CMPT 459 Data Mining
Spring 2019
Martin Ester
TAs Ruijia Mao and Ruchita Rozario

Course project: Milestone 1

In this course project, we will use the Rental Listing Inquiries dataset from Kaggle (see https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/overview for more information). You will work with a challenging dataset consisting of rental listing data, kindly provided by renthop.com, an apartment listing website consisting of data like description, photos, number of bedrooms, price, etc. The main goal of this project is predicting how popular an apartment rental listing will be on the basis of data mentioned above. However, there are some intermediate stages which enhance your analytical skills and prepare you for the final goal.

Phase1. Exploratory data analysis and data pre-processing

Like all the data mining and machine learning pipelines, we expect you to perform the initial analysis and exploration on the dataset to summarize its main characteristics. This step is a great practice to see what the data can tell you beyond the formal modelling or hypothesis testing task, like discovering the potential patterns, spotting outliers and so on. To this aim, you can apply any meaningful visualization methods and statistical tests. It is worth noting, that you may incrementally improve your exploration section, as you work on the next steps.

In addition, in this phase, you need to perform data pre-processing, which is the practice of detecting and correcting corrupt or inaccurate records from the dataset, by identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Finally, you need to extract features from the unstructured text and images associated with the dataset. You can use traditional feature extraction methods such as those presented in class. You are not expected to use neural network-based methods that have recently become very popular for processing natural language and images.

Submissions

- Exploratory data analysis (Use the training dataset (. . . train) to perform EDA)
    - Plot histograms for the following numeric columns: Price, Latitude & Longitude.
    - Plot hour-wise listing trend and find out the top 5 busiest hours of postings.
    - Visualization to show the proportion of target variable values.

- Dealing with missing values, outliers
    - Find out the number of missing values in each variable.
    - Find out the number of outliers in each variable. Plot visualizations to demonstrate them. You can either remove the outliers or provide a short argument as to why outlier detection is not meaningful for that attribute.
    - Can we safely drop the missing values? If not, how will you deal with them?

- Feature extraction from images and text
    - Extract features from the images and transform it into data that's ready to be used in the model for classification.
    - Extract features from the text data and transform it into data that's ready to be used in the model for classification.

Submit your solution in pdf-format as Milestone1.<GroupName>.pdf in CourSys. Also, provide a link to the code for milestone (which may reside on GitHub etc.) on coursys.

Deadline: February 6, 2020

Marking scheme

Task	Marks
Exploratory data analysis	10 for each sub-task = 3 * 10 = 30
Dealing with missing values, outliers	5 + 10 + 15 = 30
Feature extraction	20 each for image and text = 2 * 20 = 40
Total	**100**