IR Class Activity 2.

Date _____ 20___

Eisha Tir Raazia.
17k-3730
Sec : GR1.

Example data (Given data) :

DataSet	docID	Features - Words in docs.	Class Fruit Yes/No.
Training Set	1	Orange, Orange, Lemon, Red	No
	2	Orange, Red, Blue, Yellow	No
	3	Apricot, Apple, Mango.	Yes
	4	Apple, Banana, Orange	Yes
	5	Blue, Orange, Yellow	No.
Test Set.	6	Orange, Mango, Melon.	?
	7	Orange, Red, Lemon, Yellow	?

(Q1)

(a) Priors :

$P(Yes) = 2/5 = 0.6$

$P(No) = 3/5 = 0.4$.

Distinct word in train set :
Orange, Lemon, Red, Blue, Yellow, Apple, Apricot, Mango, Banana, Melon

(b) Multinomial Naive Bayes to estimate probabilities of each term (feature).

$\therefore \hat{p}(w|c) = \frac{count(w,c)+1}{count(c)+|V|}$ ($|V|$: unique/distinct vocab)

Conditional probabilities of features in data set :

- $P(Orange|Yes) = (1+1/6+9) = 0.167$.
- $P(orange|No) = (4+1/11+9) = 5/20$
- $P(Mango|Yes) = (1+1/6+9) = 2/15$
- $P(Mango|No) = (0+1/11+9) = 01/20$
- $P(Melon|Yes) = (0+1/6+10) = 1/16$
- $P(Melon|No) = (0+1/12+10) = 1/22$
- $P(Red|Yes) = (0+1/6+9) = 1/20$
- $P(Red|No) = (2+1/11+9) = 3/20$
- $P(Lemon|Yes) = (0+1/6+9) = 1/15$
- $P(Lemon|No) = (01+1/11+9) = 2/20$

- $P(\text{Yellow} \mid \text{Yes}) = (0+1/6+9) = 1/15$
- $P(\text{Yellow} \mid \text{No}) = (2+1/11+9) = 3/10$

(c) Predicting the class labels :

for doc 6 :

$P(\text{Yes} \mid \text{doc6}) = P(\text{Yes}) \times P(\text{Orang} \mid \text{Yes}) \times P(\text{Mango} \mid \text{Yes}) \times P(\text{Melon} \mid \text{Yes})$

$= 0.4 \times (0.133)^2 \times 0.058$

$\boxed{P(Y \mid d6) = 4.10 \times 10^{-4}}$

$P(\text{No} \mid \text{doc6}) = P(\text{No}) \times P(\text{O} \mid \text{No}) \times P(\text{Mango} \mid \text{No}) \times P(\text{Melon} \mid \text{No})$

$= 0.6 \times 0.25 \times (0.05)^2$

$\boxed{P(N \mid d6) = 3.75 \times 10^{-4}}$

∴ Doc6 has class label "Yes"

for doc 7 :

$P(\text{Yes} \mid \text{doc7}) = P(\text{Yes}) \times P(\text{Orange} \mid Y) \times P(\text{Red} \mid Y) \times P(\text{lemon} \mid Y) \times P(\text{Yellow} \mid Y)$

$= (0.4) \times 0.133 \times 0.05 \times 0.066^2$

$\boxed{P(Y \mid d7) = 1.58 \times 10^{-5}}$

$P(\text{No} \mid \text{doc7}) = P(\text{No}) \times P(\text{Orange} \mid \text{No}) \times P(\text{Red} \mid \text{No}) \times P(\text{Lemon} \mid \text{No}) \times P(\text{Yellow} \mid \text{No})$

$\boxed{P(N \mid d7) = 3.375 \times 10^{-4}}$

∴ Doc7 has class label "No".

(Q2)

(a) Concept drift :

Changes (in underlying data distribution) make the model, built on old data, inconsistent with new data. Therefore, regular updating of model is necessary. This problem is known as "concept drift".

⇒ Bernoulli's model is particularly robust with concept drift as in bernoulli's distribution we use fewer than dozen features which are the most important

Date _____ 20___

indicators and therefore, are less likely to change. Thus, a model that only relies of these features are more likely to maintain a certain level of accuracy in concept drift.

(b) Naive bayes classifier assumes that presence (or absence) of a particular feature in class is independent or unrelated to the presence of any other feature, given the class variable.
It's called naive because this classifier assumes strong, or naive, independence between features.

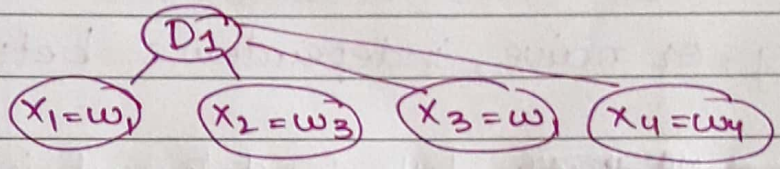
⇒ In case of NB classifier, following statement is considered true:
"Correct estimation implies accurate prediction, but accurate prediction doesn't imply correct estimation"
because,
the probability estimates of NB classifier are very low quality but classification decisions are suprisingly good.
Winning class in NB usually have much larger probability than other classes and estimates diverge very significa-ntly from true probabilities. But, classification is based on which class has highest score so despite bad estimation it makes accurate prediction.

(c) If document terms do not provide clear evidence to choose between 2 classes then we choose the one that has higher prior probability.
→ Also, if categorical has category, not present in traing set, we don't consider to provide it zero probability because it might cause problem and can reduce all terms to zero. This problem

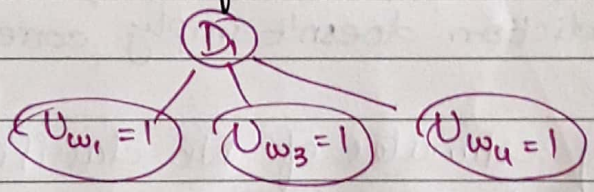
of 'zero frequency' can be solved by a soothing technique called Laplace's soothing.

- NB guarantees to perform best attempt for classificationas it's based of Baye's theorem, which help us compute the conditional probabilities of different attributes and therefore it's useful in classification task.

(d) Multinomial model for doc1

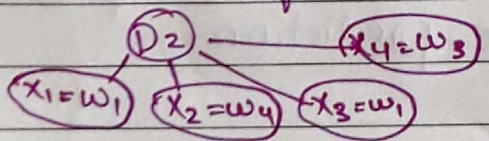


Bernouli's model for doc D3

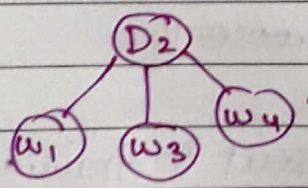


(a) The 3 docs have identical bag of word representation for the bernouli's model as it considers the presence of word regardless of frequency.

Multinomial for D_2:

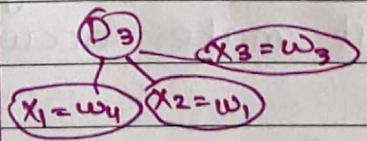


~~Bionomia~~ Bernouli of D_2:

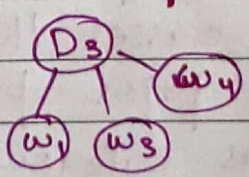


Both representations of D_2 are similar to D_1

Multinomial for D_3:



Bernouli for D_3



Different multinomial representation than D_1 and D_2

(b) The documents donot have identical bag of representation as multinomial also takes word frequency under consideration.

$\therefore d_3 = f_{w_1} = 1$

$d_1, d_2 = f_{w_1} = 2.$

so, multinomial models are different for docs.

(Q3)

(a)

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	TF $= 1+\log_2(5/df)$ d_1	d_2	d_3	d_4	d_5	d_6	d_7	IDF $\log_2(5/df)$
Yellow	0	1	0	0	1	0	1	0	1	0	0	1	0	1	1.32
Orange	2	1	0	1	1	1	1	2	1	0	1	1	1	1	0.32
Lemon	1	0	0	0	0	0	1	1	0	0	0	0	0	1	2.32
Red	1	1	0	0	0	0	1	1	1	0	0	0	0	1	1.32
Blue	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1.32
Apricot	0	0	1	0	0	0	0	0	0	1	0	0	0	0	2.32
Apple	0	0	1	1	0	0	0	0	0	1	1	0	0	0	1.32
Mango	0	0	1	0	0	1	0	0	0	1	0	0	1	0	2.32
Banana	0	0	0	1	0	0	0	0	0	0	1	0	0	0	2.32
Melon	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0

	TF * IDF d_1	d_2	d_3	d_4	d_5	d_6	d_7	$\overrightarrow{Yes}$	$\overrightarrow{NO}$
Orange	0.64	0.32	0	0.32	0.32	0.32	0.32	0.16	0.42
Lemon	2.32	0	0	0	0	0	2.32	0	0.77
Red	1.32	1.32	0	0	0	0	1.32	0	0.88
Blue	0	1.32	0	0	1.32	0	0	0	0.88
Apricot	0	0	2.32	0	0	0	0	1.16	0
Apple	0	0	1.32	1.32	0	0	0	1.32	0
Mango	0	0	2.32	0	0	2.32	0	1.16	0
Banana	0	0	0	1.32	0	0	0	1.16	0
Yellow	0	1.32	0	0	1.32	0	1.32	0	0.88
Melon	0	0	0	0	0	0	0	0	0

Table 3-1

(b) Computation of centroids

(i) $\mu_{fruit=yes} = \frac{1}{|D_{fruit=yes}|} \sum_{d \in D_{fruit=yes}} \vec{v}(d)$

$= \frac{1}{2}(\vec{d_3} + \vec{d_4})$

$= \frac{1}{2}(\langle 0,0,0,0,2.32,1.32,2.32,0,0,0\rangle + \langle 0.32,0,0,0,0,1.32,0,2.32,0,0\rangle)$

$\mu_{fruit=yes} = \langle 0.16,0,0,0,1.16,1.32,1.16,1.16,0,0\rangle$

(ii) $\mu_{fruit=no} = \frac{1}{|D_{fruit=no}|} \sum_{d \in D_{fruit=no}} \vec{v}(d)$

$= \frac{1}{3}(\vec{d_1} + \vec{d_2} + \vec{d_5})$

$= \frac{1}{3}(\langle 0.64,2.32,1.32,0,0,0,0,0,0,0\rangle + \langle 0,32,0,1.32,1.32,0,0,0,0,1.32,0\rangle + \langle 0.32,0,0,1.32,0,0,0,0,1.32,0\rangle)$

$\mu_{fruit=no.} = \langle 0.42, 0.77, 0.88, 0.88, 0, 0, 0, 0, 0.88, 0\rangle$

(c) Classification using Rocchio approach

For doc6 :

(a) for class 'Yes' :

$|\vec{\mu}_{f=yes} - \vec{d_6}| = \sqrt{(0.16-0.32)^2 + 0+0+0 + (1.16-0)^2 + (1.32-0)^2 + (1.16-2.32)^2 + (1.16-0)^2 + 0 + 0}$

$= \sqrt{0.0256 + 0 + 1.345 + 1.742 + 1.345 + 1.345}$

$= 2.403$

(b) For class 'No' :

$$|\vec{M}_{f=no} - \vec{d_6}| = \sqrt{(0.42-0.32)^2 + 0.77^2 + 0.88^2 + 0.88^2 + 2.32^2 + 0.88^2}$$

$$= \sqrt{0.01 + 0.59 + 0.77 + 0.77 + 0.1024 + 0.77}$$

$$= 2.80$$

∴ class label of doc6 is "Yes" as min (2.80, 2.403) = 2.403.

For doc7 :

(a) for class 'Yes' :

$$|\vec{M}_{f=yes} - \vec{d_7}| = \sqrt{(0.16-0.32)^2 + (2.32)^2 + (1.32)^2 + (1.16)^2 + (1.32)^2 + (1.16)^2 + 1.16^2 + 1.32^2 + 0}$$

$$= \sqrt{0.0256 + 5.38 + 1.742 + 1.34 + 1.742 + 1.34 + 1.34 + 1.742}$$

$$= 3.83$$

(b) $$|\vec{M}_{f=no} - \vec{d_7}| = \sqrt{(0.42-0.32)^2 + (0.77-2.32)^2 + (0.88-1.32)^2 + 0.88^2 + (0.88-1.32)^2 + 0}$$

$$= \sqrt{0.01 + 2.40 + 0.1936 + 0.774 + 0.1936}$$

$$= 1.9702$$

∴ class label of doc7 is "No" as min (3.83, 1.97) = 1.97.

(d) Classification using KNN. → values from tfidf table Table 3.1

(a) KNN for doc 6 :

$$\text{Cos}(d_1, d_6) = \frac{\vec{d_1} \cdot \vec{d_6}}{\|\vec{d_1}\| \|\vec{d_6}\|}$$

$$= \frac{\langle 0.64, 2.32, 1.32, 0, 0, 0, 0, 0, 0, 0 \rangle \cdot \langle 0.32, 0, 0, 0, 0, 0, 2.32, 0, 0, 0 \rangle}{2.7448 \times 2.341}$$

$$= \frac{0.2046}{6.425} = 0.0318$$

Date _____ 20___

- $\cos(d_2, d_6) = \frac{0.1024}{2.308 \times 2.3419} = 0.0189$

- $\cos(d_3, d_6) = \frac{5.382}{3.5365 \times 2.3419} = 0.6498.$

- $\cos(d_4, d_6) = \frac{0.32^2}{2.688 \times 2.3419} = 0.016.$

- $\cos(d_5, d_6) = \frac{0.32^2}{1.8939 \times 2.3419} = 0.0230.$

for 3-NN doc6 belongs to class "No".

(b) KNN (3-NN for doc7):

- $\cos(d_1, d_7) = \frac{1.74}{2.7448 \times 2.994} = 0.8916$

- $\cos(d_2, d_7) = \frac{3.5872}{3.3085 \times 2.9949} = 0.5188$

- $\cos(d_3, d_7) = \frac{0}{3.565 \times 2.9949} = 0.$

- $\cos(d_4, d_7) = \frac{0.32^2}{2.688 \times 2.9949} = 0.0127$

- $\cos(d_5, d_7) = \frac{0.32^2 + 1.32^2}{1.8939 \times 2.9949} = \frac{1.8448}{1.8939 \times 2.9949} = 0.4159.$

for 3NN doc 7 belongs to class 'No'