Have You Met The 1? A Machine's Approach to Human Relationships

Jiayu Lou, Hang Yang

Introduction

We have all pondered about the same thing in a relationship: is he/she the right one for me? Countless articles, quizzes, suggestions, tips, and consultants out there are trying to answer this question for us. They bring out big words to make it logical; they tell us to mind the differences, but also note that some differences matter and some do not; they place their own arbitrary weights on terms such as ambition, core values, intelligence, emotional intelligence, spiritual beliefs, etc. It makes us wonder: if we are already talking about logic in romance, why don't we take a step further? What if, instead of reading these sources and still remaining puzzled about the relationship, we can tell you with a high confidence level exactly how likely you are going to be with your significant other for the rest of your life with the help of a machine learning algorithm? Thanks to How Couples Meet and Stay Together dataset (HCMST) by Rosenfeld, Michael J., Reuben J. Thomas, and Maja Falcon, it is now possible, based on your everyday habit, your usage of Internet, your marriage status, and other miscellaneous information about your life, to predict if you will stay with him/her for at least the next few years.

Related Work

The HCMST dataset has been utilized in various ways concerning different social study fields, yet most of the papers are based on a few arbitrarily chosen measurements, and no Machine-Learning-related work is found. Early work entails the Internet-related data to argue increasing Internet coverage has increased chance of meeting partners for GLB (Gay, Lesbian, Bisexual) individuals, and Internet as a social intermediary has partly replaced traditional dating spaces [2]. On a similar note, data on the respondents' sexualities and their relationship longevity has been used to support the assertion that same-sex couples' break-up rates are comparable to heterosexual couples' [3].

More recent working papers discuss the topic of relationship stability as a whole, following the main motives behind the HCMST project. However, restricted by the limits of human, these papers tend to focus on selected areas of the whole dataset, drawing conclusions from partial observations and features, thus indirectly putting arbitrary weights on the subject matters. One such example chooses to investigate the relation between relative earnings in the household and the relationship stability [4], while the others select subjects of gender and marital status in order to explore the break-up rates of heterosexual couples [5].

Potentially due to the novelty of the data and the fact that it has only finished its budgeted five waves very recently in 2015, until now there exists minimal efforts to attempt bringing all parameters together under the roof of machine learning algorithms. We are here to pioneer.

Dataset and Feature Selections

I. Initial Data

HCMST conducted 5 sequential rounds of surveys in 2009, 2010, 2011, 2013 and 2015, respectively. The initial dataset consists of two parts: the respondents' answers to the original questions (e.g., recorded answers to question "How old are you?"), and features generated from the collected raw data (e.g., the categorization of cases based upon age division). The dataset includes 4002 respondents with 370 features, supplemented by additional 62 features from wave 4 survey and 78 features from wave 5 survey results.

II. Data Elimination

Since this project aims at predicting the future relationship status based on current info, we believe that at this point only the data collected in the first wave is relevant to our purpose. We have hence dropped data collected from wave 2 to wave 5, keeping only the survey results on their relationship status at each milestone.

The survey assumed that some couples who didn't respond to follow-up surveys still stayed together. For the sake of precision, however, we only kept data of respondents with partners in the beginning of the timespan who continuously responded to the surveys in following periods until wave 5 or breakup. Based on the results from the 4 follow-up surveys, we generated feature "final_relationship_status" (Boolean) to indicate the final status of couples after 6 years. After deleting all the redundant features and observations, we are left with 1569 respondents with 269 features.

III. Preprocessing Non-Standard Data Values

The initial data contains many features with a substantial amount of missing values. While some bear minimal relevance to our goal (e.g. gender of the 15th member in your family) and can be dropped without significant impact, other missing values are indicative of important information and dropping them will result in high bias. Therefore, for those question answers that have already been processed, we only kept their corresponding feature. For example, for question 34 "how would you describe the quality of your relationship?", we dropped feature "q34" and kept the corresponding processed feature "relationship_quality". This will prevent problems generated by singular matrix in future prediction models. Secondly, some of the missing values are results of branching questions. For example, if question 12b is only required for people who answered "yes" for question 12a, then feature "q12b" will contain lots of missing values. For these features, we integrated them into the main branch question by generating more categorical classes in features like "q12a". Thirdly, we dropped clearly unrelated features with high level of missing values or non-numerical and non-categorical values. These operations leave us with 148 features to work with.

IV. Feature Selection

Considering that the size of our datasets after processing stays around 1600, we decided to include fewer features to avoid potential overfitting. Therefore, we have implemented the forward-based sequential feature selection based on Logistic Regression model with cross-validation of 10-folds. Features were selected based on misclassification rate using Logistic Regression model and feature selection terminates after the misclassification error no longer improves. This leaves us with 47 features, with 10 most important features:

1	"coresident": if the couple lives together
2	"relationship_quality": how the respondent ranks relationship quality
3	"s1": the status of current relationship: married, sexual or romantic partner
4	"age_difference": the difference in age between respondent and partner
5	"How_long_ago_first_met": how long ago the respondent meets his partner
6	"Distancemoved_10mi": the distance between hometown and current home
7	"Papevangelical": if the respondent identifies as evangelical christian
8	"Children_in_hh": number of children in household
9	"How_long_ago_first_cohab": how long ago the respondent starts to live with partner
10	"Ppincimp": the household income categorized into 7 classes

Prediction Models

I. Logistic Regression

We first applied logistic regression model trying to minimize the loss function with L1-regularization:

$$J(\theta) = \frac{1}{m}[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)^2} + \lambda \sum_{j=1}^{n}|\theta_j|]$$

$$and\ h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$$

Any output of logistic regression is in the range {0,1}, where output smaller than 0.5 will be categorized as 0 and the rest categorized as 1. This method generates misclassification error of 11.1% for train set and 12.39% for test set.

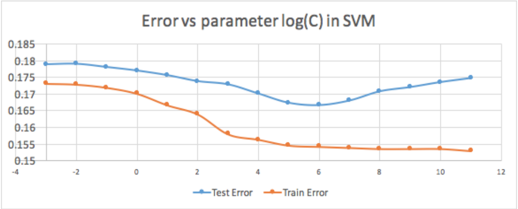
II. Support Vector Machine

We implemented the the SVM without kernel:

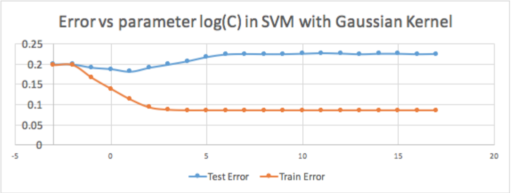
$$min\frac{1}{2}w^T w + C\sum_{i=1}^{n}\xi_i$$

$$subject\ to\ y_i(w^T\varphi(x_i) + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1, \ldots, n$$



Then we integrated the Gaussian kernel into SVM, which didn't improve the result. Observe that the train error using Gaussian kernel is substantially smaller than normal SVM without kernel. Because the number of observations is relatively small, we believe that using kernel would further complicate the method and therefore result in overfitting.

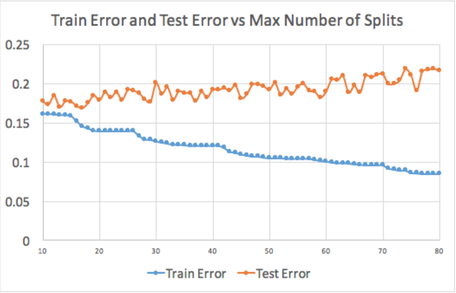


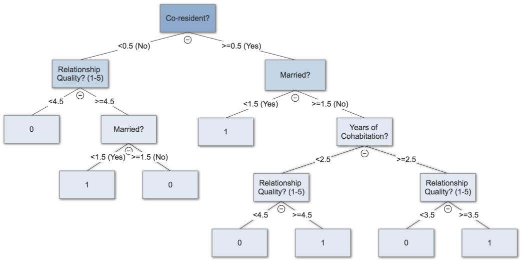
III. Decision Tree

Many of the survey questions have sequential correlation with each other and some features' existence are entirely based on others' (i.e., only respondents who have answered "yes" to question "have your religion changed since 16?" will be asked to answer "what is your religion at 16?"), therefore we believe that the decision tree model would be a proper representation of the set of if-then choices and would replicate the design logic behind the survey.

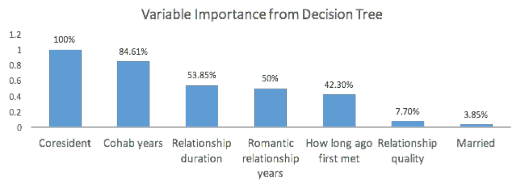
We generated the top-down binary decision tree by examining the optimal statistical improvement brought about by each feature at each split. To capture the optimal improvement, we ordered both the categorical and the continuous attributes from the smallest to the largest and measured the improvement in misclassification error by dividing at each consecutive pairs. When a missing value is encountered, we used surrogate split because many alternative features with high variance can be found.

In actual prediction, we first generated the binary decision tree with minimum branch size of 10 observations and unlimited depth, and this results in misclassification error of 4.07% for train set and 22.67% for test set after 10-fold cross-validation. As certain level of overfitting is shown, we decided to pre-prune the tree by adding the maximum number of splits in our decision tree. After trial and error, 13 splits generate the lowest test error and achieves a relative good balance between the test set and the train set, with misclassification error of 14.96% and 17.75%, respectively. For the convenience of representation, we have pruned the tree to have at most 13 splits, as shown below.





Despite decision tree helps to visualize the primary features at each split, surrogate variables may never be used in actual splitting. Therefore we also calculated the variable importance trying to capture all the highly important variables by measuring the improvement attributable to each variable either as a primary or a surrogate splitter. Below is a graph showing the importance of features with non-zero importance value. Note that the 2nd to the 5th variable doesn't appear in splitting at all but might serve as surrogates for "Coresident". Also, despite that there are over 40 features, the variable importance table shows that only very few of them are decisive features and actually have an effect in splitting.



IV. Naive Bayes

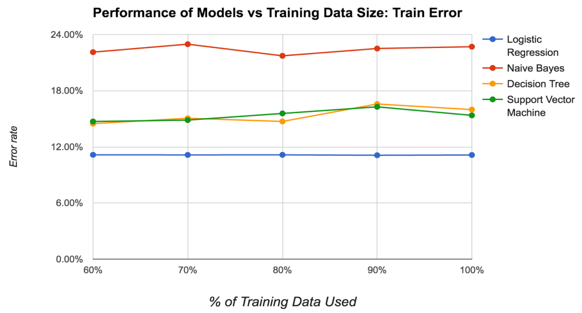
Our final dataset is left with fewer than ten numeric values: the distance from the respondent home to the current home, how long ago they first met and how long ago the respondent first lives together with partner, etc. In order to apply Naive Bayes model, we have discretized the data by converting these numeric features into several categorical classes. Because the assumption of the Naive Bayes is that every conditional probability is independent of each other, we also calculated the covariance matrix between all the features and dropped features with a covariance over 0.3. We then applied the Naive Bayes model trying to maximize

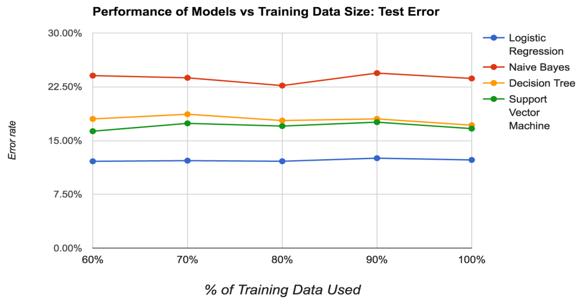
$$L = \prod_{i=1}^{m} p(x^i, y^i) = \prod_{i=1}^{m} (\prod_{j=1}^{n_i} p(x_j^i|y))p(y^i)$$

Because some of the survey questions have relatively small amount of responses or unbalanced results, we also added Laplace smoothing to ensure at least one data point per feature per class. Using the obtained probabilities $\phi_{k|y=1}, \phi_{k|y=0}, \phi_y$ we then cross validated by partitioning the data in 10-folds and obtained the averaged misclassification error 22.27% for train set and 23.19% for test set.

Results and Analysis

At this step, we have graphed the test error and train error regarding each method after cross-validation. As the graph have shown, the errors generated by all the methods range from 13% to around 24%.





As the graphs have shown, generally logistic regression produces the best result while Naive Bayes performs the worst. Decision tree and support vector machine have very close performance in both test and train dataset.

Model	False Positive	False Negative	Precision
Logistic Regression	23.67%	9.05%	87.70%
SVM	29.20%	9.24%	83.33%
Naive Bayes	17.26%	25.15%	76.34%
Decision Tree	32.96%	8.35%	82.85%

One possible reason that Naive Bayes doesn't generate good precision is that some of the input features are not completely mutually independent. Despite that we have used sequential feature selection and later removed features with correlation greater than 0.3 when using NB, some features left are still vaguely related with each other and this violates the basic independent assumption of Naive Bayes. The correlation between variables, however, helped to boost the precision in Decision Tree Model because the missing values can be replaced with their correlated alternatives using surrogate splits.

We believe that one explanation for the logistic regression to generate better result than decision tree is the high dimensionality of the dataset compared to the number of observations. As trees always tends to overfit in presence of high dimensionality since it has high freedom degree, we had to limit the number of splits to prevent overfitting. However, some important information are lost in this process and bias is sacrificed to obtain lower variance. Logistic Regression, though very simple, does draw information from the basis of the entire set of features and therefore could perform better than tree-based model.

Note that the false positive rate is almost about 3 times as high as the false negative rate in logistic regression, SVM and decision tree. This is partially because the initial dataset is imbalanced with the ratio between positive data and negative data being 2:1. We then decided to rebalance the dataset by adding a weight vector to assign more weights to negative classes. The table below shows false positive and false negative results generated using the re-balanced dataset. Notice that the the precision rate doesn't stay stable for all the four models and false-positive and false-negative rates are more balanced than before for logistic regression, SVM and decision tree. However, the false-negative rate from NB is still substantially higher than the other three, meaning that Naive Bayes model is very pessimistic about the couple's relationship: it tends the believe a couple would break up even indeed they will very likely not.

Model	False Positive	False Negative	Precision
Logistic Regression	18.81%	13.42%	86.60%
SVM	16.81%	14.12%	81.94%
Naive Bayes	16.37%	27.14%	77.96%
Decision Tree	18.81%	17.00%	81.15%

We also tried to apply the logistic regression and decision model to predict the data points that have been deleted from our dataset due to (1) missing values; (2) unknown labels. These observations were first removed from our samples because the respondents stopped to respond to the survey from the second, third or fourth round. Interestingly, the pattern demonstrated by our prediction shows that the earlier the couple stops to respond to the survey, the more likely they will get a "0" in prediction. In other words, for couples that stop to respond to the survey since round 2, our model predicts that very large portion of them will break up in 6 years. This verifies our guess: people don't just quit in the middle of the survey for random reasons; the absence of a couple's voices in later surveys might already indicates a deceased romance, and the pain and embarrassment to admit this usually makes people shun away.

Conclusion

While our models express >10% test errors, it is rather reasonable given that relationships are still indeed based on one of the most complex systems in the known universe: human mind. Our results provide insights of what otherwise remains mysterious and unquantified, and will potentially help sociologists and everyday individuals alike.

In process of fitting samples through the models, we have also discovered that adding more samples don't always result in higher accuracy. This is likely rooted in the nature of relationships and romance: it's the surprise and unexpected turn of events that highlight their beauty. When consensus deems long-distance relationships hard to maintain, there are always outliers who prove it wrong, and same goes for other difficulties in love. For this exact reason, while increasing data size from additional surveys will theoretically improve our prediction, we believe that it may not be necessary.

References

[1] Rosenfeld, Michael J., Reuben J. Thomas, and Maja Falcon. 2015. "How Couples Meet and Stay Together." Stanford, CA: Stanford University Libraries. waves 1, 2, and 3 version 3.04; wave 4 supplement version 1.02; wave 5 supplement version 1.0. http://data.stanford.edu/hcmst.

[2] Rosenfeld, Michael J. and Reuben J. Thomas. 2012. “Searching for a Mate: The Rise of the Internet as a Social Intermediary.” American Sociological Review 77:523–47.

[3] Rosenfeld, Michael J. 2014. “Couple Longevity in the Era of Same-Sex Marriage in the United States.” Journal of Marriage and Family 76: 905–918.

[4] Weisshaar, Katherine. 2014. “Earnings Equality and Relationship Stability for Same-Sex and Heterosexual Couples.” Working paper.

[5] Rosenfeld, Michael J. 2015. “Who wants the Breakup? Gender and Breakup in Heterosexual Couples.” Working paper.