Machine “Gaydar”: Using Facebook Profiles to Predict Sexual Orientation

Nikhil Bhattasali[1], Esha Maiti[2]
[1]Symbolic Systems, [2]Mathematical and Computational Science, Stanford University

CS 229: Machine Learning
Fall 2015

Introduction & Motivation

What can an average college student’s sparse Facebook profile tell you? A surprising number of things.

With the rise of social media, we have all cultivated our online identities based on what we share and how we communicate with others. Even if we think that we have not shared much information on our profiles, the large, rich datasets that are accessible allow us to draw conclusions about different individuals from implicit information in the form of trends and patterns in a population.

We chose to concentrate on sexual orientation. On one level, a learning algorithm to distinguish non-heterosexuals is a lighthearted way to study the power of implicit information and allow interested parties to connect with similar others. On a deeper level, such a project would allow us to learn more about each other in an organic way: a learning algorithm allows us to minimize prejudices and prior beliefs in the study; it is very much initialized as a “blank slate” and develops a model based on the data.

Feature Selection & Labeling

What do we look at on a Facebook profile?

About:
- Hometown
- Religion
- Political views
- Interested in
- Relationship status

Profile Photos:
- Rainbow filter (2014)
- Equality sign (2012)
- Number with 1 female
- Number with 1 male

Friends List:
- Ratio of male to female friends
- Number of non-heterosexual friends
- Ratio of non-heterosexual to heterosexual friends

Timeline:
- Shared links
- Photo comments
- Status updates

Each training example was assigned a label y = 1 (non-heterosexual) or y = 0 (heterosexual) based on its source: the Stanford LGBT secret groups or the official class groups on Facebook, respectively. We used a simple random sample of undergraduate Stanford students to create our training set.

Model Components

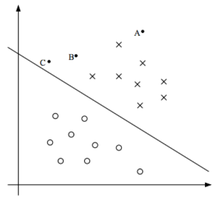
Naïve Bayes Classifier

A Naïve Bayes classifier is especially useful for determining the probability of a profile’s non-heterosexuality given discrete data features, such as the explicit attributes or the timeline text.

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}$$
$$\phi_{j|y=0} = \frac{\sum_{i=1}^{m} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{m} 1\{y^{(i)} = 0\}}$$
$$\phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)} = 1\}}{m}$$

Support Vector Machine

A Support Vector Machine is flexible and natural for classification problems, and admits the use of kernels to transform the data to higher dimensional spaces, possibly making the classification more accurate (if a linear boundary is insufficient to separate the classes).

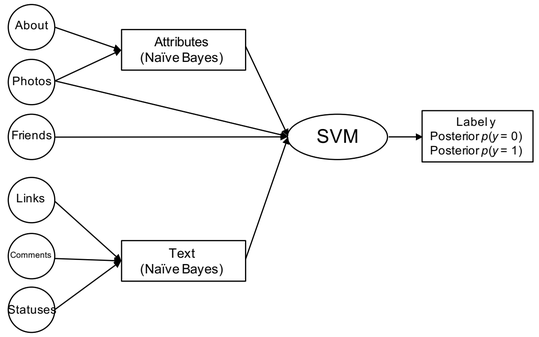


Prediction Model

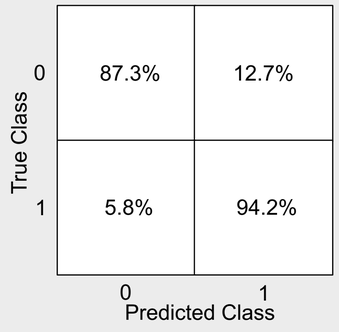
How does our classifier create a prediction model?

The model has two Naïve Bayes classifiers: one that analyzes attributes in a profile’s About and Photos section, and one that analyzes the text from the profile’s Timeline. Each of these computes a score that is fed into the Support Vector Machine and combined with other continuously-valued features to create a final classification.

This is illustrated in the diagram below:



Results I

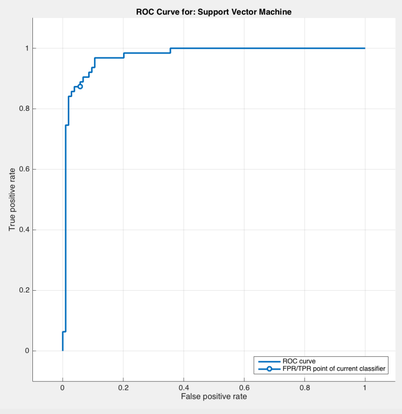


Confusion Matrix

Also known as a “contingency table” or an “error matrix”, this visualization shows the predictive performance of the classification algorithm. The top-left box shows the true positive rate, and the bottom-right box shows the true negative rate. The top-right box shows the Type I error (false positive error), and the bottom-left box shows the Type II error (false negative error).

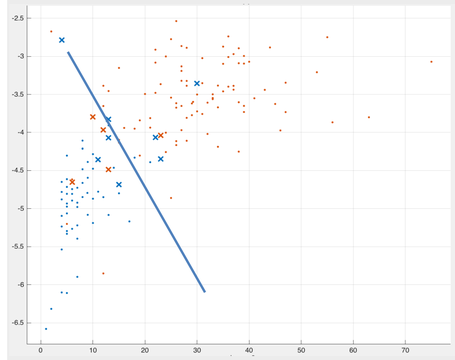
ROC Curve

The ROC curve summarizes the performance of a classifier over all possible thresholds. The optimal curve has an area underneath it equal to 1. This curve has curve has an area underneath it equal to 0.973.



Results II

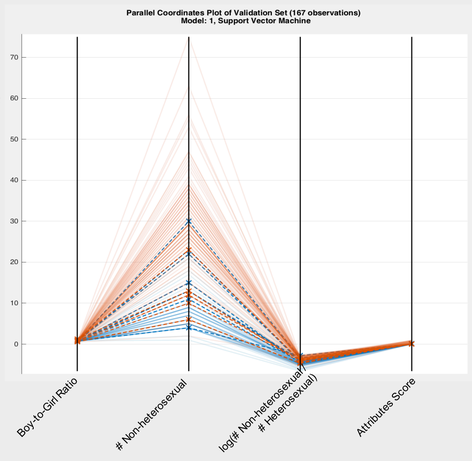
Decision Boundary



The feature vectors in this model are too high-dimensional to visualize meaningfully, even with a dimension reduction algorithm such as PCA. Therefore, this plot shows the decision boundary for just two features: the number of non-heterosexual friends (x-axis) and the log ratio of non-heterosexual to heterosexual friends. A remarkably clear separation between the two classes can be observed.

Parallel Coordinates Plot

This is another method of visualizing high-dimensional geometry and analyze multivariate data. The ordering of the data points suggests that a linear boundary is reasonable.



Implications & Future Work

Although an individual profile may not share much information explicitly, it is possible to use the aggregation of sparse data from many profiles to train a learning algorithm to tell us whether certain present (or missing) data contributes to the probability of that student being heterosexual or non-heterosexual.

In the future, we could consider using a neural network on the same data, as it may be interesting to study the features that the algorithm learns and how they confirm or challenge different stereotypes.

Acknowledgments

We would like to thank our professor Andrew Ng and our project mentor Sam Corbett-Davies for guidance, as well as Stanford’s Department of Computer Science for the opportunity to perform and present this research.