How a robot can use basic probability theory to classify things:

The basics of statistical classification theory
We are going to classify test into topics. We assume a random model for generating documents.

- Given the topic the words are independent.
- We do not consider the order the words appear in sentence nor how often they appear. Only if they appear or not.
Text-classification: the word $\times$ document matrix

We consider the document $\times$ word matrix $X$ which records which words appear in which document:

<table>
<thead>
<tr>
<th></th>
<th>dog</th>
<th>cat</th>
<th>car</th>
<th>home</th>
<th>run</th>
<th>be</th>
<th>nice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Document1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Document2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Document3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Document4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Document5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Document6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
we consider two topics:

\[ T_0 = \text{animal life}, T_1 = \text{home and car} \]
We assume independence of words given topic

First document:
\[
\text{document 1} = (\text{dog}, \text{run}, \text{home})
\]

has probability under topic 0:
\[
P(\text{dog, run, home} | \text{topic 0}) = P(\text{dog} | \text{topic 0}) \cdot P(\text{run} | \text{topic 0}) \cdot P(\text{home} | \text{topic 0})
\]

and under topic 1:
\[
P(\text{dog, run, home} | \text{topic 1}) = P(\text{dog} | \text{topic 1}) \cdot P(\text{run} | \text{topic 1}) \cdot P(\text{home} | \text{topic 1})
\]
Conditional probabilities of words given topic

Let the conditional probabilities under topic 0 be:

\[\bar{p} = (P(dog|T0), P(cat|T0), P(car|T0), P(home|T0), P(run|T0), P(be|T0), P(nice|T0))\]

Let the conditional probabilities under topic 1 be:

\[\bar{q}(P(dog|T1), P(cat|T1), P(car|T1), P(home|T1), P(run|T1), P(be|T1), P(nice|T1))\]
Bayse classification

Given document \((\text{dog}, \text{run}, \text{cat})\) compare the probabilities and chose the one with higher probability as the topic. IF

\[
\pi_0 \cdot P(\text{dog}|T0) \cdot P(\text{run}|T0) \cdot P(\text{cat}|T0) > \pi_1 \cdot P(\text{dog}|T1) \cdot P(\text{run}|T1) \cdot P(\text{cat}|T1)
\]

we classify as topic 0. Here \(\pi_0\) is probability of topic 0 and \(\pi_1\) is probability of topic 1.
Example of estimation of condition probabilities

<table>
<thead>
<tr>
<th>Topic</th>
<th></th>
<th>dog</th>
<th>cat</th>
<th>car</th>
<th>home</th>
<th>run</th>
<th>be</th>
<th>nice</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T0$</td>
<td>Document1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$T0$</td>
<td>Document2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$T0$</td>
<td>Document3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$T1$</td>
<td>Document4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$T1$</td>
<td>Document5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$T1$</td>
<td>Document6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Estimated conditional probabilities given $T0$:

$$\hat{p} = \left( \frac{2}{3}, \frac{1}{3}, 0, \frac{2}{3}, \frac{1}{3}, \frac{2}{3}, 0 \right)$$

Estimated conditional probabilities given $T1$:

$$\hat{q} = \left( 0, 0, 1, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}, \frac{2}{3} \right)$$
Summary text classification naive bayse

- Estimate the conditional probabilities $p$ and $q$.
- Use the estimates $\hat{p}$ and $\hat{q}$ to estimate the probabilities of a new document which you wish to classify.
- Choose the topic classification which has higher estimated probability.