# ETC3250 Lab 9

*Di Cook*

*SOLUTION*

## Purpose

This lab will fit a variety of classifiers (support vector machines, trees and forests) to two different data sets, and compare results.

## Data

- chocolates data used in the previous lab
- Bob Ross paintings

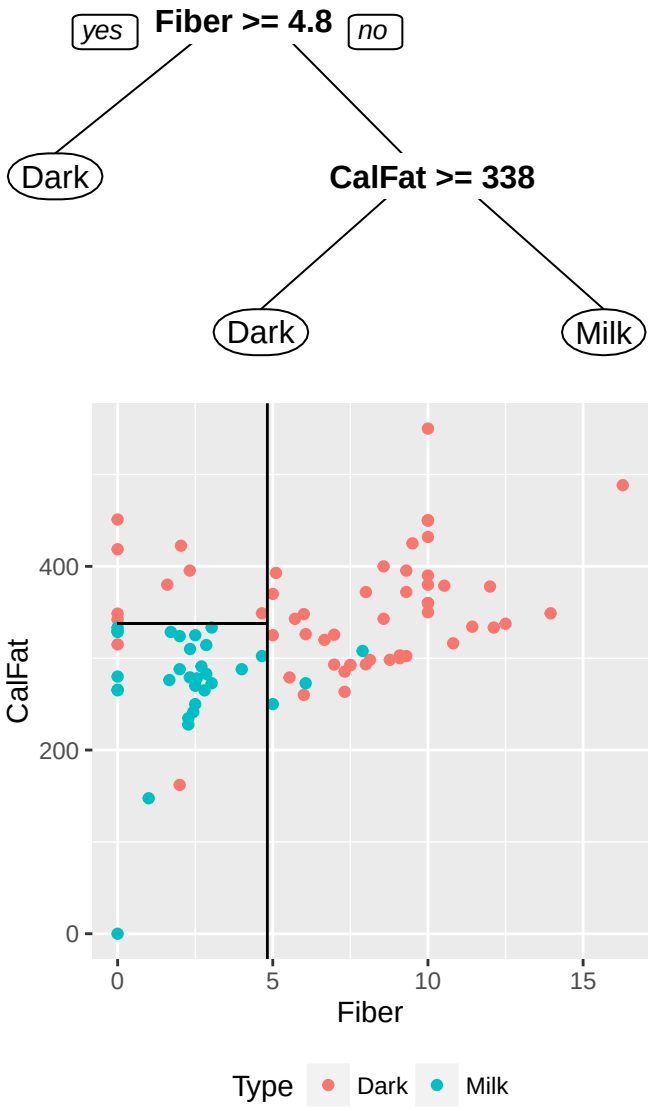
## Question 1

```
#      Calories CalFat TotFat SatFat  Chol   Na Carbs Fiber Sugars Protein
# [1,]   -0.085   0.61   0.55   0.39 -0.24 -0.7 0.097   1.3  -0.64   -0.41
# [1] -1.2
#
#        Dark Milk
#   Dark   53    2
#   Milk    4   28
#
#        Dark Milk
#   Dark   53    2
#   Milk    4   28
```

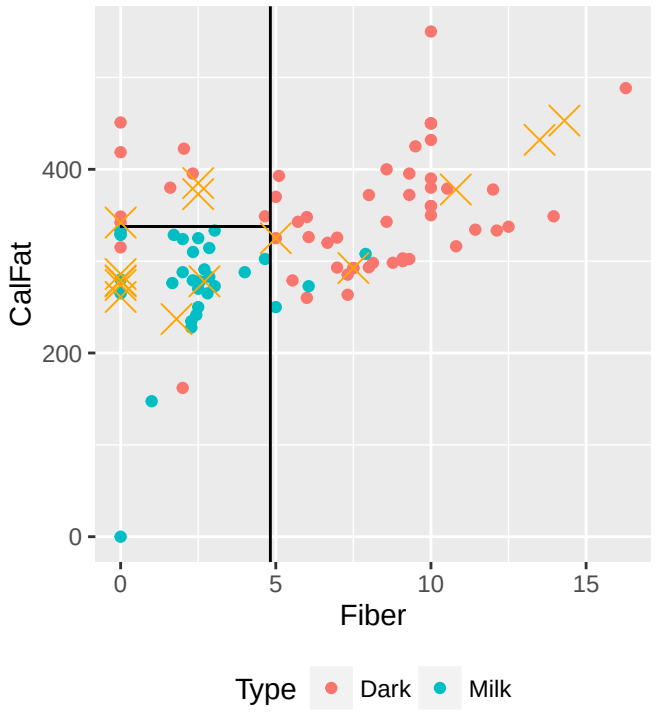
a. Read in the chocolates data, from the class web site.
b. Fit a linear kernel support vector machine. Report the equation of the separating hyperplane. `The coefficients are` -0.08, 0.61, 0.55, 0.39, -0.24, -0.7, 0.1, 1.25, -0.64, -0.41 `and the contant is` -1.19
c. Compute the error. `6/87=0.069`
d. Does the error get smaller if you use a different kernel?`Other kernels don't really improve predictions for this data. The error with the linear kernel is pretty small.`
e. Predict the new data.

```
#    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
# Dark Milk Milk Milk Dark Milk Dark Dark Dark Milk Dark Dark Dark Milk Milk
# Levels: Dark Milk
```

## Question 2




```
#
#        Dark Milk
#   Dark   53    2
#   Milk    3   29
```




```
#
#        Dark Milk
#   Dark   55    0
#   Milk    0   32
```

a. Fit a tree classifier to the data, using the default settings. Print the tree and write down the decision rule.`The rule is: If Fiber is greater than 4.8 assign new observation to Dark, otherwise if CalFat is greater than or equal to 338 assign to Dark, else assign to Milk.`

b. Compute the error.`5/87=0.057`

c. Make a plot that shows the boundary.

d. Plot (on the training data) and predict the new data.

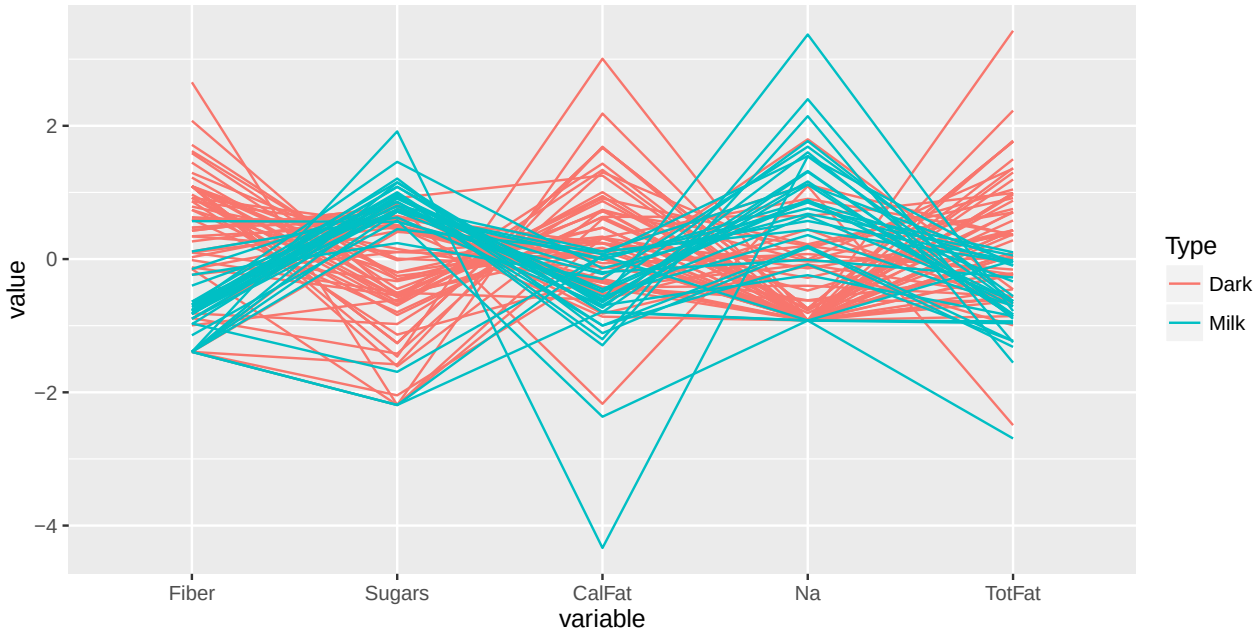
```
#    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
# Dark Dark Milk Milk Dark Dark Dark Milk Dark Milk Dark Milk Milk Milk Dark
# Levels: Dark Milk
```

e. Try adjusting the controls (e.e. minimum split), to get a lower error. `By adjusting the controls we can get the model to perfectly fit this data. Not necessarily a good idea, because the model may not work well with future data.`

## Question 3

```
#
# Call:
#  randomForest(formula = Type ~ ., data = choc_sub, importance = TRUE,      ntree = 500, mtry = 4)
#                Type of random forest: classification
#                      Number of trees: 500
# No. of variables tried at each split: 4
```

```
#
#         OOB estimate of  error rate: 10%
# Confusion matrix:
#      Dark Milk class.error
# Dark   51    4       0.073
# Milk    5   27       0.156
#          Var     Dark   Milk MeanDecreaseAccuracy MeanDecreaseGini
# 1      Fiber  0.06101 0.1548               0.0929             10.0
# 2     CalFat  0.03061 0.0842               0.0483              5.3
# 3     TotFat  0.03775 0.0594               0.0448              5.0
# 4     Sugars  0.03557 0.0436               0.0378              6.4
# 5         Na  0.02209 0.0613               0.0365              5.2
# 6       Chol  0.00711 0.0169               0.0107              2.2
# 7      Carbs  0.00457 0.0095               0.0063              1.7
# 8    Protein  0.00585 0.0035               0.0048              1.2
# 9     SatFat -0.00167 0.0148               0.0043              1.3
# 10 Calories  0.00018 0.0074               0.0028              1.6
```




a. Fit a random forest to the chocolates data.
b. Report the error.`About 13% depending on the sample of trees`
c. Use a parallel coordinate plot to display the data using the importance to order the variables.
d. Predict the new data.

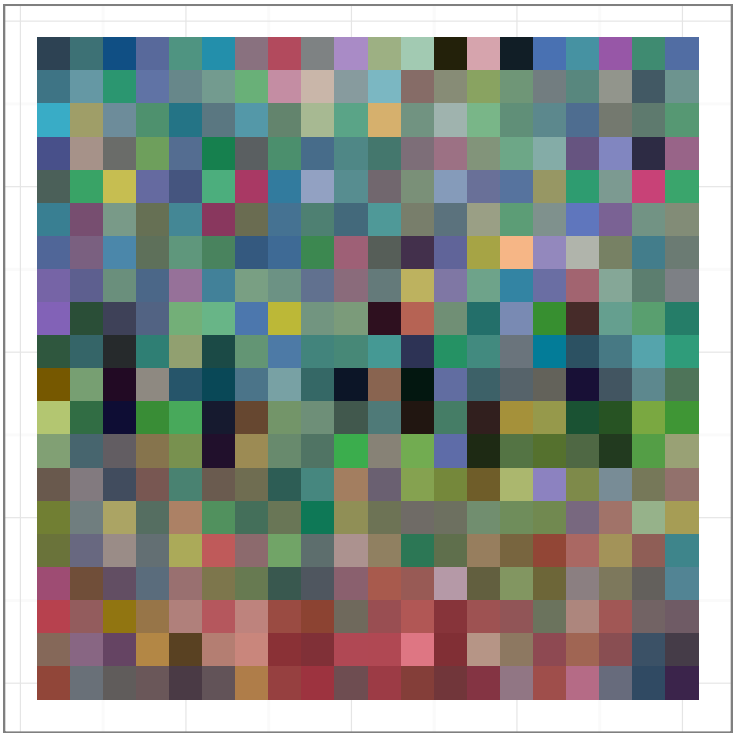
```
#    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
# Dark Dark Milk Milk Dark Dark Dark Milk Dark Milk Dark Milk Milk Milk Dark
# Levels: Dark Milk
```

## Question 4

a. Which of the new cases do the methods all agree on? On which ones is there disagreement?
b. Plot the cases where there is disagreement on the full data, in a parallel coordinate plot (as used in Q3).

## Question 5

```
#
# Call:
#  randomForest(formula = class ~ ., data = p_sub[, -c(1, 2)], ntree = 10000,      importance = TRUE)
#                Type of random forest: classification
#                      Number of trees: 10000
# No. of variables tried at each split: 34
#
#         OOB estimate of  error rate: 22%
# Confusion matrix:
#         cold flowers class.error
# cold      19       4        0.17
# flowers    6      16        0.27
#    Var   cold flowers MeanDecreaseAccuracy MeanDecreaseGini
# 1 b317 0.0059  0.0050               0.0053             0.36
# 2 b295 0.0039  0.0050               0.0042             0.32
# 3 g316 0.0044  0.0035               0.0038             0.25
# 4 g373 0.0039  0.0029               0.0033             0.20
# 5 b316 0.0034  0.0029               0.0030             0.22
# 6 b385 0.0035  0.0026               0.0030             0.20
#  [1] cold    cold    cold    cold    cold    cold    cold    cold
#  [9] cold    cold    cold    cold    cold    cold    cold    cold
# [17] cold    cold    cold    cold    cold    cold    cold    flowers
# [25] flowers flowers flowers flowers flowers flowers flowers flowers
# [33] flowers flowers flowers flowers flowers flowers flowers flowers
# [41] flowers flowers flowers flowers flowers
# Levels: cold flowers
#       1       2       3       4       5       6       7       8       9
#    cold    cold    cold    cold    cold    cold    cold    cold    cold
#      10      11      12      13      14      15      16      17      18
#    cold    cold    cold    cold    cold    cold    cold    cold    cold
#      19      20      21      22      23      24      25      26      27
#    cold flowers flowers flowers flowers    cold    cold flowers flowers
#      28      29      30      31      32      33      34      35      36
# flowers flowers flowers flowers flowers flowers flowers    cold    cold
#      37      38      39      40      41      42      43      44      45
# flowers flowers flowers    cold    cold flowers flowers flowers flowers
# Levels: cold flowers
# # A tibble: 1 x 5
#      id            name   class    r1    g1
#   <int>           <chr>  <fctr> <dbl> <dbl>
# 1   188 a-summers-place flowers  0.18  0.26
```



a. Explain the difference between the long and the wide format of the data.

`The wide form has the r, g, b values for each pixel in a painting in a column of the matrix. One row corresponds to one painting. We need this for fitting the classifier. The long form has the pixel location in the painting, r, g, b and hex value for each pixel in columns, with one row corresponding to a pixel in a painting.`

b. Subset the data to focus on two classes, `flowers` and `cold`.
c. Build a random forest for the training data.
d. Predict the class of test set, report the error.`The error is around 25%`
e. Which pixels are the most important for distinguishing these two types of paintings? `b317,b295,g373,... These should be pretty stable from one forest fit to another, if the model is a reliable classifier for future data. There are more variables than cases, which makes it possible that we are simply classifying noise.`
f. Plot one of the `flower` paintings that was misclassified as `cold`. Can you see any reasons why this might be? `The one I've chosen to plot is simply a hodge podge of color, could be anything!`