SNLP 2009
Exercise 2

• Use set theory and the axioms defining a probability function to show that:
  \[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]

• Suppose we are interested in a test to detect a disease which affects one in 100000 people on average. A lab has developed a test which works but is not perfect. If a person has the disease it will give a positive result with probability 0.97; if they do not, the test will be positive with probability 0.007. You took the test, and it gave a positive result. What is the probability that you actually have the disease?

• Are X and Y as defined in the following table independently distributed? How did you check?

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>0</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>p(X = x, Y = y)</td>
<td>0.32</td>
<td>0.08</td>
<td>0.48</td>
<td>0.12</td>
</tr>
</tbody>
</table>

• Cross-entropy between two discrete probability distributions \( p \) and \( q \) is defined as:
  \[ H(p, q) = H(p) + D_{KL}(p||q) \]
  where \( H(p) \) is entropy of \( p \) and \( D_{KL}(p||q) \) is the Kullback-Leibler divergence between \( p \) and \( q \). Show that this definition leads to the following formulation:
  \[ H(p, q) = E_p [-\log q(\cdot)] \]

• Use the frequency counts from exercise 1. Compute the entropy of the empirical frequency rank distributions for the three versions of the text you chose.

Please send your solutions to snlp@lsv.uni-saarland.de by Friday, 7 May.

Important: Please use PDF as a document format. If you need to compress files, use ZIP or GZIP.