SI 671/721 (Fall 2020) Data Mining: Methods and Applications

Instructor: Paramveer Dhillon

Homework 2:
Speed-dating: Match Prediction and Recommendation

Due: 10/21/2020 (Wednesday)

1 Summary

We will use the speed-dating dataset described in Fisman et al. (2006) for this homework. The dataset was gathered from 552 participants in several speed dating events from 2002-2004. As is typical, in each of these speed-dating events each attendee met every participant of the opposite gender for a four-minute first-date in which they got to know each other's interests and determined their compatibility for going out on subsequent dates. At the end of four minutes, all the attendees were asked if they would like to see their date again as well as to rate their date on six attributes: attractiveness, sincerity, intelligence, fun, ambition, and shared interests. The dataset also contains several other features/covariates regarding the participants and their speed-dates.

This homework involves predicting whether an attendee matched with another attendee, and further predicting the missing rating given by an attendee to their dates.

We recommend that you use Jupyter Notebooks and Python libraries Numpy, Sci-kit learn, and Pandas for this homework.

2 Details

This homework is divided into three parts.

1. Data Exploration.
2. Predicting matches.
3. Recommending potentially "good" matches.

2.1 Part 1: Data Exploration [15 Points]

This part of the homework is designed to help you familiarize yourself with the dataset and the context in which it was collected. The insights from this part of the homework can help you in building the prediction and recommendation models for the Parts 2 and 3 of the homework.

a). Read Fisman et al. (2006) closely to understand the context of the data collection.

b). Read `SpeedDatingDataKey.pdf` to familiarize yourself with the various variable definitions.

c). Perform basic exploratory data analysis, for example,

- Visualize and describe some or all of the variables. You might choose to graph combinations of variables. You might also choose to graph variables in isolation.
- Perform correlation analysis.
- Deal with missing values and outliers.

This part of the assignment has a written conclusion. In the written conclusion please detail how missing values and outliers were handled and justify your decisions made. Consider how your handling alters the distribution of the dataset.

The main deliverable for this part of the homework is a step-by-step exploration of data along with text describing your conclusions in your Jupyter Notebook. The analysis pipeline described above is just a suggestion and you are welcome to add more or fewer steps as long as it helps you carefully analyze the properties of the dataset.

2.2 Part 2: Predicting Matches [35 Points]

You will build and compare machine learning models to predict if participants X and Y who met for the speed-date actually "matched" i.e. after the event finished they did choose to see their date again. Note that our data contains both the observations, when participant X is the "ID" and Y is the "partner ID", and when Y is the "ID" and X is the "partner ID". It's easy to see that this is an asymmetric prediction task—X might choose to see Y again but Y might not choose to see X again.

We provide you with training dataset which you should use judiciously to train your models. We also provide a test dataset `testML.csv` where the "match" label is missing.

We recommend you try at least 4 different machine learning methods/models before choosing the final model. The error-metric that we will use for evaluating your match labels on the test dataset is weighted f1 score (higher is better). Some of the models that you can consider using include,

- Random Forests
- Lasso Regression
- Logistic Regression
- Support Vector Machine (SVM)
- Multi-layer perceptron (MLP)
- Ensemble Methods

You may use feature selection techniques to find important features.

For the written analysis part of this homework we would like you to interpret the results from at least two machine learning models. One of these machine learning models should be the final model that you have chosen in this part of the homework. The written analysis should contain a short paragraph for each model and include a comparison of model results.

The main deliverable for this part of the homework is a step-by-step analysis of your model building exercise describing clearly why you chose one model over the other. Your Jupyter notebook should contain the reproducible code for training various models as well as text descriptions of your conclusions after each step.

Your grade on this part of the homework will depend on the accuracy of your model on the test dataset as well as your step-by-step description of how you arrived at your final model. We will evaluate your model using weighted F1-score (e.g. `sklearn.metrics.f1_score(y_true, y_pred, average='weighted')`).

2.3 Part 3: Build recommendation Engine [50 Points]

For the final part of the homework you have to build a recommendation engine where you have to predict the missing ratings (1-10 scale) of a participant X for their date Y. You are encouraged to use the surprise Python library (http://surpriselib.com) and compare at least 4 algorithms for generating the missing ratings e.g. SVD, SVD++, NMF, k-NN, SGD, etc.

We provide you with training dataset which you should use judiciously to train your models. We also provide a test dataset `testRec.csv` where the partner ratings are absent.

As earlier, your Jupyter Notebook should contain a step-by-step analysis of your model building exercise describing clearly why you chose one model over with supporting evidence.

Your grade on this part of the homework will depend on the accuracy of your model on the test dataset as well as your step-by-step description of how you arrived at your final model. We will evaluate your model using the Mean Absolute Error (MAE) metric.

3 Data Description

Here's the description of files included with this homework.

1. `trainML.csv`: This file contains the training data for Part 2 of the homework. It contains 3808 observations and 61 columns/features. The column keys are described in the file `SpeedDatingDataKey.pdf`.
2. `testML.csv`: This file contains the test data for Part 2 of the homework. Please insert your prediction results in the **match** column in the file.
3. `trainRec.csv`: This file contains the training data for Part 3 of the homework. The complete training data set has 38340 observations. It contains ratings on a scale of 1-10 given by an attendee with id `''iid''` to their partner with id `''pid''`.

4. `testRec.csv`: This file contains the test data for Part 3 of the homework. Please insert your predicted ratings in the **rate** column in the file.

Note: Please do not use the data in `trainRec.csv` to test or evaluate Part 1 or Part 2 of the homework. We use different groups of users, thus, their ID won't match with each other.

4 Submission

All submissions should be made electronically by 11:59 PM EST on October 21, 2020.

Here are the main deliverable files:

- HTML version of your Jupyter notebook.
- The actual Jupyter notebook with "step-by-step analysis" for all the three parts of the homework. It's fine to submit everything in a single notebook. (So that we could replicate your results.).
- File `testML.csv` with your predicted 0/1 labels for Part 2 of the homework. Keep all the columns in the file `testML.csv` which we shared with you, as they are. Just update the file with your predictions in the correct column. (We will use automated scripts to evaluate the performance of your model, so please do not make any other changes to the file).
- File `testRec.csv` with your predicted 1-10 ratings for Part 3 of the homework. Again, keep all the columns in the file `testRec.csv` which we shared with you, as they are. Just update the file with your ratings in the correct column. (We will use automated scripts to evaluate the performance of your model, so please do not make any other changes to the file).

5 Academic Honesty

Unless otherwise specified in the homework, all submitted work must be your own original work. Any excerpts, statements, or phrases from the work of others must be clearly identified as a quotation, and a proper citation provided. Any violation of the University's policies on Academic and Professional Integrity may result in serious penalties, which might range from failing a homework, to failing a course, to being expelled from the program. Violations of academic and professional integrity will be reported to the concerned authorities. Consequences of academic misconduct are determined by the faculty instructor; additional sanctions may be imposed.

References

Raymond Fisman, Sheena S Iyengar, Emir Kamenica, and Itamar Simonson. Gender differences in mate selection: Evidence from a speed dating experiment. *The Quarterly Journal of Economics*, 121(2):673–697, 2006.