ETC3250 Lab 9

Di Cook

Week 9

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

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.
c. Compute the error.
d. Does the error get smaller if you use a different kernel?
e. Predict the new data.

Question 2

a. Fit a tree classifier to the data, using the default settings. Print the tree and write down the decision rule.
b. Compute the error.
c. Make a plot that shows the boundary.
d. Plot (on the training data) and predict the new data.
e. Try adjusting the controls (e.e. minimum split), to get a lower error.

Question 3

a. Fit a random forest to the chocolates data.
b. Report the error.
c. Use a parallel coordinate plot to display the data using the importance to order the variables.
d. Predict the new data.

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).

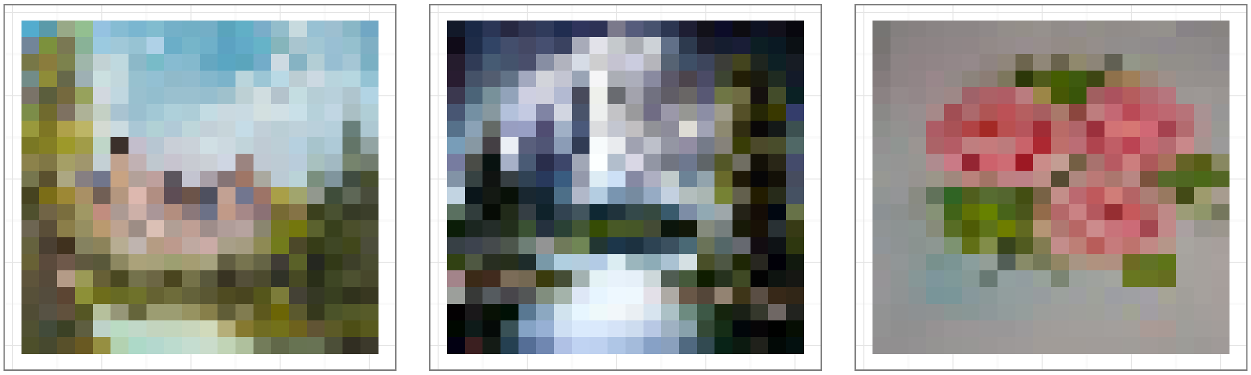
Question 5

This last question is to analyse the happy paintings by Bob Ross. This was the subject of the 538 post, “A Statistical Analysis of the Work of Bob Ross”.

We have taken the painting images from the sales site, read the images into R, and resized them all to be 20 by 20 pixels. Each painting has been classified into one of 8 classes based on the title of the painting. This is the data that you will work with.

It is provided in wide and long form. Long form is good for making pictures of the original painting, and the wide form is what you will need to use for fitting the classification models. In wide form, each row corresponds to one painting, and the rgb color values at each pixel are in each column. With a 20×20 image, this leads to $400 \times 3 = 1200$ columns.

Here are three of the original paintings in the collection, labelled as “scene”, “water”, “flowers”:



a. Explain the difference between the long and the wide format of the data.
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.
e. Which pixels are the most important for distinguishing these two types of paintings?
f. Plot one of the `flower` paintings that was misclassified as `cold`. Can you see any reasons why this might be?

WHAT TO TURN IN

Turn in two items: a `.Rmd` document, and the output `.pdf` or `.docx` from running it. No need to include the R output in your output, but the code should be in the Rmd file. Include your plots in your output.