Problem Set 3

D Keck

February 19, 2018

Objectives

- Understand information entropy
- Experience a Kaggle competition
- Increase your understanding of image (character) recognition
- Increase your understanding of classification and regression trees
- Understand the k means clustering and Naive Bayes methods

Due March 5

Requirements

- Problem sets are done in groups of 3 (21 groups)
- Each group member must be skilled with the language chosen at the start of problem set
- Each student must do approximately $\frac{1}{3}$ of the development
- The deliverable is single PDF or HTML that meets the standard of a reproducible research report with professional quality communication. Deployable code will not be a deliverable for any problem set.
- All code must be displayed
- Identify and explain each answer. Don't just print a number.
- Numeric output should be easy to read e.g., not 10 decimal places
- Graphics must be easy to read i.e., titles, axis labels, legends, colors, etc.
- The rows and columns of tables and the columns of matrices and data frames must be labeled
- These problem statements provide less detailed guidence, but you are expected to follow the data science process, especially exploratory data analysis and good communication of process and results.
- A maximum of 7 points will be awarded for minimalist correct answers. All 9 points will be awarded for nice explorations and descriptions that include correct answers. Some bonus points will be available as specified below.

Problem 1

Write a general function that computes the information entropy for a data set in a parent node *and* the aggregate information entropy and information gain for any number of partitions (subsets) of that data set in child nodes. The response variable is Bernoulli e.g., play tennis or do not play tennis.

Test your function and output results for *each* partitioning in the example in this video. Note that Prof Patrick Winston is a super famous MIT professor. And yes, you will have to watch the entire lecture.

(https://www.youtube.com/watch?v=SXBG3RGr_Rc (https://www.youtube.com/watch?v=SXBG3RGr_Rc))

Problem 2

Rework the Kaggle Titanic prediction and submit your test set result to Kaggle.

Your work should be done according to the specifications above, *plus* submit your predictions to Kaggle and display your accuracy and ranking from the leaderboard.

You must continue to improve and submit predictions until you surpass the test set accuracy achieved in class: .78947.

You may use any methods and solution parameters in the decision tree family of methods. (Check last slide in the Decision Tree lectures.) The top two teams will get an extra point.

Problem 3

Do the Kaggle / MNIST digit recognizer challenge (or at least a subset that will run on your computer)

Lecun (http://yann.lecun.com/exdb/mnist/)

Kaggle (https://www.kaggle.com/c/digit-recognizer)

- Download the Kaggle training set which includes labels for the digits
- Print the labels using the table() function to assure a random distribution of digits
- Split the 28,000 images in the Kaggle training set into your own training and test sets. Keep 75% in the training set. (The full training and test sets may exceed your computing capacity without a feature reduction scheme, parallel processing, etc.)
- Center and scale the images between 0 and 1
- Choose a method (library function), solution parameters, and training - test set sizes, then predict labels on your test set.
- Continue to develop your method until you reach at least 94% accuracy. Display and label your confusion table. The team with the highest accuracy will get an extra point.
- If your team is able to complete the entire Kaggle challenge with a training set of 28,000 and test set of 10,000, submit and get an accuracy score, then your team will get 2 extra points on this assignment.
- Which predicted digit was most often different from the actual digit ? For example a predicted 1 was most confused with an actual 7.
- Summarize your insights from the development process.

Problem 4

Use k means clustering to analyze the Iris data set. (You may substitute another data set if you prefer, perhaps the UCI Wholesale customers Data Set. (https://archive.ics.uci.edu/ml/datasets/Wholesale+customers (https://archive.ics.uci.edu/ml/datasets/Wholesale+customers))
Write code from scratch, as well as use the kmeans() function. An extra point will be awarded for the team that has the best home grown code performance (time to solution) relative to kmeans() run on the same computer.

Provide a complete narrative of your data science and machine learning solution process. Provide a study of the optimal number of clusters using the *total within-ness* mean squared error. Show all code and display the clusters using 3-D scatterplots.

Problem 5

Use the *mushroom data set* from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/mushroom (https://archive.ics.uci.edu/ml/datasets/mushroom)) to predict whether mushrooms are edible or poisonous.

What are the dimensions of the data set?

What are the response and explanatory variables ? What type are they?

How many mushrooms in the data set are edible and poisious ?

Should the explanatory variables be scaled ?

Split the data set into training and test sets with 75% of data in the training set.

Print the dimensions of each set.

Develop a model using naiveBayes()

Print the conditional probability tables. Choose one and explain the contents.

'Predict' the training labels. Print the confusion matrix and accuracy.

Predict the test labels. Print the confusion matrix and accuracy.

Why is this a good data set for Naive Bayes despite *mushrooms* not being especially interesting ?