Homework 1

Congcheng Yan (cy2550)

February 18, 2020

Problem 1

1.

$$P(spam) = 3/5 = 0.6$$

$$P(ham) = 2/5 = 0.4$$

2.

$$P(word|class) = \frac{count(word \quad in \quad class)}{count(class)}$$

Using this function, we can calculate all the conditional probabilities.

$P(buy\|spam)$	1/12
$P(car\|spam)$	1/12
$P(Nigeria\|spam)$	1/6
$P(profit\|spam)$	1/6
$P(money\|spam)$	1/12
$P(home\|spam)$	1/12
$P(bank\|spam)$	1/6
$P(check\|spam)$	1/12
$P(wire\|spam)$	1/12
$P(money\|ham)$	1/7
$P(bank\|ham)$	1/7
$P(home\|ham)$	2/7
$P(car\|ham)$	1/7
$P(Nigeria\|ham)$	1/7
$P(fly\|ham)$	1/7

3.

(a)

$$P(Nigeria, spam) = P(spam) * P(Nigeria|spam) = 0.1$$

$$P(Nigeria, ham) = P(ham) * P(Nigeria|ham) = 0.057$$

so the label of Nigeria is spam

(b)

$$P(Nigeria, home, spam) = P(spam) * P(Nigeria|spam) * P(home|spam) = 0.00833$$

$$P(Nigeria, home, ham) = P(ham) * P(Nigeria|ham) * P(home|ham) = 0.0163$$

so the label of Nigeria home is ham

(c)

$$P(home, bank, money, spam) = P(spam)*P(home|spam)*P(bank|spam)*P(money|spam) = 0.00694$$

$P(home, bank, money, ham) = P(ham) * P(home|ham) * P(bank|ham) * P(money|ham) = 0.00233$

so the label of home bank money is ham

Problem 2

Use induction to prove it.
Base case It is obvious that when $n = 1$, $\sum_{w_1} P(w_1) = \sum_{w_1} P(w_1|start) = 1$
Induction hypothesis Assume we have

$$\sum_{w_1,w_2,\ldots w_{n-1}} P\left(w_1, w_2, \ldots, w_{n-1}\right) = \sum_{w_1,w2,\ldots w_{n-1}} P\left(w_1|start\right) \cdot P\left(w_2|w1\right) \cdots P\left(w_{n-1}|w_{n-2}\right) = 1$$

Then it is also correct below:

$$\sum_{w_1,w_2,\ldots w_n} P\left(w_1, w_2, \ldots, w_n\right) = \sum_{w_1,w_2,\ldots w_n} P\left(w_1|start\right) \cdot P\left(w_2|w_1\right) \cdots P\left(w_n|w_{n-1}\right) = 1$$

Induction step Given $\sum_{w_1,w_2,\ldots w_{n-1}} P\left(w_1, w_2, \ldots, w_{n-1}\right) = \sum_{w_1,w_2,\ldots w_{n-1}} P\left(w_1|start\right) \cdot P\left(w_2|w_1\right) \cdots P\left(w_{n-1}|w_{n-2}\right) = 1$

$$\begin{aligned}
\sum_{w_1,w_2,\ldots w_n} P\left(w_1, w_2, \ldots, w_n\right) &= [\sum_{w_1,w_2,\ldots w_n} P\left(w_n|w_1, w_2, \ldots, w_{n-1}\right)][\sum_{w_1,w_2,\ldots w_{n-1}} P\left(w_1, w_2, \ldots, w_{n-1}\right)] \\
&= [\sum_{w_{n-1},w_n} P\left(w_n|w_{n-1}\right)][\sum_{w_1,w_2,\ldots w_{n-1}} P\left(w_1|start\right) \cdot P\left(w_2|w_1\right) \cdots P\left(w_{n-1}|w_{n-2}\right)]\text{by} \\
&= \sum_{w_1,w_2,\ldots w_n} P\left(w_1|start\right) \cdot P\left(w_2|w_1\right) \cdots P\left(w_n|w_{n-1}\right) \\
&= 1
\end{aligned}$$