Decision Making Final

Andrew Cash

2023-12-10

1. Using the data in wine.csv, use one of the supervised classification algorithms we learned (logistic regression or KNN) to determine the class of the wine (white or red) based on the variables given. Use an 80/20 split to test your model.

The guessed types of wine using KNN:

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 1 2 1 2
```

The correct types:

```
##  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2
```

The percent of correctly guessed wine types:

```
## [1] 92.30769
```

2. Use the data in wine.csv again, this time ignore the class and run a clustering algorithm to see if you can find the 2 clusters, Does the clustering algorithm think 2 is the correct number of clusters?

Note that indexes 1-95 are actually white wine, and 96-128 are red wine

The 1st cluster produced by k-means clustering:

```
##  [1]  23  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 113 114
## [20] 115 117 118 119 120 121 122 123 124 125 126 127 128
```

The 2nd cluster produced by k-means clustering:

```
##  [1]   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
## [20]  20  21  22  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39
## [39]  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58
## [58]  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77
## [77]  78  79  80  81  82  83  84  85  86  87  88  89  90  91  92  93  94  95 112
## [96] 116
```

The actual indexes of the types of wines:

white:

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## [26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## [51] 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
## [76] 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95
```

red:

```
##  [1]  96  97  98  99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
## [20] 115 116 117 118 119 120 121 122 123 124 125 126 127 128
```

Or more useful,

The percentage of white wine in cluster 1 vs cluster 2:

```
## [1] 0.01052632
```

```
## [1] 0.9894737
```

The percentage of red wine in cluster 1 vs cluster 2:

```
## [1] 0.9393939
```

```
## [1] 0.06060606
```

The main take-away is that it is inconsistent, which is in-line for an unsupervised model. The red wine data almost always stays together, no matter which cluster it ends up in. From my observations, one of two things tends to happen: either one cluster is mostly red wine and the other is mostly white wine, or one cluster has mostly red wine but also about half of the white wine, and the other cluster has the other half of white wine. This also seems to be the case for R's built in k-means test. In the former situation, yes, two clusters is absolutely correct, that is what you would want from this algorithm. For the later situation it's trickier, but I would still say yes due to the nature of k-means clustering.

3. A mouse is placed in a 4×4 square grid. The mouse can vertically and horizontally through the grid, but can't leave the grid. The mouse moves with equal probably to any adjacent cell (the mouse has to move). The mouse is placed into location (1,1). In location (4,4) is a large piece of cheese. In cell's (3,2) and (2,4) is a trap in which the mouse will get stuck in that cell.

The steady state for each of the 16 rooms. The only non-zeros are the absorbing states.

```
##  [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [7] 0.00000000 0.31419514 0.00000000 0.63085594 0.00000000 0.00000000
## [13] 0.00000000 0.00000000 0.00000000 0.05494893
```

(a) What is the probability the mouse makes it to the the Cheese.

```
## [1] 0.05494893
```

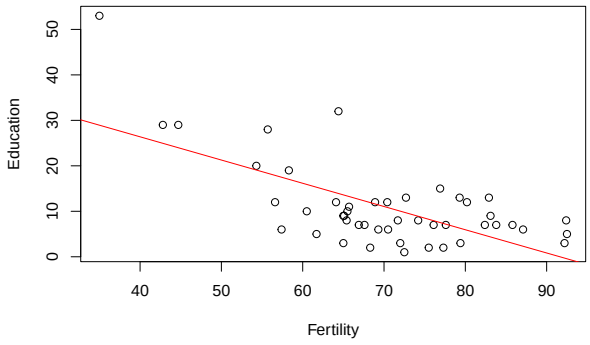
(b) What is the expected number of steps until the mouse is either trapped or gets to the cheese?

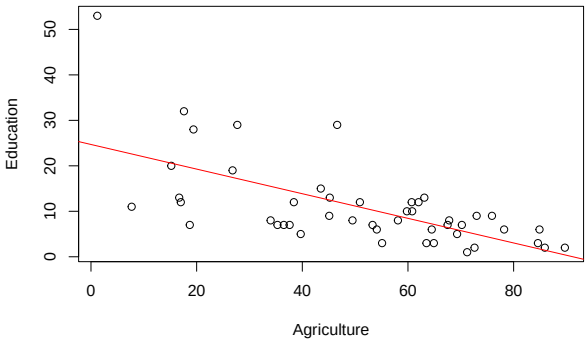
```
## [1] 9.623811
```

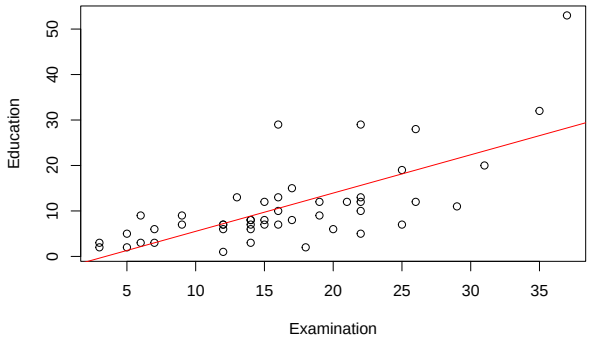
4. Using the swiss data set built into r,

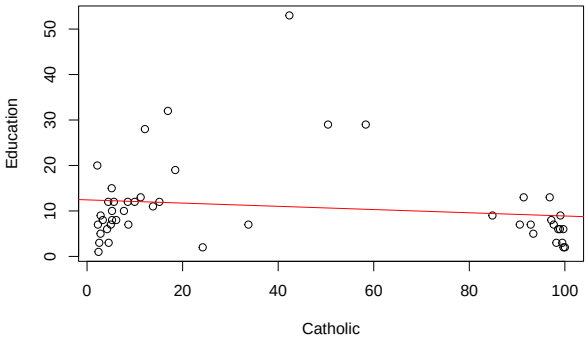
(a) Determine which of the following (Fertility, Agriculture, Examination, Catholic, Infant.Mortality) are significant when estimating Education with a linear model.

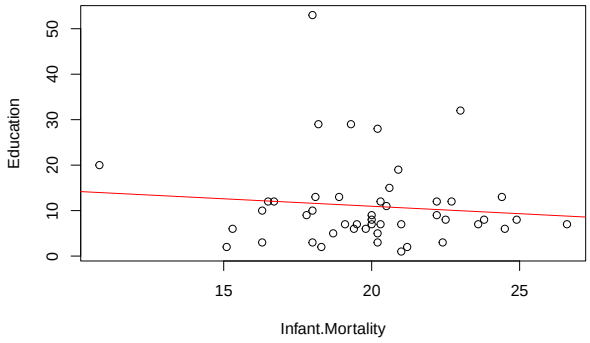
Plot of Education vs Fertility, Agriculture, Examination, Catholic, and Infant.Mortality respectively:











The p-values of Fertility, Agriculture, Examination, Catholic, and Infant.Mortality respectively within the linear model:

```
## [1] 2.430605e-05
```

```
## [1] 0.0008038196
```

```
## [1] 0.01392183
```

```
## [1] 3.293778e-05
```

```
## [1] 0.4763052
```

Fertility, Agriculture, and Examination and Catholic are significant, as they are less than the alpha value (0.05).

(b) Using only the variables that are significant, construct a multiple regression model to estimate Education.

```
##
## Call:
## lm(formula = Education ~ Fertility + Agriculture + Examination +
##     Catholic, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.4844  -2.5083  -0.3969   2.8415  11.3295
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.94784    8.28480   4.218 0.000128 ***
## Fertility   -0.38246    0.07738  -4.943 1.28e-05 ***
## Agriculture -0.16689    0.04419  -3.777 0.000493 ***
## Examination  0.43273    0.16146   2.680 0.010466 *
## Catholic     0.10151    0.02130   4.766 2.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.879 on 42 degrees of freedom
## Multiple R-squared:  0.7649, Adjusted R-squared:  0.7425
## F-statistic: 34.17 on 4 and 42 DF,  p-value: 1.066e-12
```

(c) Estimate the education level for a draftee using only the significant variables from the full list of variables and the values below.

Fertility 60.00
Agriculture 55.00
Examination 20.00
Catholic 4.15
Infant.Mortality 15.20

Significant variables:

```
##   Fertility Agriculture Examination Catholic
## 1        60          55          20     4.15
```

Estimated education level:

```
##        1
## 11.89742
```

5. Let X1, X2, . . . be iid normal random variables with mean 100 and sd 15.

(a) What is P (85 < X1 < 115)?

```
## [1] 0.6826895
```

(b) If we take a random sample of 20 of these rv's, what is the probability that at least 15 of these random variables will fall between 85 and 115?

```
## [1] 0.3521636
```

(c) If we take one rv at a time and we let Y be the first time that a rv's value falls within [85, 115].

i. What is the distribution of Y ?

Y is a geometric distribution as it is performed until the 1st success (a value within [85, 115]), the trials are independent, the result is either success or failure, and each trial has the same probability of success.

ii. What is the mean number of rv's needed to get a value in [85,115]?

The mean of a geometric distribution is given by 1/p, p being the probability of success:

```
## [1] 1.464795
```