 7   |
|     | 2375 Jo            | M  | 54 40           | 106 |                                    | 5356 |    | Ju<br>Ju<br>Ri                   | 34  | 63  |
|     | 6673 Ka            | Er<br>Ra   | 25 54           | 107 |                                    | 2585 |    | Ri                               | 9   | 34  |
|     | 6702 Ka            | Ra   | 71 25           | 108 |                                    | 7364 |    | M                                | 37  | 9   |
|     | 8343 Ki            | М  | 52 71           | 109 |                                    | 7045 |    | Ri                               | 80  | 37  |
|     | 3293 Kd<br>4132 Ku | N:   | 43 52<br>29 43  | 110 |                                    | 6532 |    | CH                               | 47  | 80  |
|     | 3377 Le            | Jo<br>Hi   | 30 29           | 111 |                                    | 6246 |    | Cr<br>At<br>Zr<br>Cr<br>Dr<br>Re | 58  | 47  |
|     | 9014 Le            | G  | 20 30           | 112 |                                    | 7611 |    | ZI                               | 13  | 58  |
|     | 2209 Li:           | GI<br>Ke   | 35 20           | 113 |                                    | 6463 |    | Cr                               | 41  | 13  |
|     | 2222 Lis           | Qi   | 120 35          | 114 |                                    | 8683 |    | Di                               | 59  | 41  |
|     | 5458 Lii           | Br   | 22 120          | 115 |                                    | 5068 |    | Re                               | 124 | 59  |
|     | 6646 Li            | Br<br>Az<br>Di<br>Ja<br>Vi<br>Ar<br>Jo<br>Di<br>Tr | 95 22           | 116 |                                    | 7627 |    | Te                               | 107 | 124 |
|     | 7832 Li            | Di   | 17 95           | 117 |                                    | 9826 |    | Fd                               | 31  | 107 |
|     | 1776 Lo            | Ja   | 101 17          | 118 |                                    | 3679 |    | Ec<br>Je<br>Ze                   | 125 | 31  |
|     | 3384 M             | Vi   | 14 101          | 119 |                                    | 4632 |    | 76                               | 92  | 125 |
|     | 2182 M             | Ar   | 111 14          | 120 |                                    | 7553 |    | M                                | 60  | 92  |
|     | 5676 M             | Jo   | 55 111          | 121 |                                    | 2938 |    | Ra                               | 26  | 60  |
|     | 9627 M             | Di   | 51 55           | 122 |                                    | 7767 |    | Rame                             | 21  | 26  |
|     | 6756 M             |  | 44 51           |     |                                    |      | 20 |                                  | 96  | 21  |
|     | 1538 M             | U:<br>Ar<br>Ar                                     | 42 44<br>12 42  |     | 123 Sample data<br>124 Sample data |      |    | Sue                              | 6   | 96  |
|     | 6780 M<br>7065 M   | Δι   | 90 12           |     |                                    |      |    | Joe                              |     |     |
|     | IVI C007           | A  | 50 12           | 125 | Sample                             | uala |    | Van                              | 88  | 6   |