IR Coursework (2)

Dell Zhang
Birkbeck, University of London
2016/17

Part 2 of the IR coursework is worth 10 marks.

1. (4 marks)

Train two models, multinomial Naïve Bayes and binarized Naïve Bayes, both with Laplace smoothing, on the following document counts for key sentiment words, with positive or negative class assigned as noted.

<table>
<thead>
<tr>
<th>doc</th>
<th>good</th>
<th>poor</th>
<th>great</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>pos</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>pos</td>
</tr>
<tr>
<td>$d_3$</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>neg</td>
</tr>
<tr>
<td>$d_4$</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>neg</td>
</tr>
<tr>
<td>$d_5$</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>neg</td>
</tr>
</tbody>
</table>

Use both models to assign a class (pos or neg) to this sentence $d_6$:

A good, good plot and great characters, but poor acting.

Do the two models agree or disagree?

2. (2 marks)

Assume a situation where every document in the test collection has been assigned exactly one class, and that a classifier also assigns exactly one class to each document. This setup is called one-of classification.

(a) In one-of classification, is the total number of false positive decisions always equal to the total number of false negative decisions? Please briefly explain your answer.

(b) In one-of classification, is micro-averaged precision always equal to micro-averaged recall? Please briefly explain your answer.
(c) In one-of classification, is micro-averaged $F_1$ always equal to accuracy?
   Please briefly explain your answer.

(d) In one-of classification, is micro-averaged $F_1$ always equal to macro-averaged $F_1$?
   Please briefly explain your answer.

3. (4 marks)

Consider the following 6 points in a two-dimensional vector space representing a small document collection:

- $a = (0.6, 1.9)$, $b = (1.8, 1.6)$, $c = (2.7, 2.0)$,
- $d = (3.0, 2.1)$, $e = (3.0, 2.6)$, $f = (4.2, 2.7)$.

Compute the clustering of those points (based on the Euclidean distance) using both single-link and complete-link HAC algorithms. Show the clustering results as dendrograms.