ISEN 613 Fall 2017 Final

Start: November 28 (Tue) 8:00am
Due: November 29 (Wed) 11:59pm

Background

Bike sharing systems are new generation of traditional bike rentals where the whole process from membership, rental to return-back has become automatic. Through these systems, the user is able to easily rent a bike from a particular position and return back at another position. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues. Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns the bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data.

Bike-sharing rental process is highly correlated with the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc., can affect the rental behaviors. The dataset is related to the two-year historical log (aggregated on daily basis) corresponding to years 2011 and 2012 from the Capital Bikeshare system, Washington D.C., USA. The variables are defined as follows:

season	1=spring, 2=summer, 3=fall, 4=winter
year	0=2011, 1=2012
month	1=Jan, 2=Feb, …., 12=Dec
holiday	whether the day is holiday or not (1=holiday, 0=not)
weekday	day of the week (0=Sun, 1=Mon, …., 6=Sat)
weathersit	weather condition of the day 1= Clear, Few clouds, Partly cloudy 2= Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3= Light Snow, Light rain + Thunderstorm + Scattered clouds 4= Heavy Rain + Ice pallets + Thunderstorm + Mist, Snow + Fog)
temp	normalized temperature in Celsius
atemp	normalized feeling temperature in Celsius
hum	normalized humidity
windspeed	normalized wind speed
count	count of users (response)

A dataset ***Bike_data.csv*** with *n*=711 samples is available. The original response ***count*** in this dataset is numerical. Please create a binary variable, ***count01***, that takes on a value of "High" if ***count*** exceeds its median and takes on a value of "Low" otherwise. **Use the new variable *count01* as the response variable in the required analysis.**

Required Analysis

Split the dataset into a training set containing 600 samples and a test set containing the remaining samples. Then conduct the following analysis:

1. Fit a logistic regression model for the training data. Interpret the fitted model. Find its prediction performance (prediction accuracy, sensitivity, specificity) on the test data. **(Note: Show formulas and calculations on each performance measure.)**

2. Fit a tree model for the training data considering both prediction performance and interpretability (that is, your model must be good in prediction and easy to interpret). Interpret the fitted model. Find its prediction performance (prediction accuracy, sensitivity, specificity) on the test data. **(Note: Justify your choices, if any, in the model building process.)**

3. List all the other methods you have learned in this course that can be used for this dataset. For each of those methods, apply it on the training data and then find its prediction performance (prediction accuracy, sensitivity, specificity) on the test data.

4. Summarize the prediction performance of all methods in a table. Which method is the best? Why?

Submission

Submit a **pdf** report about your analysis (including snapshots of R outputs) through the given link

eCampus -> Final_Part B ->Final Exam Report submission

Instructions

(1) Use what you have learned in **Units 1~34** to solve the problems.

(2) To upload a csv dataset in R, you may use the following command or search for help online:
data1 = read.csv("file name.csv", header=TRUE)

(3) Before starting analysis in R, you may need to change the working directory of R (in "File" menu) to the folder containing the dataset.

(4) Report must follow the order of questions. First show the question and then your answer (including R snapshots). Points will be deducted for violations of the required format.

(5) Page limit = **10**. Materials beyond will be ignored.

(6) Show your **INDEPENDENT** work. Collaboration would be considered a direct violation of the Aggie Honor Code and such instances would be dealt accordingly.