ETC3250 Lab 10

Di Cook

Week 10

Purpose

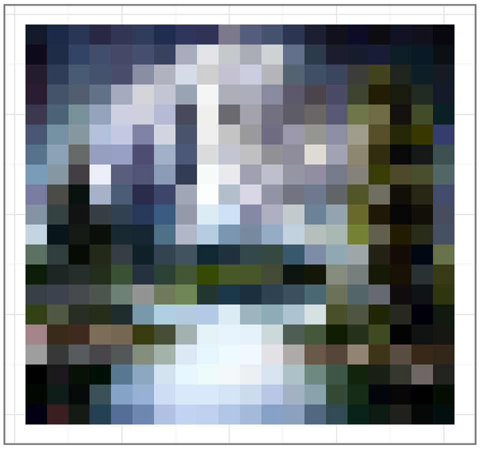
In this lab we will fit a variety of classifiers including penalized LDA and xgboost and compare the performance, for a high dimension, low sample size data set.

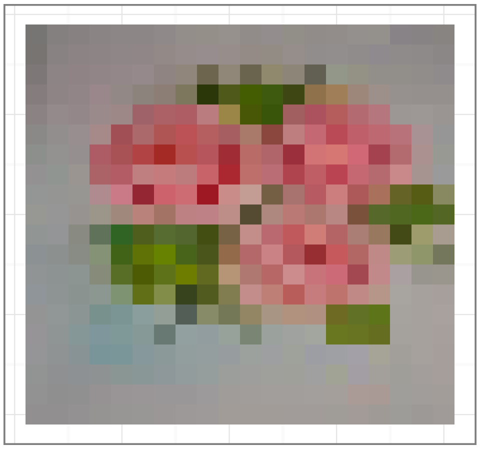
Data

This lab will examine the happy paintings by Bob Ross, using several different classifiers. The paintings were 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. 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 × 3 = 1200 columns.

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





The data has been subsetted to contain just two classes: flowers and cold. The files are `paintings_training_sub.csv` and `paintings_test_sub.csv`.

Reading

Do a little reading about the xgboost algorithm. Here is a starting place: R-bloggers explanation of the xgboost algorithm

Question 1

a. Build a linear discriminant analysis model to predict whether the painting is about flowers or cold theme.
b. Compute the error of the model for the test data.
c. Summarise the coefficients of the LDA classifier.

Question 2

a. Build a penalised linear discriminant analysis model to predict whether the painting is about flowers or cold theme.
b. Compute the error of the model for the test data.
c. Summarise the coefficients of the Penalised LDA classifier.
d. Discuss how these differ fomr the LDA coefficients.

Question 3

a. Build a support vector machine to predict whether the painting is about flowers or cold theme.
b. Compute the error of the model for the test data.

Question 4

a. Build a random forest classifier to predict whether the painting is about flowers or cold theme.
b. Compute the error of the model for the test data.
c. Compare the ten most important variables from random forest with that of penalizedLDA. Is there much overlap in the subset of variables?

Question 5

a. Write a paragraph describing the xgboost algorithm, in your own words.
b. Build an xgboost model to predict whether the painting is about flowers or cold theme.
c. Compute the error of the model for the test data.
d. Tweak the inputs to predict the test as best as you can.

Question 6

Write a couple of paragraphs to compare and contrast the different classifiers for building a model on the paintings data.

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.