Proactively Discouraging Cyberbullying using SVMs

Michael Kossyrev, mkossyre@stanford.edu

Stanford University 2017

Motivation

1. As of 2012, over 3 million kids per month are absent from school, and approximately 4,500 kids kill themselves each year at least in part because of cyberbullying [1]
2. Goal: build a text parser that takes in a text conversation, parses it into a boolean feature vector using a learned dictionary, then trains an SVM to identify bullying.
3. Results: The SVM correctly identified bullying vs non-bullying instances 95% of the time, and missed bullying instances (identified bullying as non-bullying) only 1% of the time on the test set.

Data

1. Myspace data retrieved from `http://chatcoder.com/DataDownload`
2. Contained 2,084 XML files of 10 interactions (lines of text) between online chat users.
3. Contained 11 XML files labeling which files contained bullying and which did not.

```
<post id="MS_5090_843779">      <user id="MS_1602656">
   <username>Stacy</username>        <sex>F</sex>
<age>24</age>        <city>CALIFORNIA</city>
<province/>        <country>US</country>      </user>
<date>1105393740</date>      <body>      Exactly.  </body>
   </post>  <post id="MS_5090_884431">      <user
id="MS_1961883">        <username>dylano</username>
<sex>M</sex>        <age>27</age>        <city>Huntington
Beach</city>        <province>California</province>
<country>US</country>      </user>
<date>1105894380</date>      <body>lol no its not, its
surrounded by an upper class area</body>
```

Figure 1: Sample input

3852727.0006	Y
3852727.0007	Y
3915113.0000	N
3915113.0001	N
3915113.0002	N

Figure 2: Sample labels

SVM

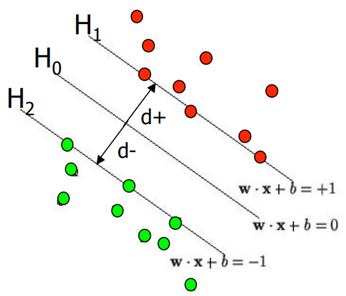


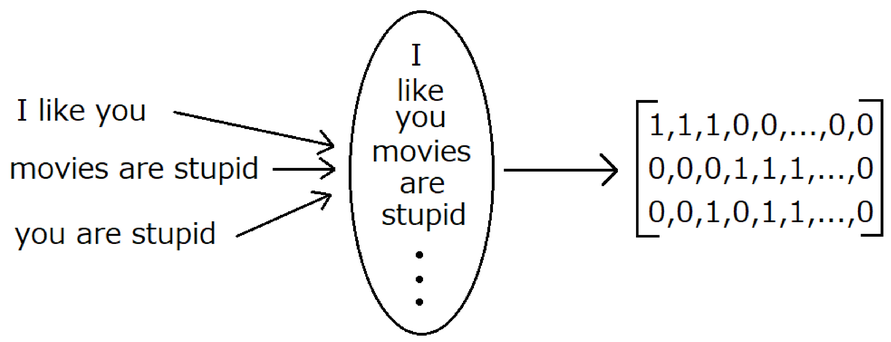
Figure 3: Graphic courtesy of MIT
http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf

Classifies data by constructing a hyperplane with the largest margin. Margin is the closest distance between plane and nearest points to it on either side (known as support vectors). Margin $\propto ||w||^{-1}$, which leads to quadratic optimization:

$$\begin{aligned} &\text{minimize} && \frac{1}{2}||w||^2 \\ &\text{s.t.} && y_i(w \cdot x_i + b) = 1 \end{aligned}$$

Much more detailed information can be found here.[2]

Feature Extraction



Feature Description

1. Feature size is number of unique words in all dialogs analyzed.
2. Each XML file transformed into a vector of length Feature_Size
3. 80% of training data (80% of all data) trains the SVM
4. SVM tested on 20% of training data as dev set
5. Final model tested on last 20% of data as test set

Test Results

Several different grams were tested before settling for the 1-word gram.

Gram Type	SVC Dev Accuracy	NuSVC Dev Accuracy	NuSVC test Accuracy
2-char	46%	97%	N/A
3-char	51%	96%	N/A
4-char	46%	97%	N/A
1-word	70%	99%	95%
2-word	43%	97%	N/A

Adaptive Re-learning

1. In simulator, "You are dumb" did not register as bullying (due to limited data)
2. Function added for users to label data on the fly (would be done by admins in the real world)
3. After conversation was done, SVM would update its dictionary and retrain with new data.
4. Results: SVM learned "You are dumb".

Analysis

1. Plain SVC did poorly on all grams, but was much improved with a Nu term of 5%
2. Nu controls how much data is forced to be support vectors. High Nu = high chance of overfitting. 5% deemed acceptably low enough.
3. Ambiguity: "I think you stupid" and "you think I stupid" output the same feature vector. This was not accounted for; bullying is occurring one way or another in each conversation. In the future, the machine could be trained to distinguish victim from aggressor.

Future Work

The first thing I would do would be to get more data, given the limited size of my set. Second, it would be interesting to run NLP sentiment analysis on the data to see if more nuanced meaning could be extracted. Perhaps the feature set could consist of only counts of emotions present, with a few identifier words. That would greatly reduce the dimensionality of the problem.

References

[1] *Cyberbullying: How Bullies Have Moved From the Playground to the Web.* OnlineCollege.org. 2012

[2] Ng, A. *Support Vector Machines.* http://cs229.stanford.edu/notes/cs229-notes3.pdf