

Data Mining I

Final Homework

SELECT ONLY TWO EXERCISES!

(a third may be selected for extra credit)

- 1) (a) Apply bagging, boosting, and random forests to a data set of your choice (not one used in the committee machines labs). Fit the models on a training set and evaluate them on a test set.

b) How accurate are these results compared to more simplistic (non-ensemble) methods (e.g., logistic regression, kNN, etc)? Use the same test/training as in part A.

c) What are some advantages (and disadvantages) do committee machines have related to the data set that you selected?
- 2) Consider the pima data. Use boosting, random forests and a single tree (CART model). Comment on your performance. Explore the partial dependence plots for those variables that have high ranking “variable importance”.
- 3) (ESL Exercise 15.6) Fit a series of random-forest classifiers to the SPAM data, to explore the sensitivity to m (the number of randomly selected inputs for each tree). Plot both the OOB error as well as the test error against a suitably chosen range of values for m .
- 4) This problem involves the OJ data set which is part of the ISLR2 package.
(a) Create a training set containing a random sample of 800 observations, and a test set containing the remaining observations.
(b) Fit a support vector classifier to the training data using $\text{cost} = 0.01$, with Purchase as the response and the other variables as predictors. Use the `summary()` function to produce summary statistics, and describe the results obtained.
(c) What are the training and test error rates?
(d) Use the `tune()` function to select an optimal cost. Consider values in the range 0.01 to 10.
(e) Compute the training and test error rates using this new value for cost.
(f) Repeat parts (b) through (e) using a support vector machine with a radial kernel. Use the default value for gamma.
(g) Repeat parts (b) through (e) using a support vector machine with a polynomial kernel. Set degree = 2.
(h) Overall, which approach seems to give the best results on this data?
- 5) From your collection of personal photographs, pick 10 images of animals (such as dogs, cats, birds, farm animals, etc.). If the subject does not occupy a reasonable part of the image, then crop the image. Now use a pretrained

image classification CNN as in Lab ISLR 10.9.4 to predict the class of each of your images and report the probabilities for the top five predicted classes for each image.

6) A pen-based handwritten digit recognition (pendigits) was obtained from 44 writers, each of whom handwrote 250 examples of the digits 0,1,2,...,9 in a random order. The raw data consists of handwritten digits extracted from tablet coordinates of the pen at fixed time intervals. The last column in the dataset are the class labels (digits). The data can be found here: (<https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits>).

a) How many PCs explain 80% and 90% of the total variation of the data? Display biplots for the first few PCs, color the plots by class (digit). Create a three-dimensional score plot for PC1, PC2 and PC3, color the samples by class.

b) Divide the data into test and training. Fit a kNN model over a range of “k” to the (a) raw data, and (b) PCs from part (A) that capture at least 80% of the variation. Comment on your results.

c) Fit another classifier of your choosing. How do the results compare to B?

7. Fit a neural network to the Default data in the ISLR2 package. Use a single hidden layer with 10 units, and regularization. Have a look at ISLR Labs 10.9.1–10.9.2, and in class, for guidance. Compare the classification performance of your model with that of logistic regression.