CMDA 3654: Assignment #4

Due on Friday, Oct 21, 2016

Jung Choi

"I have neither given nor received unauthorized assistance on this assignment."

October 21, 2016

1 CLASSIFICATION.

(a) Load the **wine** dataset from the **rattle** package in R, and describe the dataset in your own words, in 2-3 lines.

Consider **Type** to be the response variable, and all other variables as features.

The wine datset contains three types of wine in the 178 samples. The three types of wine are labeled in numeric 1 3 in column 1(response variable) and rest of the 13 columns(features) represent the wine results of chemical analyses. All data will be represented in numeric values.

(b) Perform classification using LDA (linear discriminant analysis) and report the classification error rate.

```
require(MASS) # library and require load and attach add-on packages.

install.packages("rattle")
library(rattle)

r1 = lda(formula = Type~., data = wine)
r1
r1$counts

r1.prop = r1$svd^2/sum(r1$svd^2) # same as proportion of trace
plda= predict(r1,wine)
table(wine$Type, plda$class)
mean(wine$Type!=plda$class, CV=TRUE)
plot(r1)


> r1
Call:
lda(Type ~ ., data = wine)

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629

Group means:
   Alcohol    Malic      Ash Alcalinity Magnesium  Phenols Flavanoids
       Nonflavanoids Proanthocyanins
1 13.74475 2.010678 2.455593   17.03729  106.3390 2.840169  2.9823729
    0.290000        1.899322
2 12.27873 1.932676 2.244789   20.23803   94.5493 2.258873  2.0808451
    0.363662        1.630282
3 13.15375 3.333750 2.437083   21.41667   99.3125 1.678750  0.7814583
    0.447500        1.153542
     Color       Hue Dilution   Proline
1 5.528305 1.0620339 3.157797 1115.7119
2 3.086620 1.0562817 2.785352  519.5070
3 7.396250 0.6827083 1.683542  629.8958

Coefficients of linear discriminants:
                         LD1           LD2
Alcohol         -0.403399781  0.8717930699
```

```
Malic             0.165254596  0.3053797325
Ash              -0.369075256  2.3458497486
Alcalinity        0.154797889 -0.1463807654
Magnesium        -0.002163496 -0.0004627565
Phenols           0.618052068 -0.0322128171
Flavanoids       -1.661191235 -0.4919980543
Nonflavanoids    -1.495818440 -1.6309537953
Proanthocyanins   0.134092628 -0.3070875776
Color             0.355055710  0.2532306865
Hue              -0.818036073 -1.5156344987
Dilution         -1.157559376  0.0511839665
Proline          -0.002691206  0.0028529846

Proportion of trace:
   LD1    LD2
0.6875 0.3125
> r1$counts
 1  2  3
59 71 48
>
> r1.prop = r1$svd^2/sum(r1$svd^2) # same as proportion of trace
> plda= predict(r1,wine)
> table(wine$Type, plda$class)

     1  2  3
  1 59  0  0
  2  0 71  0
  3  0  0 48
> mean(wine$Type!=plda$class, CV=TRUE)
[1] 0
```

(c) Perform classification using QDA (quadratic discriminant analysis) and report the classification error rate.

```
r2 = qda(formula =Type~. ,data = wine)
r2
pqda=predict(r2,wine)
table(wine$Type, pqda$class)
mean(wine$Type!=pqda$class, CV=TRUE)

> r2
Call:
qda(Type ~ ., data = wine)

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629

Group means:
   Alcohol    Malic      Ash Alcalinity Magnesium  Phenols Flavanoids
       Nonflavanoids Proanthocyanins
1 13.74475 2.010678 2.455593   17.03729  106.3390 2.840169  2.9823729
    0.290000        1.899322
2 12.27873 1.932676 2.244789   20.23803   94.5493 2.258873  2.0808451
    0.363662        1.630282
3 13.15375 3.333750 2.437083   21.41667   99.3125 1.678750  0.7814583
    0.447500        1.153542
     Color       Hue Dilution   Proline
1 5.528305 1.0620339 3.157797 1115.7119
2 3.086620 1.0562817 2.785352  519.5070
```

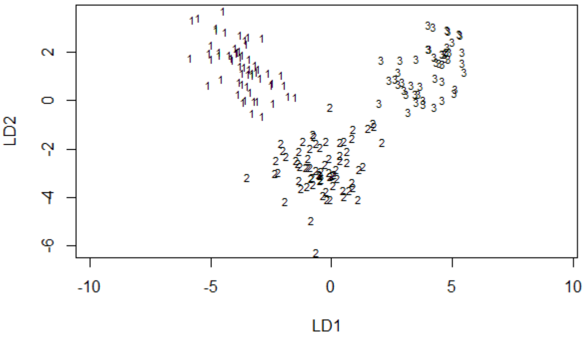
```
3 7.396250 0.6827083 1.683542  629.8958
> pqda=predict(r2,wine)
> table(wine$Type, pqda$class)

     1  2  3
  1 59  0  0
  2  1 70  0
  3  0  0 48
> mean(wine$Type!=pqda$class, CV=TRUE)
[1] 0.005617978
```

(d) Perform classification using SVM (support vector machines) and report the classification error rate.

```
> svm(Type~., data = data.wine)

Call:
svm(formula = Type ~ ., data = data.wine)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.07692308

Number of Support Vectors:  69

> r3

Call:
svm(formula = Type ~ ., data = data.wine)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.07692308

Number of Support Vectors:  69

> psvm = predict(r3, data = data.wine)
> psvm
  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  27
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
      1   1   1   1   1   1   1
 28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46  47
     48  49  50  51  52  53  54
  1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
      1   1   1   1   1   1   1
 55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
     75  76  77  78  79  80  81
  1   1   1   1   1   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
      2   2   2   2   2   2   2
 82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99 100 101
    102 103 104 105 106 107 108
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
      2   2   2   2   2   2   2
```

```
109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128
    129 130 131 132 133 134 135
  2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2   2
      2   2   3   3   3   3   3
136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155
    156 157 158 159 160 161 162
  3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
      3   3   3   3   3   3   3
163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178
  3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
Levels: 1 2 3
> mean(data.wine$Type!=psvm)
[1] 0
```

(e) Rank the classification methods in your order of preference for this dataset, and justify your preference.



All of the three methods make quiet excellent assumption for the classification error rates. As you can see on the graph fig1. All three types of wine are well separated and one small over lapping with type two and three. LDA would go straight line between the datasets while qda would sort of go around the datasets. Due to the data type two and three, I would personally rate qda as first then lda and svm.