EE 660 Homework Week 7: Project Proposal

Posted: Wed., 10/3/2018 Due: Fri., 10/12/2018, 2:00 PM

Please fill in both the Project Proposal form (pp. 1-2) and the Dataset Information Form (p. 3). This is required of everyone (each team submits one HW7 with all their names on it). All fields except "other comments" are required. In each field, replace instructions (black text) with your descriptions. Preferred format is to enter your answers into the Word version of this form, then convert to pdf before submission. If you prefer to use another app instead of Word, then submit a typed version with each field labeled with its title ("Dataset", etc.), and submit as a pdf file.

Please note that this proposal will not be graded like a regular homework. The primary purpose is to give you some feedback on your project topic and plans; the scoring on this homework will be primarily based on whether you put in a reasonable effort and whether the content makes good technical sense.

Insert Project Title Here

Project team: Your name(s) and email address(es)

Omer Solakli: solakli@usc.edu
Tamoghna Chattopadhyay: tchattop@usc.edu
Rohan Amarapurkar: amarapur@usc.edu

Project type (specify which):

(1) Our own design, using real-world data.

Clear statement of the problem and/or goals.

Problem Statement

1. Determine the daily (per night) price for Airbnb Listening in Los Angeles County, which could be the optimal price for the listings
2. Classify the Airbnb Listings based on Customer ratings, estimated price, locality, amenities (if possible) etc., into a maximum of 5 start system for "value for money"
3. Compare the performance of models for the dataset and draw conclusions based on them

A plan of preprocessing and feature extraction (if applicable)

Pre-Processing:

1. Identify categorical features

2. Identify missing data points and assess how to fix them (using mean or median of features, deleting data points, etc.,)
3. Standardize/Normalize necessary features/Discretization

Feature Extraction:

1. PCA, FDA
2. Identify most and least important features

A plan of your approach

After preprocessing the data, we will divide the dataset into four parts: 2000 datapoints for pretraining, the rest divided in the ratio of 60% training data, 20% cross validation data and 20% test data.

We will do pretraining on around 2000 datapoints by the algorithms : SVM, Linear Regression, Lasso , Ridge, Random Forest, Decision Tree Regression, Boosting etc. and select the five best algorithms to apply on the training dataset based on training accuracy. Check how the model generalizes to unknown data.

Then we will use these algorithms for training on the training dataset and then use them for cross validation and testing. We will also evaluate the model's performances based on three metrics, namely, R^2 value, mean squared error and median absolute error.

A description of any other work of yours that is related to your class project

None

If yours is a team project, roughly describe how work will be divided

Yes

We will do the preprocessing together. Everyone will look at implementing one algorithm and check for their effects. Each teammate will also implement one type of error function on their own on the final result.

Other Comments

A potential problem which can arise is the computing time due to the huge dataset available. We will try to reduce the computation time by better feature selection and pre processing.

Include one form for each dataset you plan to use. (For each dataset's form, you may continue onto an additional page if necessary.)

Dataset or competition title: Inside Airbnb

Link: http://insideairbnb.com/get-the-data.html

Student name(s): Omer Solakli, Tamoghna Chattopadhyay, Rohan Amarapurkar

Brief description of dataset and problem domain:
Airbnb listing datasets for Los Angeles County, containing location, description, reviews, nightly, weekly, and monthly availability and pricing, amenities included. The problem is a Regression based problem with an end goal of estimating the nightly cost of Airbnb Listing in LA County.

Number of data points: 43204

Number of features or input variables: 95

Feature or input-variable types: numeric: continuous and discrete - 35
categorical - 7, other (textual data eg. Reviews, location etc) - 53

Label (output) type: numeric

If Label Type is Categorical, is the number of samples significantly unbalanced (maximal variation of more than a factor of 2)? No.

Problem type: regression

Has Missing Data? Yes
The textual features will be removed. For missing data in numerical features, steps will be taken to fix them as discussed in preprocessing step of Project Proposal.