HOMEWORK -1 by AMOGH MISHRA (am5323)

Problem1

Ques 1: Estimate the prior probability for a random email to be spam or ham if we don't know anything about its content, i.e. P(Class)?

Ans 1: P(spam) = Total number of spam class instances / Total number of instances of all class

P(spam) = 3 / 5

Similarly, P(Ham) = 2 / 5

Ques 2: Estimate the conditional probability distributions for each word given the class, i.e. P(Word | Class). You can write down these distributions in a table.

Ans 2:

Word	P(word|Spam)	P(word|Ham)
buy	1/3	0
car	1/3	1/2
Nigeria	2/3	1/2
profit	2/3	0
money	1/3	1/2
home	1/3	1
bank	2/3	1/2
check	1/3	0
wire	1/3	0
fly	0	1/2

Ques 3:

i) P(Spam | Nigeria) and P(Ham | Nigeria)

Throughout the solution, I have removed the denominator from the Baye's theorem as it doesn't affect the result.

Ans i) P(Spam | Nigeria) = P(Nigeria | Spam) * P(Spam) = 2/3x3/5= 2/5

P(Ham | Nigeria) = P(Nigeria | Ham) * P(Ham) = 1/2x2/5 = 1/5

As P(Spam | Nigeria) > P(Ham | Nigeria), the output class is Spam.

Ans (ii) P(Spam | Nigeria Home) = P(Nigeria | Spam) * P(home | Spam) * P(Spam) =

= 2/3x1/3x3/5 = 0.13

P(Ham | Nigeria Home) = P(Nigeria | Ham) * P(home | Ham) * P(Ham) = 1/2x1x2/5 = 0.2
As P(Ham | Nigeria Home) > P(Spam | Nigeria Home), the output class is Ham.

Ans (iii) P(Spam | home bank money) = P(home|Spam) * P(bank|Spam) * P(money|Spam) * P(Spam)

= 1/3x2/3x1/3X3/5 = 2/45 = 0.044

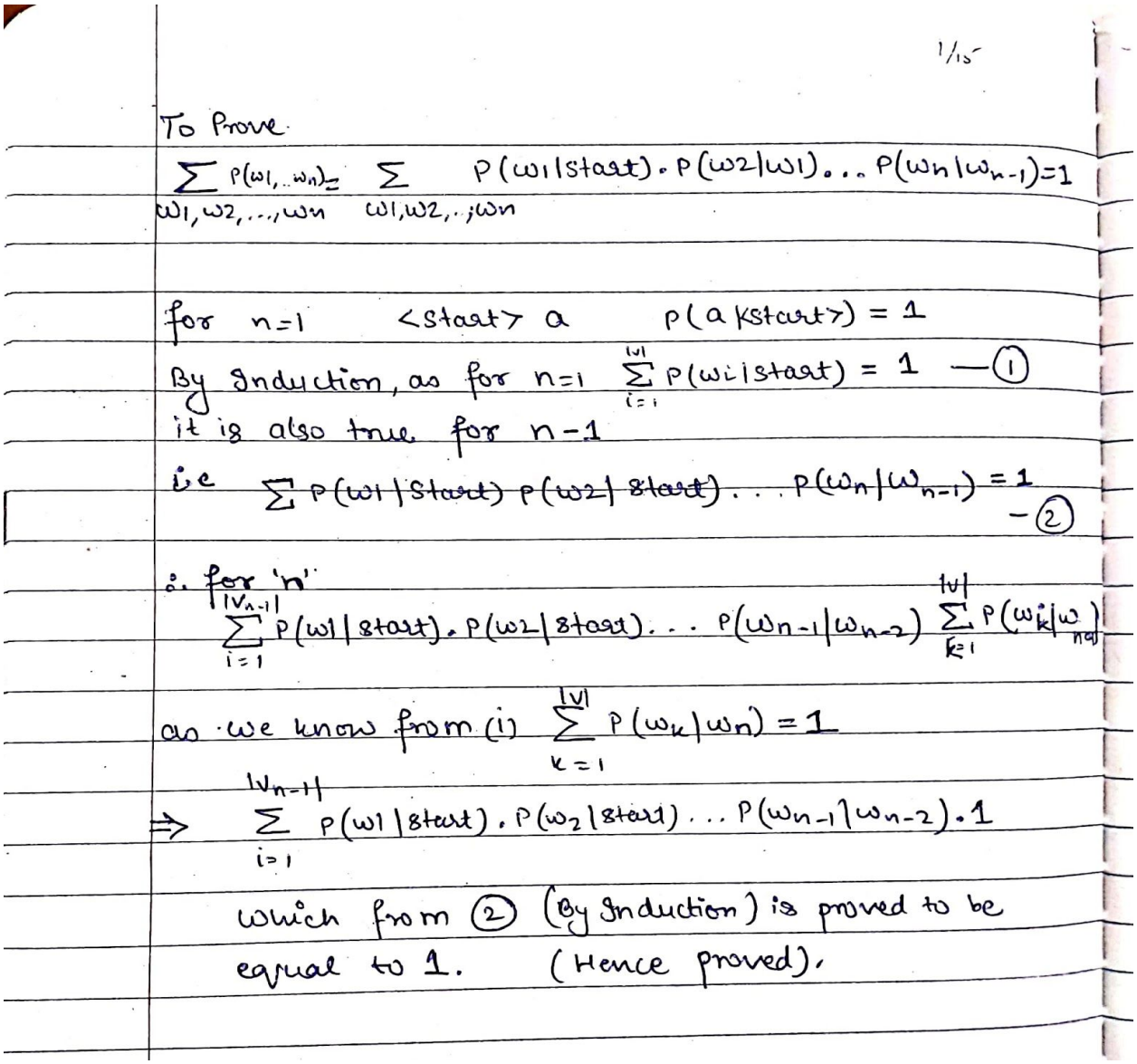
P(Ham | home bank money) = P(home|Ham) * P(bank|Ham) * P(money|Ham) * P(Ham)

= 1x1/2x1/2X2/5 = 1/10 = 0.10

As P(Ham | home bank money)>P(Spam | home bank money), the output class is Ham.

Problem 2:

To prove:



1/15

To Prove

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

for $n=1$ $\quad <start> \; a \quad P(a|<start>) = 1$

By Induction, as for $n=1$ $\sum_{i=1}^{|V|} P(w_i|start) = 1$ — (1)

it is also true for $n-1$

i.e $\sum P(w_1|start)\, P(w_2|start) \ldots P(w_n|w_{n-1}) = 1$ — (2)

∴ for 'n'

$$\sum_{i=1}^{|V_{n-1}|} P(w_1|start) \cdot P(w_2|start) \ldots P(w_{n-1}|w_{n-2}) \sum_{k=1}^{|V|} P(w_k|w_{n-1})$$

as we know from (i) $\sum_{k=1}^{|V|} P(w_k|w_n) = 1$

$$\Rightarrow \sum_{i=1}^{|V_{n-1}|} P(w_1|start) \cdot P(w_2|start) \ldots P(w_{n-1}|w_{n-2}) \cdot 1$$

which from (2) (By Induction) is proved to be equal to 1. (Hence proved).