

College of Engineering Pune

Test)

Subject: Data Science (Div1 + Div 2)
TY-Computer Engineering

Date : 27/2/2023

Functions:

Scientific calculators are only permissible.

Written answers would not be considered for evaluation.

Time: 10 to 11 a.m.

Which of the following is/are instance(s) of lemmatization?

- a) eat -> ate
- b) conflation->conflate
- c) Studies -> Studi
- d) walked -> walk

How many different lexemes are there in the following list?
man, men, girls, girl, mouse

- a) 5
- b) 4
- c) 3
- d) 2



Select categorical type of data attributes:

- a) Nominal
- b) Ordinal
- c) Numeric
- d) Discrete



[Fill in the blanks: 4-8]

IQR is used to identify the outlier ✓

Transformation of a continuous attribute into ordinal categorical attribute is called as

In case of many outliers in the dataset, the most representative value is maximum

Matching strings that should not have matched _____ [which error]

Low value for chi-square means there is a high similarity between two sets of data.

Reasons of noisy data [Select the appropriate] :

- a) faulty data collection instruments
- b) data entry problems
- c) data transmission problems
- d) inconsistent with other recorded data
- e) technology limitation



PRESENTED		CLASS		TOTAL
	SICK	NO SICK	TYPE	
SICK	654 SICK 654 REHABILITATED	46 SICK 102 102 REHABILITATED		1000
NO SICK	412 REHABILITATED	558 REHABILITATED 162 REHABILITATED		720
TOTAL	1066	782		1848

For more:

	Year	Month	Day	Hour	Min	Sec	Season	Time	Fast
Dec1	-	-	-	-	-	-	-	-	-
Dec2	0	0	2	15	0	0	-	-	-
Dec3	1	0	0	0	0	0	-	-	-

11. Let data be in the range 10,000 to 80,000 normalized to [0.0, 1.0]. Apply min-max normalization to determine the normalized value of 50,500.

$$\text{Ans} \quad V_1^* = \frac{V_1 - P_{\text{gas}} n_R}{P_{\text{gas}} n_R - P_{\text{gas}} n_R} \left[\rho_{\text{gas},T_0} n_R - \rho_{\text{gas},T_1} n_R \right]$$

$$= \frac{\$16,560 - \$6,600}{\$10,600 - (\$6,600)} [1 - e]$$

$$V_1 = 0.6571$$

- 12 Let there be two Cricket players, Rohit and stardie, and you have to select one for the cricket world cup. The scores of both the players in the last four one day matches are as follows:

Robot	Hardik
28	42
36	73
45	2
59	17
63	63

finding mean, Q₁, Q₃, min, max
for both of them.

$$\text{mean}(rabit) = \frac{28+38+45+59+63}{5} = 46.6$$

$$\text{mean}(\text{marks}) = \frac{92 + 59 + 2 + 17 + 13}{5} = 48.6$$

from seeking diff. bel'd.

Min μ = 28 Max μ = 63

$$\min H = 2 \quad \max H = 83$$

By taking diff. betⁿ. mean & min also mean & max it is found that difference is more in case of Hardik while for Rabi diff. is quite same.

Rohit is consistent.

$$C \cdot (\text{mean}_R - \text{min}_R) = 14.6 = R_1$$

$$(-(\text{mean}_R) + \text{max}_R) = 16.4 = R_2$$

less diff. in R_1 & R_2 .

$$\text{mean}_H - \text{min}_H = 36.6 = H_1$$

$$\text{max}_A - \text{mean}_A = 44.4 = H_2$$

more diff in H, & β_2 also more

"diff. bet" mean & max
After plotting box plot can be

A box plot can be drawn i.e. Rohit is more consistent.

3. Consider the corpus C of sentences
1. there is a big garden
 2. children play in a garden
 3. they play inside beautiful garden

Calculate $P(\text{Children play in a big house})$

- i) Maximum likelihood estimation (MLE), and
- ii) Add one smoothing.

[Hint: Augment each sentence with sentence beginning and end markers]

sol from above

Using CORPUS, bigram model.

$$P(\text{play} \mid \text{children}) = \frac{1}{4} = 0.25$$

$$P(\text{in} \mid \text{play}) = \frac{1}{2} = 0.5$$

$$P(\text{big} \mid \text{in}) = \frac{1}{2} P(\text{big} \mid \text{a}) = \frac{1}{2} = 0.5$$

Here considering only one before word, i.e.
by using MLE, bigram mode.

the probability of house follows

big is 0, \therefore

$$\begin{aligned} P(\text{children play in a big house}) &= P(\text{play} \mid \text{children}) \times P(\text{in} \mid \text{play}) \times P(\text{big} \mid \text{in}) \\ &\times P(\text{big} \mid \text{a}) P(\text{house} \mid \text{big}) \\ &= 1 \times 0.5 \times 1 \times 0.5 \times 0 \\ &= 0 \end{aligned}$$

$$\begin{aligned} \therefore P(\text{children play in a big house}) &= \frac{1}{16} \times \frac{2}{13} \times \frac{1}{6} \times \frac{2}{13} \times \frac{1}{12} \\ &= \frac{1}{18252} \\ &= 0.000054783 \end{aligned}$$

$(\frac{1}{2})$

Incomplete

4
5
ore

Add-one smoothing:

$$P(w_i \mid w_1, w_2, \dots, w_{i-1}) = \frac{(w_{i-\text{big}} + 1)}{(w_{i-1}) + V}$$

$$P(\text{in} \mid \text{play}) = \frac{1+1}{1+11} = \frac{2}{12} = \frac{1}{6}$$

$$P(\text{in} \mid \text{a}) = \frac{1+1}{2+11} = \frac{2}{13}$$

$$P(\text{big} \mid \text{a}) = \frac{1+1}{1+11} = \frac{1}{6}$$

$$P(\text{house} \mid \text{big}) = \frac{0+1}{1+11} = \frac{1}{12}$$

- 14 Suppose you have the following data of time in seconds for different cyclothon runners
 59, 65, 61, 62, 53, 55, 60, 70, 64, 56, 58, 59, 62, 62, 54, 65, 56, 59, 53, 61, 67. Find median
 and estimated median values.

Sol Arranging in increasing order = 52, 53, 54, 55, 56, 56, 58, 58,
 59, 59, 60, 61, 61, 62, 62, 62, 62, 64, 65, 65, 66, 67, 67, 70.
 Here $N = 21$,

$$\text{median} = x \left[\frac{N+1}{2} \right] \\ = x \left[\frac{21+1}{2} \right] \\ = x [11]$$

$$\therefore \text{median} = \underline{\underline{60}}$$



- 15 Suppose you are fishing in a lake with 8 species (bass, carp, catfish, eel, perch, salmon, trout, whitefish) and you have seen 6 species with the following counts: 10 carp, 3 perch, 2 whitefish, 1 trout, 1 salmon, and 1 eel (so you haven't yet seen the catfish or bass). What are the Good Turing probabilities ($P_{GT}^*(\text{trout})$) for catching the next fish as "trout" or the Good Turing probabilities ($P_{GT}^*(\text{bass})$) for catching the next fish as "bass"?

Sol N_c - frequency of frequency - for bass (which is unseen)
 Here $N_1 = 3$, $N_2 = 1$, $N_3 = 1$, $N_4 = N_5 = N_6 = N_7 = N_8 = N_9 = 0$, $N_{10} = 1$

Total no. of fish found = $18 = N$

So the probability that next fish is trout = $\frac{1}{18}$
 (since term freq. of trout = 1)

Now for bass, as in 18 above - only found fishes. $\#tf_{\text{bass}} = 0$.

But $N_1 = 3$, pf_{bass}

$$\therefore P_{\text{bass}} = \frac{2-N_1}{N} = \frac{3}{18}$$

But now finding next trout i.e. $P_{GT}^*(\text{trout})$ will change

It is less than $\frac{1}{18}$.

$$P_{GT}^* = \frac{N_1}{N}$$

$$c^* = \frac{(c+1)N_{c+1}}{N_c}$$

$$c = 0 \quad c^* = \frac{N_1}{N} = \frac{3}{18}$$

$$\therefore P_{GT}^*(\text{bass}) = \frac{N_1}{N} = \frac{3}{18}$$

for seen trout,
 $c = 1$ (1 time seen)

$$c^* = \frac{2 \times 1}{3}$$

$$\therefore P_{GT}^*(\text{trout}) = \frac{\frac{2}{3} \times \frac{2}{3}}{\frac{1}{18}}$$

$$\therefore P_{GT}^*(\text{trout}) = \frac{2}{3 \times 18}$$

$$\therefore P_{GT}^*(\text{trout}) = \frac{1}{27}$$

$$\cos(d_2, d_4) = \frac{1 \times 1 + 1 \times 1 + 1 \times 1}{\sqrt{1^2+1^2+1^2} \sqrt{1^2+1^2+1^2}} \\ = \frac{3}{\sqrt{6} \cdot \sqrt{6}} = \frac{\sqrt{2}}{4} = 0.86$$

$$C(d_2, d_4) = \cos'(0.86) \approx 50 \text{ % SL}$$

$$\cos(d_3, d_4) = \frac{1 \times 1}{\sqrt{1^2+1^2+1^2} \sqrt{1^2+1^2+1^2}} = \frac{1}{\sqrt{3} \cdot \sqrt{3}} = \frac{1}{3} = 0.33$$

$$(d_3, d_4) = \cos'(0.33) \approx 7.1$$

2/2

Column 1
JItem = 0

Column 2
JItem =

Column 3
JItem =

;

- 3 Given the following data. Fill the matrix by computing the minhash signature for each column using the given hash functions. 5

Element	S1	S2	S3	S4	$2x+1 \pmod{6}$	$3x+2 \pmod{6}$	$5x+2 \pmod{6}$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

Determine signatures matrix? [Note: Show steps while updating signature matrix starting from initialization]

Justify, how close are the estimated Jaccard similarities for all pairs of columns to the true Jaccard similarities?

Signatures matrix for $2x+1 \pmod{6}$.
1st step:- We select 0th row, in which s_2 & s_4 are 1.

1	0	1
---	---	---

2nd step:- Now, select 2nd row, in which s_1 & s_4 are 1.

1	1	1
---	---	---

3rd step:- Now, select 4th row, s_3 & s_4 are 1.

1	1	3	1
---	---	---	---

Signature matrix:-
For $3x+2 \pmod{6}$.
1st step:- Select 4th row.

1	2	2
---	---	---

2nd step:- Select 5th row

5	2	2
---	---	---

3rd step:- Select 2nd row.

5	5	2	2
---	---	---	---

Signature matrix:-
For $5x+2 \pmod{6}$.
1st step:- Select 0th row.

1	1	1
---	---	---

2nd step:- Select 1st row

1	4	4
---	---	---

3rd step:- Select 3rd row

3	1	4	4
---	---	---	---

2023.04.20 14:59
Final - 13 19 4

Column 1 & 2.
 $J_{sim} = 0$

Column 2 & 3.
 $J_{sim} = 0$.

Column 3 & 4

$$J_{sim} = \frac{1}{4} = 0.25.$$

\therefore Columns 1 & 2 and Columns 2 & 3 have the Jaccard similarity. i.e. zero.

(2)

Y

(1)

[Fill in the blanks: 4-7]

- 4 Anti-monotone property is: Support of an itemset does not exceed the support of its subset. It is called anti-monotone property.
- 5 State Apriori principle: Support of an itemset is frequent, then the support of its subset is frequent. Apriori principle holds the property: $\forall X, Y : X \subseteq Y \text{ then } S(X) \geq S(Y)$.
- 6 The Prune step of Apriori algorithm is: There are 3 steps:-
1) Brute-force 2) Merge 3) Prune
- 7 If minsup is set too low then the frequent itemset will be more.

- 8 Dataset given below shows weight of students in a class. Draw box plot with whiskers for the same. Analyze the box plot.
Weight: 13, 12, 9, 11, 14, 12, 10, 20, 11, 10, 7

Sohed data: 7, 9, 10, 10, 11, 11, 12, 12, 13, 14,

$$\text{median} = 11$$

$$Q_1(1^{\text{st}} \text{ quartile}) = \text{median}(7, 9, 10, 10, 11) \\ Q_1 = 10$$

$$Q_3(3^{\text{rd}} \text{ quartile}) = \text{median}(12, 12, 13, 14, 20) \\ Q_3 = 13$$

$$IQR = Q_3 - Q_1 = 13 - 10 \\ IQR = 3$$

- 9 Investigate whether there's a relationship between the number of students and the marks scored. Which plot will help and how? Assume the data below:

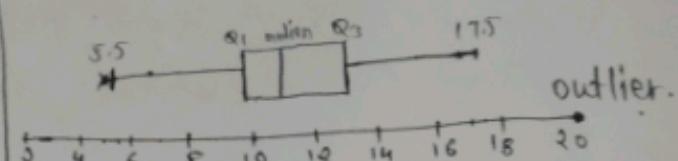
No of students	Marks Obtained	Percentage Of Students
5	40	2.5
6	60	3
25	70	12.5
11	65	5.5
30	80	15
4	50	2
6	55	3
10	75	5
14	90	7
18	45	9
20	40	10
22	95	11
2	100	1
11	35	5.5
16	25	8

Lower limit: $Q_1 - 1.5 \times IQR = 5.5$

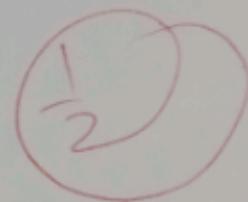
Upper limit: $Q_3 + 1.5 \times IQR = 17.5$

There's outlier, which is greater than 17.5 is 20.

Box plot:-



For the given relationship, the line plot will be helpful.



1. Scenario calculations are only permissible
2. Over written answers would not be considered for evaluation

from the

given the

1. Find the edit distances using only insertions [cost 1] and deletions [cost 1] between the following pairs of strings.

Source: abcdaabc, Target: aabcbab

	a	b	c	d	a	b	c
a	0	1	2	3	4	5	6
b	1	D	1	2	3	4	5
c	2	1	2	3	4	5	6
d	3	2	1	2	3	4	5
a	4	3	2	1	2	3	4
b	5	4	3	2	1	2	3
c	6	5	4	3	2	1	2
d	7	6	5	4	3	2	1

Minimum Cost: 5

2. Give the formula for Cosine Similarity and determine the Cosine Similarity between the given document collection D and test document D4 specifying the document closer to test document.

[Apply stop-word removal, lemmatization] [Stop words to, from, is]

Document Collection D:

D1: Jack travelled to London

D2: Jack travelled from Oakland to London

D3: Travel to Oakland is wonderful.

Test Document D4: Jack travelled to Oakland

$$\text{Cosine Similarity} = \frac{|a \cap b|}{|a \cup b|}$$

$$\cos(D_1, D_2) = \frac{a}{6} = 0.667$$

$$\cos(D_2, D_3) = \frac{2}{9} = 0.22.$$

$$\cos(D_3, D_4) = \frac{2}{9} = 0.22.$$

$$\cos(D_1, D_4) = \frac{1}{8} = 0.125$$

$$\cos(D_1, D_1) = \frac{4}{5} = 0.8$$

$$\cos(D_2, D_4) = \frac{4}{6} = 0.667$$

$$\cos(D_3, D_4) = \frac{2}{7} = 0.285$$

$$\cos(D_1, D_3) = \frac{2}{7} = 0.285$$

$$\cos(D_2, D_3) = \frac{2}{7} = 0.285$$

$$\cos(D_1, D_2) = \frac{2}{7} = 0.285$$

$$\cos(D_3, D_2) = \frac{2}{7} = 0.285$$

$$\cos(D_1, D_4) = \frac{1}{8} = 0.125$$

$$\cos(D_2, D_4) = \frac{1}{8} = 0.125$$

$$\cos(D_3, D_4) = \frac{1}{8} = 0.125$$

2023.04.20 14:59

- 3 Given the following data. Fill the matrix by computing the minhash signature for each column using the given hash functions. 5

Element	S1	S2	S3	S4	$(2x+1) \text{ mod } 6$	$(3x+2) \text{ mod } 6$	$(5x+2) \text{ mod } 6$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	4

Determine signatures matrix? [Note: Show steps while updating signature matrix starting from initialization] ✓ X

Justify, how close are the estimated Jaccard similarities for all pairs of columns to the true Jaccard similarities?

(x)

[Fill in the blanks: 4-7]

$$s(x) \geq s(y) \quad x \subset y$$

①

- 4 Anti-monotone property is: Candidate does not have more support than its subset. ✓ ②
- 5 State Apriori principle: If any set is frequent then its subset is also frequent itemset. ③
- 6 The Prune step of Apriori algorithm is: Candidate elimination ✓
used to infrequent itemsets & prevent searching for association in infrequent itemsets. ④
- 7 If minsup is set too low then nearly all candidate itemsets becomes frequent itemsets. ✓ ⑤
all candidate itemsets

⑥

all candidate itemsets

✓ 1+1+1+1

2023.04.20 14:59

- 8 Dataset given below shows weight of students in a class. Draw box plot with whiskers for the same. Analyze the box plot.

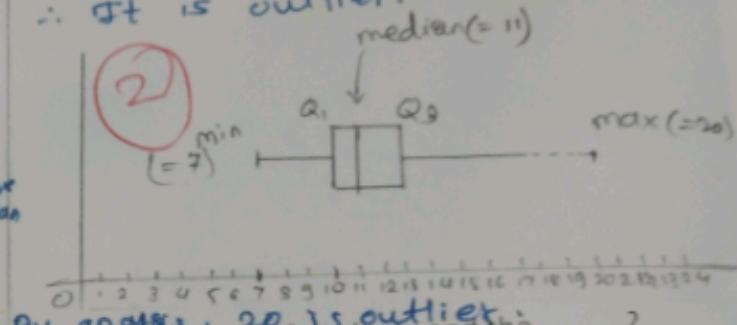
Weight: 13, 12, 9, 11, 14, 12, 10, 20, 11, 10, 7

Arranging weights in increasing order
weight: 7, 9, 10, 10, 11, 12, 12, 13, 14, 20
 $n = 11$

$$\min = 7, \max = 20, \\ \text{median} = n \left[\frac{n+1}{2} \right] \\ = n [6] \\ = 11$$

$$Q_1 = n \left[\frac{n+1}{2} \right], (n \text{ below median}) \\ = n [3] \\ Q_1 = 10 \\ Q_3 = n \left[\frac{n+1}{2} \right], \text{ above median} \\ = n [9] \\ Q_3 = 13$$

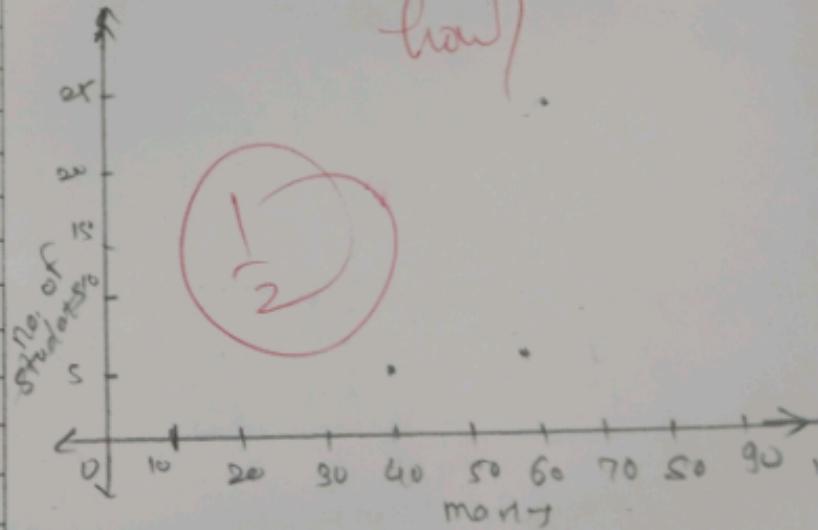
Using 5 no. summary.
Box plot: $IQR = 13 - 10 = 3$
outlier is less than $Q_1 - 1.5IQR = 5 - 5$
more than $Q_3 + 1.5IQR = 17.5$
now $\max(20)$ is more than 17.5
 \therefore It is outlier.



- 9 Investigate whether there's a relationship between the number of students and the marks scored. Which plot will help and how? Assume the data below:

No of students	Marks Obtained	Percentage Of Students
5	40	2.5
6	60	3
25	70	12.5
11	65	5.5
30	80	15
4	50	2
6	55	3
10	75	5
14	90	7
18	45	9
20	40	10
22	95	11
2	100	1
11	35	5.5
16	25	8

We need to find out relationship
b/w no. of students and
marks. \rightarrow Scatter plot.



Add-k -

$$P(\text{Add-k} \neq w_{i-1}, w_i) = \frac{c(w_{i-1}, w_i) + k}{v}$$

$$= \frac{c(w_{i-1}, w_i) + m \left(\frac{1}{v} \right)}{c(w_{i-1}) + m}$$

8) * Good-Turing smoothing -

N_c - frequency of frequency c

N_c - count of things we've seen c times

Sam I am I am Sam \neq do not eat

how many words are occurring once

I 3

$N_1 = 3$ words (distinct)

Sam 2

$N_2 = 2$

am 2

$N_3 = 1$

do 1

↑

no 1

how many words occurring thrice
(only one - I)

eat 1

IMP

next eg. suppose find 18 fish.

for trout + $y_{18} \rightarrow$ 10 carp, 3 perch, 2 trout, 1 ^{cat} salmon = 18

now new species finding probability.

so here $N_1 = 3$, (^(catfish) no. of fishes occurred only once)

$$\text{so } N_1/18 = 3/18$$

Now probability of next trout is not y_{18} (less than it)
How to find new estimate?

$$P_{GT}^* \text{ (thing with zero freq)} = \frac{N_1}{N}$$

$$C^* = \frac{(c+1)N_{c+1}}{N_c}$$

Unseen (bass or catfish) Seen once (trout)

$$c=0$$

$$c=1$$

$$\text{MLE} = p=0/18 = 0$$

$$\text{MLE} = p=1/18$$

$$P_{GT}^* = \frac{3}{18}$$

$$C^*(\text{trout}) = \frac{2 \times N_2}{N_1}$$

$$= 2 \times \frac{1}{3}$$

$$= 2/3$$

$$\text{So } P_{GT}^* (\text{trout}) = \frac{2}{3} \times \frac{1}{18} = \frac{1}{27}$$

2023.04.22 18:53

College of Engineering Pune

Test 1

Subject: Data Science (Div 1 + Div 2)
TY-Computer Engineering

Time: 10 to 11 a.m.

Date : 27/2/2023

Instructions:

1. Scientific calculators are only permissible
2. Over written answers would not be considered for evaluation.

1. Which of the following is/are instance(s) of lemmatization?

- a) eat -> ate
- b) conflation -> conflate
- c) Studies -> Studi
- d) walked -> walk

(1)

2. How many different lexemes are there in the following list?

man, men, girls, girl, mouse

- a) 5
- b) 4
- c) 3
- d) 2

(1)

3. Select categorical type of data attributes:

- a) Nominal
- b) Ordinal
- c) Numeric
- d) Discrete

(1)

[Fill in the blanks: 4-8]

4. IQR is used to identify the outliers

(1)

5. Transformation of a continuous attribute into ordinal categorical attribute is called as discretisation

(1)

6. In case of many outliers in the dataset, the most representative value is mean

X 1

7. Matching strings that should not have matched Type II [which error]

X 1

8. Low value for chi-square means there is a high similarity between two sets of data.

1

9. Reasons of noisy data [Select the appropriate]:

- a) faulty data collection instruments
- b) data entry problems
- c) data transmission problems
- d) inconsistent with other recorded data
- e) technology limitation

(0.25)

ata

- 14 Suppose you have the following data of time in seconds for different cyclothon runners:
59, 65, 61, 62, 53, 55, 60, 70, 64, 58, 58, 62, 62, 54, 65, 56, 59, 58, 52, 61, 67. Find median
and estimated median values.

Sol: data: 52 53 54 55 56 57 58 58 59 59 60 61 61 62
64 65 65 67 70 n=21 ~~n+2 =~~

since n is odd median = $x_{\frac{n+1}{2}} = \underline{\underline{60}}$

mean = $\frac{\sum x_i}{n} = \frac{1259}{21} = 59.95$

(5)
2

estimated median = median - mean ~~not m~~
= ~~60 - 59.95~~ 0.05

- 15 Suppose you are fishing in a lake with 8 species (bass, carp, catfish, eel, perch, salmon, trout, whitefish) and you have seen 6 species with the following counts: 10 carp, 3 perch, 1 whitefish, 1 trout, 1 salmon, and 1 eel (so you haven't yet seen the catfish or bass). What are the Good Turing probabilities ($P^*_{GT}(\text{trout})$) for catching the next fish as "trout" or the Good Turing probabilities ($P^*_{GT}(\text{bass})$) for catching the next fish as "bass"?

Sol:

0

11

Consider the corpus C of sentences

1. there is a big garden
2. children play in a garden
3. they play inside beautiful garden

Calculate $P(\text{children play in a big house})$ assuming a bigram language model using

- i) Maximum likelihood estimation (MLE), and
- ii) Add one smoothing.

[Hint: Augment each sentence with sentence beginning and end markers]

So: 1 <ss> there is a big garden <ss>

2 <s> children play in a garden <ss>

3 <s> they play inside beautiful garden <ss>

$P(\text{children play in a big house})$

$P(w_i | w_{i-1}) = \frac{\text{count}(w_{i-1}, w_i)}{\text{count}(w_{i-1})}$

no of variable: c.v
+2
= 8
(2SS + 1SS)

$$P = P(\text{children} | \text{ss}) \times P(\text{play} | \text{children}) \times P(\text{in} | \text{play}) \times P(\text{at} | \text{in}) \times P(\text{big} | \text{at}) \times P(\text{house} | \text{big}) \times P(\text{beautiful} | \text{house})$$

$$\times \frac{c(\text{ss}, \text{children})}{c(\text{ss})} \times \frac{c(\text{children}, \text{play})}{c(\text{play})} \times \frac{c(\text{play}, \text{in})}{c(\text{in})} \times$$

$$\frac{c(\text{in}, \text{a})}{c(\text{a})} \times \frac{c(\text{a}, \text{big})}{c(\text{big})} \times \frac{c(\text{big}, \text{house})}{c(\text{house})} \times \frac{c(\text{house}, \text{beautiful})}{c(\text{beautiful})}$$

$$P = \frac{1}{1} \times \frac{1}{2} \times \frac{1}{1} \times \frac{1}{2} \times \frac{1}{1} \times 0 \times 0 = 0$$

Probability by adding one smoothing = $\frac{c(w_i | w_{i-1}) + 1}{c(w_{i-1}) + V}$

$$P = \frac{1+1}{9} \times \frac{1+1}{2+8} \times \frac{1+1}{1+8} \times \frac{1+1}{2+8} \times \frac{1+1}{8+8} \times \frac{0+1}{0+8} \times \frac{0+1}{1+8}$$

$$\text{working} = \frac{2}{9} \times \frac{2}{10} \times \frac{2}{9} \times \frac{2}{10} \times \frac{2}{9} \times \frac{1}{8} \times \frac{1}{9}$$

$$= \frac{92}{5249800} = \frac{1}{164025} = \frac{6.0966 \times 10^{-6}}{1}$$

one
smoothing

10 Match the following

1. Term Document matrix	a)																																																
d ✓	<table border="1"> <thead> <tr> <th colspan="2">PREDICTED</th> <th colspan="2">CLASS</th> <th rowspan="2">TOTAL</th> </tr> <tr> <th>A</th> <th>SICK</th> <th>NO SICK</th> <th>NO SICK</th> </tr> </thead> <tbody> <tr> <td>SICK</td> <td>6954 SICK IDENTIFIED SICK</td> <td>46 SICK IDE- NTIFIED HEALTHY</td> <td>7000</td> <td>10000</td> </tr> <tr> <td>NO</td> <td>412 HEALTHY IDENTIFIED SICK</td> <td>3588 HEALTHY IDENTIFIED SICK</td> <td>3600</td> <td>3600</td> </tr> <tr> <td>Total</td> <td>7366</td> <td>2234</td> <td>10000</td> <td>10000</td> </tr> </tbody> </table>	PREDICTED		CLASS		TOTAL	A	SICK	NO SICK	NO SICK	SICK	6954 SICK IDENTIFIED SICK	46 SICK IDE- NTIFIED HEALTHY	7000	10000	NO	412 HEALTHY IDENTIFIED SICK	3588 HEALTHY IDENTIFIED SICK	3600	3600	Total	7366	2234	10000	10000																								
PREDICTED		CLASS		TOTAL																																													
A	SICK	NO SICK	NO SICK																																														
SICK	6954 SICK IDENTIFIED SICK	46 SICK IDE- NTIFIED HEALTHY	7000	10000																																													
NO	412 HEALTHY IDENTIFIED SICK	3588 HEALTHY IDENTIFIED SICK	3600	3600																																													
Total	7366	2234	10000	10000																																													
2. Transactional data	b)																																																
c ✓	<p>Healthy person: -GTCGCTGGCCAT...</p> <p>Person with β-thalassemia: -GTCGGGCCAT...</p>																																																
3. Ordered data	c)																																																
b ✓	<table border="1"> <thead> <tr> <th>ID</th> <th>Item</th> </tr> </thead> <tbody> <tr> <td>1</td> <td>Bread, Coke, Milk</td> </tr> <tr> <td>2</td> <td>Bread, Bread</td> </tr> <tr> <td>3</td> <td>Bread, Coke, Diaper, Milk</td> </tr> <tr> <td>4</td> <td>Bread, Bread, Diaper, Milk</td> </tr> <tr> <td>5</td> <td>Coke, Diaper, Milk</td> </tr> </tbody> </table>	ID	Item	1	Bread, Coke, Milk	2	Bread, Bread	3	Bread, Coke, Diaper, Milk	4	Bread, Bread, Diaper, Milk	5	Coke, Diaper, Milk																																				
ID	Item																																																
1	Bread, Coke, Milk																																																
2	Bread, Bread																																																
3	Bread, Coke, Diaper, Milk																																																
4	Bread, Bread, Diaper, Milk																																																
5	Coke, Diaper, Milk																																																
4. Cross tabs	d)																																																
a ✓	<table border="1"> <thead> <tr> <th></th> <th>team</th> <th>coach</th> <th>play</th> <th>ball</th> <th>game</th> <th>win</th> <th>lost</th> <th>score</th> <th>season</th> <th>time</th> <th>first</th> </tr> </thead> <tbody> <tr> <td>Doc1</td> <td>1</td> <td>3</td> <td>0</td> <td>0</td> <td>5</td> <td>6</td> <td>0</td> <td>0</td> <td>1</td> <td>0</td> <td>2</td> </tr> <tr> <td>Doc2</td> <td>0</td> <td>0</td> <td>2</td> <td>4</td> <td>0</td> <td>1</td> <td>0</td> <td>2</td> <td>0</td> <td>4</td> <td>1</td> </tr> <tr> <td>Doc3</td> <td>1</td> <td>0</td> <td>0</td> <td>3</td> <td>4</td> <td>0</td> <td>2</td> <td>0</td> <td>1</td> <td>1</td> <td>2</td> </tr> </tbody> </table>		team	coach	play	ball	game	win	lost	score	season	time	first	Doc1	1	3	0	0	5	6	0	0	1	0	2	Doc2	0	0	2	4	0	1	0	2	0	4	1	Doc3	1	0	0	3	4	0	2	0	1	1	2
	team	coach	play	ball	game	win	lost	score	season	time	first																																						
Doc1	1	3	0	0	5	6	0	0	1	0	2																																						
Doc2	0	0	2	4	0	1	0	2	0	4	1																																						
Doc3	1	0	0	3	4	0	2	0	1	1	2																																						

1 Let data be in the range 10,000 to 80,000 normalized to [0.0, 1.0]. Apply min-max normalization to determine the normalized value of 56,500.

ii:

$$\text{normalization} = \frac{v + \min}{\max - \min} \times (\text{newmax} - \text{newmin}) + \text{newmin}$$

$$\text{normalized} = \frac{56500 + 10000}{80000 - 10000} \times (1 - 0) + 0 = 0.95$$

Let there be two Cricket players: Rohit and Hardik, and you have to select one for the cricket world cup. The score of both the players in the last five one-day matches are as follows:

Rohit X	Hardik Y
28	32
38	79
45	2
59	17
63	83

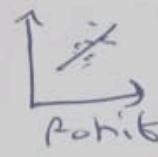
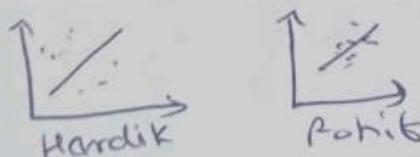
Find the player who is more consistent

$$\frac{28+38+45+59+63}{5} = 46.6$$

$$\frac{32+79+2+17+83}{5} = 42.6$$

$$\frac{(x_i - \bar{x})^2}{n} = 169.04$$

$$\frac{(y_i - \bar{y})^2}{n} = 1064.112$$



from the 6 values
s.d of hardik is more

than the s.d of
Rohit

1 1
2 2

so Rohit is more
consistent ✓

College of Engineering Pune

Test2

Subject: Data Science (Div1 + Div 2)
TY-Computer Engineering

Date : 20/3/2023

Time: 10 to 11 a.m.

Instructions:

1. Scientific calculators are only permissible
2. Over written answers would not be considered for evaluation.

- 1 Find the edit distances using only insertions (cost 1) and deletions(cost 1) between the following pairs of strings. 3

Source: abccdabc , Target: acbdcab

#	a	b	c	c	d	a	b	c	
#	0	01	02	03	04	05	06	07	08
a	01	0	1	2	3	4	5	6	7
c	2	1	2	31	42	53	64	75	6
b	3	2	1	2	3	4	5	4	5
d	4	3	2	3	4	3	4	5	6
c	5	4	3	2	3	4	5	6	5
a	6	5	4	3	4	5	4	5	6
b	7	6	5	4	5	6	5	4	5

Minimum Cost= 5

- 2 Give the formula for Cosine Similarity and determine the Cosine Similarity between the given document collection D and test document D4 specifying the document closer to test document. 4

[Apply stop-word removal, lemmatization] [Stop words: to, from, is]

Document Collection D:

D1: Jack travelled to London

D2: Jack travelled from Oakland to London

D3: Travel to Oakland is wonderful

Test Document D4: Jack travelled to Oakland

D		D4	
SUPPORT COUNT		SUPPORT COUNT	
Jack	2	Jack	1
travel	3	travel	1
London	2	London	0
Oakland	2	OAKLAND	1
wonderful	1	wonderful	0

$$[2, 3, 2/2, 1] \cdot [1, 1, 0/1, 0] = 1.61 / 1.41 \cos \theta$$

$$D1: [1, 1, 1, 0, 0] \cdot [1, 1, 0, 1, 0] = 1.01 / 1.41 \cos \theta_1$$

$$\therefore \cos \theta_1 = \frac{1+1}{\sqrt{3} \cdot \sqrt{2}} = \frac{2}{3}$$

$$\theta_1 = 48.16^\circ$$

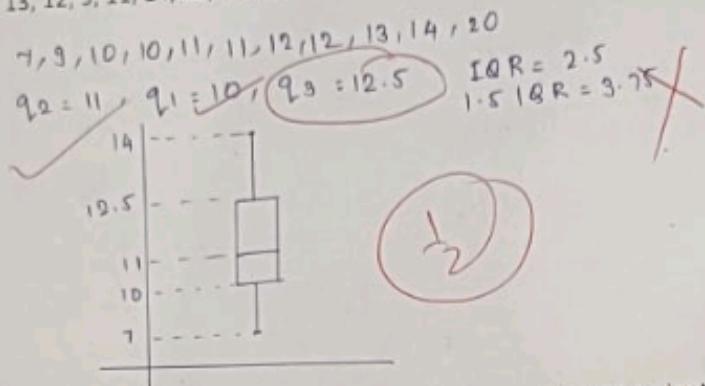
Ctrl



Alt

2

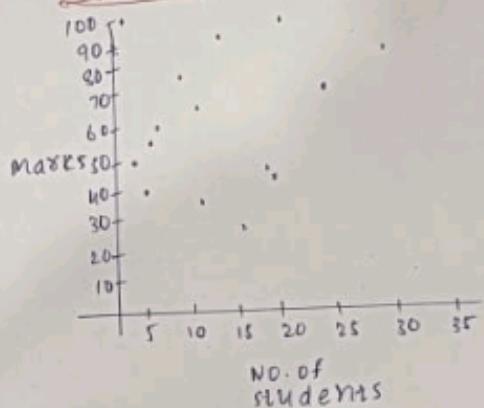
- 8 Dataset given below shows weight of students in a class. Draw box plot with whiskers for the same. Analyze the box plot.
Weight: 13, 12, 9, 11, 14, 12, 10, 20, 11, 10, 7



- 9 Investigate whether there's a relationship between the number of students and the marks scored. Which plot will help and how? Assume the data below:

No of students	Marks Obtained	Percentage Of Students
5	40	2.5
6	60	3
25	70	12.5
11	65	5.5
30	80	15
4	50	2
6	55	3
10	75	5
14	90	7
18	45	9
20	40	10
22	95	11
2	100	1
11	35	5.5
16	25	8

Scatter plot will help using Regression.



∴ There's no relation between them.

	Estimated Jaccard similarity	Time Tardard similarity
S1, S2	1/3	0
S1, S3	1/3	0
S1, S4	2/3	1/4
S2, S3	2/3	1/4
S2, S4	2/3	1/4
S3, S4	2/3	

4½

- [Fill in the blanks: 4-7]
- 4 Anti-monotone property is: support count of itemset is \leq support count of its subsets 1
 - 5 State Apriori principle: All the subsets of frequent itemset are also frequent itemsets. 1
 - 6 The Prune step of Apriori algorithm is: If support count of itemset $<$ minsup or an itemset is a superset of an itemset whose support count $<$ minsup prune it. 1
 - 7 If minsup is set too low then too many values and calculations, large memory needed. 1

4

Lock

Z e X e C e V e B e N e M e < e > e

Shift

$$D_2 : [1, 1, 1, 1, 0] \cdot [1, 1, 0, 1, 0] = 1D211D41 \cos\theta_2$$

$$1+1+1 = \sqrt{4} \cdot \sqrt{3} \cdot \cos\theta_2$$

$$\cos\theta_2 = \frac{3}{2\sqrt{3}} = \frac{\sqrt{3}}{2} \quad \theta_2 = 30^\circ$$

$$D_3 : [0, 1, 1, 0, 1, 1] \cdot [1, 1, 1, 0, 1, 0] = 1D311D41 \cos\theta_3$$

$$1+1 = \sqrt{3} \cdot \sqrt{3} \cdot \cos\theta_3$$

$$\cos\theta_3 = \frac{2}{3} \quad \theta_3 = 48.16^\circ$$

As $\theta_2 < \theta_1, \theta_3$
 D_2 is closer to D_4 .

(4)

- 3 Given the following data. Fill the matrix by computing the minhash signature for each column using the given hash functions. 5

Element	S1	S2	S3	S4	$2x+1 \text{ mod } 6$	$3x+2 \text{ mod } 6$	$5x+2 \text{ mod } 6$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	4
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

Determine signatures matrix? [Note: Show steps while updating signature matrix starting from initialization]

Justify, how close are the estimated Jaccard similarities for all pairs of columns to the true Jaccard similarities?

	S1	S2	S3	S4
$2x+1 \text{ mod } 6$	00	10	10	00
$3x+2 \text{ mod } 6$	00	00	00	00
$5x+2 \text{ mod } 6$	00	00	00	00

00	1	00	1	00	1	00	1
00	2	00	2	00	2	00	2
00	2	00	2	00	1	00	2

5	1	00	1	5	1	01	1	5	1	1	1
2	2	00	2	2	2	00	2	2	2	2	2
0	1	00	0	0	1	0	1	4	0	1	4

S1	S2	S3	S4
1	1	1	1
2	2	2	2
0	1	4	0

signature matrix

Shot on OnePlus

By Jay 2023.04.20 14:54