Social Dating Matching Prediction using SVM
&
Clustering / SVD analysis for logistic regressions

SangHyeon (Alex) Ahn, Jin Yong Shin

1. Abstract

Speed dating have become a popular method for seeking partners especially for the people who have limited access to the pool of people. Finding the right partner whom aligns with the set of characters that a person seeks can be beneficial in many aspects: reduced time and effort, selecting a candidate from a larger pool, quantitative/qualitative guarantee for the search of a true love. Being able to cluster set of people who have similar traits/characteristics can help to narrow down the pool of potential partner. Also, non-linear (or generalized) regression model that gives either the predictive probability of developing into a relationship or binary classification indicating a good match can significantly improve in terms of speed dating service. In our paper, we will explore both k-means(or lambda-means) clustering and Naive Bayes clustering methods for clustering, and Logistic Regression, Support Vector Machine for classification prediction. Lastly, we will combine two methods to use cluster ID's as features and compare them with Primary Components.

2. Method

a. Algorithms
    i. Clustering
        1. Lambda-means
        2. Naive Bayes
    ii. Evaluation / Classification
        1. Logistic Regression
        2. Support Vector Machine
b. Cross Validation
    i. Using method of Train/Dev/Test datasets for both clustering and classification
    ii. Performance Testing
        1. Clustering
            a. Variation of Information (VI)
            b. Number of clusters
        2. Evaluation/Classification
            a. Accuracy Test on Test set labels
c. Feature Selections
    i. Purpose
        1. We are going to look at the data that has the most relevant data
        2. Look for correct attributes to remove complexity
    ii. Feature includes:

1. Binary label indicating ("Yes" or "No" for a match described as 1 or 0)
2. Attributes: gender, age, race, profession, level of education, income, attractiveness, sincerity, intelligence, humor, interests.
3. Candidate's attributes
4. Candidate's desirable attributes on opposite sex
5. Partner's attributes
6. Partner's desirable attributes on opposite sex
7. Shared Interest Correlation

iii. Regularization (overfitting issues)

1. One reason doing feature selections is to reduce the possible overfitting problem. We may consider L1, L2 regularization as we compare the results

3. Resources

a) Data

The data set that we are using is called "Speed Dating Data Key" and is obtained from the website(https://www.kaggle.com/annavictoria/speed-dating-experiment). This data set contains actual experiment data of dating experiment of total 21 waves

b) Libraries

As we are planning on implementing learning algorithm ourselves, we will not make use of libraries outside of standard python libraries (and some for mathematical operations i.e. numpy, scipy)

4. Milestone

a. Binary prediction (supervised learning)
   i. Support Vector Machine
      1. Using feature matrix of candidate and partner to make match prediction.
      2. Sampling datasets: Train (40% of the data), Test (60% of the data)
      3. Training prediction model based on refined/sampled data
   ii. Regularization
      1. L1 (Lasso) vs. L2 (Ridge)
      2. Choice of Lambda
   iii. (Potential) Reasonable prediction accuracy on binary classification
b. Clustering (unsupervised learning)
   i. Clustering of the speed date candidates using Lambda Means Clustering
   ii. Generate Cluster ID's
   iii. Some number of clusters with a reasonable Variation of Information
   iv. (Potential) Appropriate, accurate, and relatively small number of clusters with a small Variation of Information
c. Combinatorial Prediction
   i. Using Cluster ID as a feature (dimension reduction)

ii. Quantitative evaluation of a good match
    1. We will return a continuous value
    2. Logistic regression function returns a value from 0 to 1 (probability)

d. Additional Research
    i. Comparing cluster means to Primary Components (SVD eigenvectors)
    ii. Finding the most impactful attribute in speed dating

5. Final Write-Up

a. Introduction
b. Background
    i. Motivation
c. Preparation
    i. Feature Selections
        1. Data refinements
    ii. Set-up files (i.e. classify.py)
d. Algorithms
    i. K-means (Lambda means)
    ii. Naive Bayes
    iii. Logistic Regression
    iv. SVM
e. Evaluations
    i. Cluster Analysis
    ii. Accuracy Testing
    iii. Cross Validation
f. Experiments
    i. Different models
g. Results
    i. Evaluation of milestones
h. Analysis
    i. Pros and cons
    ii. Efficiency
    iii. Utilization
i. Conclusion
    i. Remarks