Lab #7: Tweaking Classifiers

Everything Data
CompSci 290.01 Spring 2014
Announcements (Tue. Feb. 25)

• **Project team formation** due today!
  – Submit `proj-team.txt` to `proj-team`
  – Don’t confuse it with submitting `team.txt` for this lab
Seat assignment

Front of D106

Course staff

A

B

C

D

E

F

G

H

I

J

K

L

M

N

Back of D106
Format of this lab

• Discussion of HW #7
• Introduction to Lab #7
• Lab #7
  – Team challenge: win prizes and extra credits!
• Discussion of Lab #7
HW #7, Part 1:

Clustering senators by tweets

• Single vs. complete vs. average linkage

• Word count vs. TF-IDF

• Cosine vs. Euclidean
  – No clear winner

Overall not impressive
HW #7, Part 2:  
Predicting party by tweets

<table>
<thead>
<tr>
<th>Classifier</th>
<th>TF-IDF</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>Train</td>
<td>78.33%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>55.00%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>Train</td>
<td>97.67%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>80.00%</td>
</tr>
<tr>
<td>SVM</td>
<td>Train</td>
<td>55.00%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>55.00%</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>Train</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>79.00%</td>
</tr>
<tr>
<td>kNN (k = 4)</td>
<td>Test</td>
<td>88.00%</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>78.00%</td>
</tr>
</tbody>
</table>

• 89.00% is pretty impressive!  
• Count works better TF-IDF  
  – Making features binary can give more improvement  
  – Why?
HW #7, Part 3:

Votes lend a helping hand

• Record linkage was easy
  – SQL to get person_id’s; match first & last names
• Simple representation of votes as features
  – One column per vote
    • 2 = positive, 0 = negative, 1 = other
  – SQL to get all votes for a given person_id
    • Record vote in the correct column!

Better results—even better if we look at votes alone!

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Tweets</th>
<th>+Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve Bayes</td>
<td>Train</td>
<td>100.00%</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>89.00%</td>
</tr>
</tbody>
</table>
Introducing Lab #7

So you are not happy with your classifier (or any prediction algorithm in general); what can you do?

• *Try a different algorithm?*
• *Try different parameters of the algorithm?*
• *Get more training examples?*
• *Try fewer features?*
• *Try more features?*
Tuning algorithm parameters

• Systematic search of parameter space
  – You don’t get to see test data, yet
  – Use cross-validation on training data

• Understanding of how algorithms (and parameters) work will help
  – E.g., if you observe overfitting, try increasing $k$ in the $k$NN classifier
Scenario 1

If you increase # training examples and see
• Test error continues to decrease
• Gap between test and training errors remains big

More training examples will help
Possible overfitting, consider fewer features

From Andrew Ng: http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf
Scenario 2

But if

- Even training error is unacceptably high
- Gap between test and training error is narrow

Try additional and/or alternative features

From Andrew Ng: http://see.stanford.edu/materials/aimlcs229/ML-advice.pdf
Feature selection

One simple approach: rank all features by some utility measure, and use only the top $k$

- A popular utility measure is $\chi^2$
  - A high $\chi^2$ means it’s unlikely that the feature value and the class label are independent

- When does this fail?
Team challenge

• To get checked off: beat f1 score of 0.65
• 10% extra credit: beat 0.79
• 20% extra credit: get the highest f1 score

Prizes for the highest f1 score, and the first team to beat 0.79
Lessons learned

• Getting started should be easy; getting really good results is hard
• So many tools and knobs, so little time!
  – Automatic searches through feature and parameter spaces can help
  – Better understanding of the tools/knobs helps
• “Careful design” vs. “build-and-fix”
  – Andrew Ng
Finally

• Remember to submit team.txt under lab07 by midnight
• And also proj-team.txt under proj-team
• Sample solutions to Homework #7 and Lab #7 will be posted by tonight