Project Proposal

Course: EE 380L-10 Data Mining (16755)
Team: Chin Wei Yeap, Ravindra Manjunatha, Saharsh Oza, Colin Maxfield, Huy Doan
Topic: Talking Data Mobile User Demographics [1]
Dataset: https://www.kaggle.com/c/talkingdata-mobile-user-demographics

Motivation

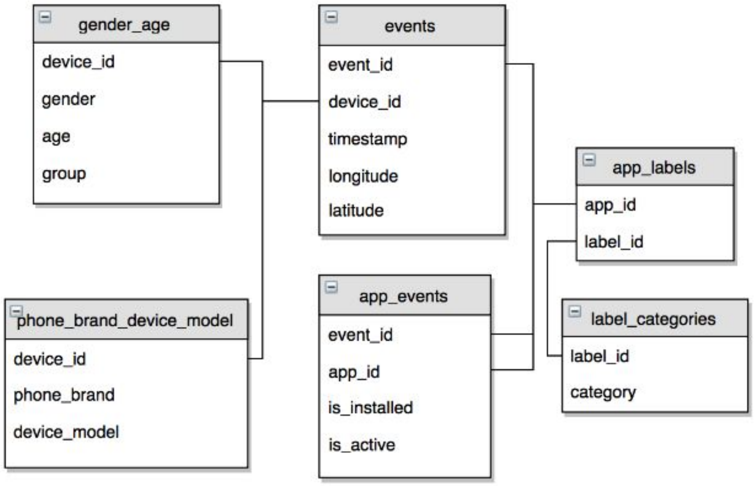
In 2016, there were 4.61 billion mobile phone users worldwide and 2.6 billion of them were smartphone users [2]. With an increasing number of smartphone apps using location and context-aware sensors, smartphones collect large amounts of data on mobile users. In this project, we propose to use Kaggle's TalkingData dataset to predict mobile user demographics (gender and age) based on downloaded app, app activity, location history and other events to detect common trends and patterns within and across mobile users, and finally draw conclusions from these observations. We can also predict mobile user behaviors and characterize probabilities of each behavior in the future.

Design and Methodology

Summary of Dataset

TalkingData is an SDK that runs on a user's mobile device. It samples the user's activities periodically. Each row in the dataset is a snapshot of the user's activity on a mobile device. Below is a brief discussion of the data.

1. Each event has attributes:
   - device_id: Unique user identity
   - Timestamp: Event time
   - Location: Latitude and longitude
   - app_id: Application identity for all the installed apps when the event_id sample is taken.
   - is_active: Specifies which of the given apps are active at a time.
   - is_installed: This is 1 for all. So, it provides no inference.
2. Each app has a category associated with it. For example: finance, gaming, and health.
3. Each device_id has attributes:
   - Phone and brand
   - Gender
   - Age
   - Group: This is a combination of gender and age interval. For example: F21-23. This means a female in the age group of 21-23.



Goal

The goal of the project is to predict the probability of a user (device_id) falling into a set of defined groups. (F21-23, F24-30, M21-23, M24-30). For example:
device_id,F23-,F24-26,F27-28,F29-32,F43+,M22-,M23-26,M27-28,M29-31,M32-38,M39+
1234,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833,0.0833

Cost Function

The cost function is a multi class logarithmic loss. The formula to compute this is given below. N is the number of devices in the test set, M is the number of class labels, log is the natural logarithm, y_{ij} is 1 if device i belongs to class j and 0 otherwise; and p_{ij} is the predicted probability that observation i belongs to class j.

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij})$$

Model Selection

Regardless of the model chosen to classify the data, we believe that splitting the prediction process into 2 stages (the first for gender and the second for age) would be an ideal choice to predict the grouped target which is the weighted aggregate of multiple targets.

1. **Bayesian Methods:** The goal of the classification problem is to predict probabilities of each class. Hence, Bayesian methods seem to be a good fit for this exercise. Techniques including Naive Bayes, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Bayesian Belief Networks (BNN) are natural choices.
2. **Discriminant Methods**: Although Discriminant methods only gives decision boundary and does not give exact prediction probability, we will use them (Support Vector Machines (SVM) (Linear, Polynomial, and Radial Basis Functions (RBF)) and k-Nearest Neighbors (kNN)) for additional high-level and macro group analysis.
3. **Ensemble Techniques:** xgboost and random forests have been popular choices on the Kaggle contest. So, we would explore them with different split metrics: Entropy and Gini.

Feature Selection

One of the most crucial steps of this exercise seems to be to identify the right features from the given data. While this process would be iterative, surveying the forums and some intuition has led us to this initial set of features:

1. Use the timestamps to arrive at weekday/weekend. (Explore time segments during the day as another feature)
2. Number of apps per device per time interval
3. Number of active apps per device per time interval
4. Number of apps in each category for a per device per time interval

It is unclear how we may use the location data, if at all. The phone/brand feature would require extensive knowledge of the location domain to be binned into fewer categorical variables.

Hyperparameters Tuning

In addition to model and feature selection, we still need to select best hyperparameter values for best accuracy. We will experiment with Grid Search CV, Randomized Search CV, Successive Halving, aggressive early stopping convergence, adaptive feature and data subsampling for fast training and high accuracy across very large search size permutation of feature space, model space, and hyperparameter space.

Resources

Software: Python, sklearn library, R, R Studio
Hardware: Laptop, Amazon Web Services (Cloud)

Schedule

We divide the workload to different tasks and models and each team member experiments on specific model and reports individual findings. We will also combine and ensemble different models to give better accuracy. We will follow the schedule as below:
Week 1 (3/20/2017-3/26/2017): Brainstorm ideas, finalize dataset, and write proposal.
Week 2 (3/27/2017-4/2/2017): Explore, feature engineering, preprocess, and clean data.
Present idea in class (3/30/2017).
Week 3 (4/3/2017-4/9/2017): Code, run multiple models, analyze, test, and debug.
Week 4 (4/10/2017-4/16/2017): Code, run multiple models, analyze, test, and debug.
Week 5 (4/17/2017-4/23/2017): Code freeze and combine results from all team members.
Week 6 (4/24/2017-4/30/2017): Present final project (Late April/Early May), and final report.
Week 7 and 8 (5/1/2017-5/10/2017): Complete final report and submit by 5/10/2017.

Reference

[1] Kaggle Talking Data Mobile User Demographics,
https://www.kaggle.com/c/talkingdata-mobile-user-demographics
[2] Device Atlas, “16 mobile market statistics you should know in 2016”,
https://deviceatlas.com/blog/16-mobile-market-statistics-you-should-know-2016