Knowledge Management and Discovery

Problem Set- 6

Name : Shalin Patel

Class Id: 21

Ans I

A):

P(Xfrancisco=true | Class=SFO) = P (Class=SFO|Xfrancisco=true)*PXfrancisco=true)/P(Class=SFO)

= (1*1/2) / (1/2) = 1

P(Xlondon=true | Class=SFO) = P(Class=SFO|Xlondon=true)*P(Xlondon=true)/P(Class=SFO)

= 0*0*2= 0

P(Xfrancisco=true | Class=JFK) = P(Class=JFK|Xfrancisco=true)*P(Xfrancisco=true)/P(Class=JFK)

= 0

B)

P(Xfrancisco|SFO)=(4+1)/(16+8)=5/24=0.21
P(X=london|SFO)=(1+1)/(16+8)=2/24=0.09
P(Xfrancisco|JFK)=(1+1)/(16+8)=5/24=0.09

Consider a standard Naive Bayes classifier trained on the training set and applied to a similar test set. How accurate is this classifier for:

(i)the Bernoulli model : for given data we can say that Bernoulli model 100% accurate

(ii)the multinomial model: for given data we can say that multinomial model 40% accurate

II

A)

In section I, I used Bernoulli Naïve Bays model without smoothing while used multinomial with smoothing. So with the reference of that we can say that, if we use without smoothing, for we get some probability 0 and 1 directly. On other hand for multinomial, I used smoothing that why I got out other than zero for that case where conditional probability is zero.

To see why consider the worst case where none of the words in the training sample appear in the test sentence. In this case, under your model we would conclude that the sentence is impossible but it clearly exists creating a contradiction.

Another extreme example is the test sentence "Alex met Steve." where "met" appears several times in the training sample but "Alex" and "Steve" don't. Your model would conclude this statement is very likely which is not true.

B) Consider an example of when a word is repeated number of times then the denominator will be much higher and it will reduce the probability for that word. And for the word which appears less time has lower denominator, so probability is higher. This means the result we get is not correct.

III)

a)

Given, total 5 calls

3 relevant documents means recalled successfully and two are irrelevant means not recalled.

So precision for the system is 3/5= 0.6 (60%)

Its recall is simply 3/8 as it retrieves 3 correct out of 8.

b)

i)

The accuracy given is defined by c/(c+i).

Take an example for spam detecting system.

And we got below results.

	Classified positive	Classified negative
Positive class	0 (TP)	100 (FN)
Negative class	0 (FP)	100 (TN)

For this output the accuracy will be 100/200 = 50% which is higher but actually the result is total different. The actual precision of the system is 0%.

We need recall and precision because, for accuracy there is no concept false positive and false negative.

ii)

Result set for A

	Classified positive	Classified negative
Positive class	50 (TP)	50 (FN)
Negative class	50 (FP)	50 (TN)

Result set for B

	Classified positive	Classified negative
Positive class	10 (TP)	50 (FN)
Negative class	50 (FP)	10 (TN)

For above system A the Precision is = 50/(50+50)=0.5

And accuracy =(50+50)/(50+50+50+50)=50%

For above system B the Precision is = 10/(10+50)=0.17

And accuracy =(50+10)/(50+10+50+10)=50%

For system A precision is higher than B but has same accuracy.