Introduction to NLP

Evaluation of Classification
Evaluation of text classification

• Microaveraging
  – average over classes

• Macroaveraging
  – uses pooled table
Well-known Datasets

- **20 newsgroups**

- **Reuters-21578**
  - Cats: grain, acquisitions, corn, crude, wheat, trade...

- **WebKB**
  - [http://www-2.cs.cmu.edu/~webkb/](http://www-2.cs.cmu.edu/~webkb/)
  - course, student, faculty, staff, project, dept, other

- **RCV1**
  - [http://www.daviddlewis.com/resources/testcollections/rcv1/](http://www.daviddlewis.com/resources/testcollections/rcv1/)
  - Larger Reuters corpus
**The 2-by-2 contingency table**

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
## Precision and Recall

- **Precision**: % of selected items that are correct
- **Recall**: % of correct items that are selected

<table>
<thead>
<tr>
<th></th>
<th>relevant</th>
<th>not relevant</th>
</tr>
</thead>
<tbody>
<tr>
<td>selected</td>
<td>tp</td>
<td>fp</td>
</tr>
<tr>
<td>not selected</td>
<td>fn</td>
<td>tn</td>
</tr>
</tbody>
</table>
Pick a card...
A combined measure: \( F \)

- A combined measure that assesses the P/R tradeoff is \( F \) measure (weighted harmonic mean):

\[
F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2P + R}
\]

- The harmonic mean is a very conservative average; see \( IIR \) § 8.3

- People usually use balanced \( F1 \) measure
  - i.e., with \( \beta = 1 \) (that is, \( \alpha = \frac{1}{2} \)):
    \[
    F = \frac{2PR}{P+R}
    \]
**Classic Reuters-21578 Data Set**

- Most (over)used data set, 21,578 docs (each 90 types, 200 tokens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
  - An article can be in more than one category
  - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

**Common categories (#train, #test)**

- Earn (2877, 1087)
- Acquisitions (1650, 179)
- Money-fx (538, 179)
- Grain (433, 149)
- Crude (389, 189)

- Trade (369,119)
- Interest (347, 131)
- Ship (197, 89)
- Wheat (212, 71)
- Corn (182, 56)
Confusion matrix $c$

For each pair of classes $<c_1, c_2>$ how many documents from $c_1$ were incorrectly assigned to $c_2$?

$c_{3,2}$: 90 wheat documents incorrectly assigned to poultry

<table>
<thead>
<tr>
<th>Docs in test set</th>
<th>Assigned UK</th>
<th>Assigned poultry</th>
<th>Assigned wheat</th>
<th>Assigned coffee</th>
<th>Assigned interest</th>
<th>Assigned trade</th>
</tr>
</thead>
<tbody>
<tr>
<td>True UK</td>
<td>95</td>
<td>1</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>True poultry</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True wheat</td>
<td>10</td>
<td>90</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>True coffee</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>True interest</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>13</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>True trade</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>14</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>
Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?

  - **Macroaveraging**: Compute performance for each class, then average.
  - **Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.
Which Classifier Works the Best?

• SVM gives the best performance
• Discriminative approaches tend to be more effective than generative approaches, but in general, the difference between different classifiers is not so significant as that between different feature extraction methods
• Need to consider other factors (e.g., efficiency, interpretability)
NLP