An Example of Text Classification with $k$NN ($k=1$)

Dell Zhang
17/11/2006

Two Classes: ham and spam

Training Data
ham $d_1$: “Shipment of gold arrived in a truck.”
spam $d_2$: “Shipment of gold damaged in a fire.”

Test Data
// Term IDF Weights
The number of documents in the collection $n = 3$.

$idf_a = \log(n / df_a) = \log(3 / 3) = 0$
$idf_{arrived} = \log(n / df_{arrived}) = \log(3 / 2) = 0.18$
$idf_{damaged} = \log(n / df_{damaged}) = \log(3 / 1) = 0.48$
$idf_{delivery} = \log(n / df_{delivery}) = \log(3 / 1) = 0.48$
$idf_{fire} = \log(n / df_{fire}) = \log(3 / 1) = 0.48$
$idf_{gold} = \log(n / df_{gold}) = \log(3 / 2) = 0.18$
$idf_{in} = \log(n / df_{in}) = \log(3 / 3) = 0$
$idf_{of} = \log(n / df_{of}) = \log(3 / 3) = 0$
$idf_{shipment} = \log(n / df_{shipment}) = \log(3 / 2) = 0.18$
$idf_{silver} = \log(n / df_{silver}) = \log(3 / 1) = 0.48$
$idf_{truck} = \log(n / df_{truck}) = \log(3 / 2) = 0.18$

// TF×IDF Document Vectors
$w_{i,j} = tf_{i,j} \times idf_i$

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>arrived</th>
<th>damaged</th>
<th>delivery</th>
<th>fire</th>
<th>gold</th>
<th>in</th>
<th>of</th>
<th>shipment</th>
<th>silver</th>
<th>truck</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>0</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>0.18</td>
</tr>
<tr>
<td>$d_2$</td>
<td>0</td>
<td>0</td>
<td>0.48</td>
<td>0</td>
<td>0.48</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$d_3$</td>
<td>0</td>
<td>0.18</td>
<td>0</td>
<td>0.48</td>
<td>0</td>
<td>0.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.96</td>
<td>0.18</td>
</tr>
</tbody>
</table>

// Document Vector Length
$|\vec{d}_j| = \sqrt{\sum_{i=1}^{m} w_{i,j}^2}$

$|\vec{d}_1| = \sqrt{0.18^2 + 0.18^2 + 0.18^2 + 0.18^2} = 0.36$

$|\vec{d}_2| = \sqrt{0.48^2 + 0.48^2 + 0.18^2 + 0.18^2} = 0.72$

$|\vec{d}_3| = \sqrt{0.18^2 + 0.48^2 + 0.96^2 + 0.18^2} = 1.10$

// Document Cosine Similarities
$sim(d_j, d_k) = \frac{\vec{d}_j \cdot \vec{d}_k}{|\vec{d}_j| \cdot |\vec{d}_k|} = \frac{\sum_{i=1}^{m} w_{i,j}w_{i,k}}{|\vec{d}_j| \cdot |\vec{d}_k|}$
\[ \text{sim}(d_3, d_1) = \frac{\sum_{i=1}^{11} W_{i,3} W_{i,1}}{\|\vec{d}_3\| \cdot \|\vec{d}_1\|} \]

\[ = \frac{0 \times 0 + 0.18 \times 0.18 + 0 \times 0 + 0.48 \times 0 + 0 \times 0 + 0 \times 0.18 + 0 \times 0 + 0 \times 0.18 + 0.96 \times 0 + 0.18 \times 0.18}{1.10 \times 0.36} \]

\[ = \frac{0.18 \times 0.18 + 0.18 \times 0.18}{1.10 \times 0.36} = 0.16 > 0 \]

\[ \text{sim}(d_3, d_2) = \frac{\sum_{i=1}^{11} W_{i,3} W_{i,2}}{\|\vec{d}_3\| \cdot \|\vec{d}_2\|} \]

\[ = \frac{0 \times 0 + 0.18 \times 0 + 0 \times 0.48 + 0.48 \times 0 + 0 \times 0.48 + 0 \times 0.18 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0.18 + 0.96 \times 0 + 0.18 \times 0}{1.10 \times 0.72} \]

\[ = \frac{0}{1.10 \times 0.72} = 0 \]

// kNN Classification (k=1)

The nearest neighbour of \( d_3 \) is \( d_1 \) because \( \text{sim}(d_3, d_1) > \text{sim}(d_3, d_2) \).
Since the class of \( d_1 \) is \textit{ham}, \( d_3 \) should be classified into the \textit{ham} class.