COMP 6721 - Artificial Intelligence
Naïve Bayes Classification

Solutions

Question 1 Assume that Cecilia receives many e-mails from her home town in Klinga, where people speak Klinish. If you do not know Klinish, don't worry. It is a simple language made up of only 1,000 words that all start with the letter "k". A Klinish document may also contain words that do not start with "k", but these are considered out-of-vocabulary words (like a proper name, for example). Jack is trying to help Cecilia sort her Inbox into 3 mail folders (Personal, Work and Promotion). However, Jack does not speak Klinish, so all he has to work from are old e-mails that Cecilia has already sorted into the right folders. The table below shows a sample of the data that Jack has gathered from Cecilia's previous e-mails. The table indicates the frequency of each Klinish word in each folder (to be complete, the table should contain 1,000 rows, corresponding to each word in Klinish). For example, the word kiki appeared 30 times in e-mails labelled Personal, 50 times in e-mails about Work,...

		Folder		
		Personal	Work	Promotion
Word	kami	45	12	17
	kawa	78	1	67
	keke	0	5	80
	kiki	30	50	9
	koko	6	10	10
	kotuku	5	27	20
	koula	17	56	3
	...			
Total Nb of Words		20,000	25,000	17,000

The table above corresponds to data collected from 50 e-mails labeled *Personal*, 65 e-mails labeled *Work* and 45 e-mails labeled *Promotion*.

Based on the data above, Jack is trying to classify the following two e-mails (note that upper and lower cases should not be distinguished).

Email 1:	Koko kami kawa koula keke
Email 2:	Keke kawa, koko Google koula keke!

(a) Use a Naive Bayes classifier without any smoothing, to classify the two e-mails above. Use the sum of logs (base 10), and show the score of each of the 3 classes (Personal, Work and Promotion) and the most likely class.
Solution:
priors:
P(Personal) = 50 /50 + 65 + 45
P(Work) = 65 /50 + 65 + 45
P(Promotion) = 45 /50 + 65 + 45

Email 1: Koko kami kawa koula keke

$score(Personal) =$
$log(P(personal)) + log(P(koko|personal)) + log(P(kami|personal)) +$
$log(P(kawa|personal)) + log(P(koula|personal)) + log(P(keke|personal))$
$= log(50/160) + log(6/20,000) + log(45/20,000) + log(78/20,000) + log(17/20,000) +$
$log(0/20,000)$
$= -\infty$

$score(work) = log(P(work)) + log(P(koko|work)) + log(P(kami|work)) +$
$log(P(kawa|work)) + log(P(koula|work)) + log(P(keke|work))$
$= log(65/160) + log(10/25,000) + log(12/25,000) + log(1/25,000) + log(56/25,000) +$
$log(5/25,000)$
$= -17.8546$

$score(promotion) = log(P(promotion)) + log(P(koko|promotion)) +$
$log(P(kami|promotion)) + log(P(kawa|promotion)) + log(P(koula|promotion)) +$
$log(P(keke|promotion))$
$= log(45/160) + log(10/17,000) + log(17/17,000) + log(67/17,000) + log(3/17,000) +$
$log(80/17,000)$
$= -15.2664$

highest score is -15.2664 $\implies$ *the most likely class is promotion*

Email 2: Keke kawa, koko Google koula keke!

notes:

- ignore the word Google.
- keke counts twice

$score(Personal) = log(P(personal))+log(P(keke|personal))+log(P(kawa|personal))+$
$log(P(koko|personal)) + log(P(koula|personal)) + log(P(keke|personal))$
$= log(50/160)+log(0/2,0000)+log(78/2,0000)+log(6/20,,000)+log(17/20,000)+$
$log(0/2,0000)$
$= -\infty$

$score(work) = log(65/160)+log(5/25,000)+log(1/25,000)+log(10/25,000)+$
$log(56/25,000) + log(5/25,000)$
$= -18.2348$

$score(promotion) == log(45/160) + log(80/17,000) + log(67/17,000) +$
$log(10/17,000) + log(3/17,000) + log(80/17,000)$
$= -14.5938$

highest score is -14.5938 $\implies$ *the most likely class is promotion*

(b) Do the same as part A above, but this time use "add 0.5 smoothing" (i.e. instead of adding the value 1 to each word frequency, add $\frac{1}{2}$ to each word frequency). Adjust the smoothing formula accordingly, and show all your work. Again, use the sum of logs (base 10), and show the score of each of the 3 classes and the most likely class.

Solution:

		Folder		
		Personal	Work	Promotion
Word	kami	45.5	12.5	17.5
	kawa	78.5	1.5	67.5
	keke	0.5	5.5	80.5
	kiki	30.5	50.5	9.5
	koko	6.5	10.5	10.5
	kotuku	5.5	27.5	20.5
	koula	17.5	56.5	3.5
	...			
Total Nb of Words		20,000 +0.5 x 1,000 = 20,500	25,000 +0.5 x 1,000 = 25,500	17,000 +0.5 x 1,000 = 17,500

Email 1: Koko kami kawa koula keke

$score(personal) = log(P(personal)) + log(P(koko|personal)) + log(P(kami|personal)) + log(P(kawa|personal)) + log(P(koula|personal)) + log(P(keke|personal))$
$= log(50/160) + log(6.5/20,500) + log(45.5/20,500) + log(78.5/20,500) + log(17.5/20,500) + log(0.5/20,500)$
$= -16.7561$

$score(work) = log(P(work)) + log(P(koko|work)) + log(P(kami|work)) + log(P(kawa|work)) + log(P(koula|work)) + log(P(keke|work))$
$= log(65/160) + log(10.5/25,500) + log(12.5/25,500) + log(1.5/25,500) + log(56.5/25,500) + log(5.5/25,500)$
$= -17.6373$

$score(promotion) = log(P(promotion)) + log(P(koko|promotion)) + log(P(kami|promotion)) + log(P(kawa|promotion)) + log(P(koula|promotion)) + log(P(keke|promotion))$
$= log(45/160) + log(10.5/17,500) + log(17.5/17,500) + log(67.5/17,500) + log(3.5/17,500) + log(80.5/17,500)$
$= -15.2227$

highest score is -15.2227 $\implies$ the most likely class is promotion

		Folder		
		Personal	Work	Promotion
Word	kami	45.5	12.5	17.5
	kawa	78.5	1.5	67.5
	keke	0.5	5.5	80.5
	kiki	30.5	50.5	9.5
	koko	6.5	10.5	10.5
	kotuku	5.5	27.5	20.5
	koula	17.5	56.5	3.5
	...			
Total Nb of Words		20,500	25,500	17,500

Email 2: Keke kawa, koko Google koula keke!

notes:

- ignore the word Google.
- keke counts twice

$score(Personal) = log(P(personal))+log(P(keke|personal))+log(P(kawa|personal))+$
$log(P(koko|personal)) + log(P(koula|personal)) + log(P(keke|personal))$
$= log(50/160) + log(0.5/20,500) + log(78.5/20,500) + log(6.5/20,500) +$
$log(17.5/20,500) + log(0.5/20,500)$
$= -18.7152$

$score(work) = log(65/160)+log(5.5/25500)+log(1.5/25500)+log(10.5/25500)+$
$log(56.5/25500) + log(5.5/25500)$
$= -17.9939$

$score(promotion) = log(45/160) + log(80.5/17,500) + log(67.5/17,500) +$
$log(10.5/17,500) + log(3.5/17,500) + log(80.5/17,500)$
$= -14.5599$

highest score is -14.5599 $\implies$ the most likely class is promotion