CS 696 Intro to Big Data
Fall Semester, 2017
Assignment 4

San Diego State University -- This page last updated 9/11/17

Classification
Due Oct 29 23:59

Titantic Dataset

Source: http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3info.txt

The titanic data set describes the survival status of individual passengers on the Titanic. It does not contain information for the crew, but it does contain actual and estimated ages for almost 80% of the passengers.

The data is in a tsv file with 14 columns which are described below

pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)
boat	Lifeboat
body	Body Identification Number
home.dest	Home/Destination

Notes:

Pclass is a proxy for socio-economic status 1st ~ Upper; 2nd ~ Middle; 3rd ~ Lower
Age is in Years; Fractional if Age less than One (1). If the Age is Estimated, it is in the form xx.5

1. Did age have any affect on the survival of the passengers? Divide the passengers into age groups spanning 5 years each - [0, 5), [5, 10), [10, 15), … . For each group compute the

number of passengers in each group. Then compute the percent of survivors in each group.

For the following problems divide the data into a training set and a test set. After you have created your models in problems 2-4 compute the percent false positives and false negatives you get from your model on the test set.

2. Logistic on age. Using logistic regression with independent variable age and dependent variable survived create a model to classify passengers as survivors.
3. Logistic on age, sex and pclass. Same as problem two but use independent variables sex, age, and pclass. Since sex and pclass are categorical they need special treatment.
4. Decision tree. Instead of using logistic regression use Decision tree with the independent variables sex, age, and pclass.
5. How do the models created in problems 2-4 compare based on the false positives & false negatives the produce on your test data.

Grading

Each problem is worth 10 points.

What to turn in

For problems 1-5 turn in a Jupyter notebook that uses Apache-Toree kernel.

Late Penalty

An assignment turned in 1-7 days late, will lose 5% of the total value of the assignment per day late. The eight day late the penalty will be 40% of the assignment, the ninth day late the penalty will be 60%, after the ninth day late the penalty will be 90%. Once a solution to an assignment has been posted or discussed in class, the assignment will no longer be accepted. Late penalties are always rounded up to the next integer value.