

MIDTERM EXAM
ANAV PRASAD ap-7152

A. True Or False

- | | |
|----------|-----------|
| 1. True | 11. False |
| 2. False | 12. False |
| 3. False | 13. True |
| 4. False | 14. True |
| 5. False | 15. True |
| 6. False | |
| 7. False | |
| 8. False | |
| 9. False | |
| 10. True | |

B.

PCA

I.

Pseudo Code:

```
public double [][] PCA (double [][] M, int n, int m)  
    double [][] MT = compute Transpose (M);  
    // MT → mxn matrix
```

```
    double [][] MTM = compute Matrix Product (MT, M);  
    // MTM = MT × M → mxm matrix
```

```
[double [][] temp;]  
temp = MTM - 1 /
```

```
double [] lambdas = compute Eigenvalues (MTM);  
// lambdas → size m
```

```
double [][] eigenvectors = compute Eigenvectors (MTM, lambdas);  
// eigenvectors → mxm
```

~~double [][]~~

Sort the columns of eigenvectors on the basis of
corresponding eigenvalues ~~order~~ in decreasing order

B.1, B.2

Discard the columns whose eigenvalues are too low comparatively.

Let the reduced matrix of eigenvectors be stored in double [] [] reducedMat.

// reduced Mat \rightarrow $m \times k$ for some k

\rightarrow double [] [] answer = computeMatrixProduct(M,
reduced Mat);

// answer \rightarrow $n \times k$

return answer;

g

2.

Let $M = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$

$\therefore M^T M = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 3 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix}$

~~$\begin{bmatrix} 30 & 26 \\ 28 & 30 \end{bmatrix}$~~

$\begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix}$

B.2

$$\therefore |MIM - \lambda I| = \begin{vmatrix} 30-\lambda & 28 \\ 28 & 39-\lambda \end{vmatrix}$$

$$\Rightarrow (30-\lambda)^2 - 28^2 = 0 \Rightarrow \lambda - 30 = \pm 28$$
$$\Rightarrow \lambda = 58, 2$$

$$\Rightarrow \lambda^2 - 69\lambda + 30 \times 39 - 31^2 = 0$$

$$\Rightarrow \lambda^2 - 69\lambda + 1170 - 961 = 0,$$

$$\Rightarrow \lambda^2 - 69\lambda + 209 = 0$$

$$\Rightarrow \lambda = \frac{69 \pm \sqrt{4761 - 4 \times 209}}{2}$$

$$= \frac{69 \pm \sqrt{3925}}{2} = \frac{69 \pm 5\sqrt{157}}{2}$$

$$= \frac{69 \pm 62.65}{2}$$

$$\Rightarrow \lambda = 65.8, 3.2$$

$$\therefore \lambda_1 \approx 65.8, \lambda_2 \approx 3.2$$

$$\therefore \lambda_1 = \frac{69 + 5\sqrt{157}}{2}$$

$$\lambda_2 = \frac{69 - 5\sqrt{157}}{2}$$

$$\therefore \lambda = 65.8$$

$$\lambda_2 = 3.2$$

B. 2

$$\therefore \lambda_1 = 58, \lambda_2 = 2$$

Since, $\lambda_2 \ll \lambda_1$, we can drop λ_2 .

Let $\mathbf{q} = \begin{bmatrix} x \\ y \end{bmatrix}$

$$\therefore \begin{bmatrix} 30 & 28 \\ 28 & 30 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = 58 \begin{bmatrix} x \\ y \end{bmatrix}$$

$$\therefore 30x + 28y = 58x \Rightarrow 28y = 28x$$

$$\therefore 28x + 30y = 58y \Rightarrow x = y$$

\therefore Normalized: $\mathbf{q}_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

\therefore Answer is

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \left(\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \right) =$$

$$\begin{bmatrix} 3/\sqrt{2} \\ 3/\sqrt{2} \\ 7/\sqrt{2} \\ 7/\sqrt{2} \end{bmatrix}$$

Ans

B. 3

SVD

3. /

→ Judging by Σ matrix, first concept can be broadly termed as "science fiction".

Second concept as "romance".

Third concept is more difficult to intuitively understand but, none thinks it arises from the women's low rating of science fiction movies.

→ Guys are more interested in science fiction movies as seen by Σ U matrix

→ Girls are more interested in romance movies as seen by Σ U matrix.

→ Σ matrix relates the peoples interest to the different concepts of movies.

B. 3, B. 4

→ V^T matrix relates movies and concepts.
That is which movie belongs to which concept.

4.

→ I'll remove the third movie concept
since it holds relatively less importance
than others.

That is, we have the following matrix

$$\begin{bmatrix} .13 & .02 \\ .41 & .07 \\ .55 & .09 \\ .68 & .11 \\ .15 & -.59 \\ .67 & -.73 \\ .07 & -.29 \end{bmatrix}_{7 \times 2} \quad \Sigma' \quad V^T$$

\downarrow

$$U'$$

B.4, B.5

- So people in dataset in concept would be represented by $U^T \Sigma$.
- Movies & and Σ up to relation would be given by ΣV^T .
- New people - movie relations would be given by $U^T \Sigma' V^T$

13/

$$\therefore \begin{bmatrix} 3.5 & 6.5 \end{bmatrix}_{1 \times 2} \times \begin{bmatrix} 12.4 & 0.9 \\ 0 & 9.5 \end{bmatrix}_{2 \times 2}$$

=

$$= \begin{bmatrix} 43.4 & 61.75 \end{bmatrix}$$

i. X likes romance movies.

C

I.

Stop Words: The, is, and, in, very

Removing stop words:

D₁: sky blue ~~the~~ clear

D₂: sun sky bright.

	blue	bright	clear	sky	sun
D ₁	1	0	1	1	0
D ₂	0	1	0	1	1

TFIDF:

~~D₁~~ = D₂

After TFIDF:

	blue	bright	clear	sky	sun
D ₁	$\frac{1}{3}$	0	$\frac{1}{3}$	0	0
D ₂	0	$\frac{1}{3}$	0	0	$\frac{1}{3}$

$$\frac{1}{3} \cdot \log\left(\frac{2}{2}\right) = 0$$

C. 2

2

After removing stop words

D_3 : sun sky

sun rocky

$\text{tf}(D_3)$

D_3

blue bright clear sky sun

0

0

0

1

After TF-IDF

$\text{tf}(D_3)$

blue bright clear sky sun

0

0

0

0

$1/2 \log(3/2)$

$1/2 \cdot \log(3/2) = 0$

$\therefore D_1, D_3$ similarity

$$= \frac{0 + 0 + 0 + 0 + 1/2 \log(3/2) \cdot 0}{\|D_1\| \|D_3\|}$$
$$= 0$$

$\therefore D_2, D_3$ similarity

$$= \frac{0 + 0 + 0 + 0 + 1/2 \log(3/2) \cdot 1/2}{\sqrt{2} \log(3/2) \cdot \frac{\sqrt{2}}{3}}$$

$$= \frac{1}{\sqrt{2}}$$

C2

$\therefore D_3$ is most similar to D_2 by cosine similarity.

Approach:

Firstly stop words were removed i.e.
~~the~~ 'You', 'can', 'see', 'the', 'in'.

Then, since lemmatization was again not required, TFIDF was done.

Now, cosine similarity was computed between D_1, D_3 ($= 0.9$) and D_2, D_3 ($= 1.0$) which resulted in D_2 being most similar to D_3 by cosine similarity.

C.3.

3.

We have 3 documents already. I'd use k-Means ($k=3$) to cluster any new document (with the pre-existing centroids denoted by the already present 3 documents) and the metric being cosine similarity).

C. 4

4.

→ NER :

NER's purpose is to find out words that belong together so as to not have misleading or incorrect teams in ~~the~~ when making the document - team matrix.

For example : New York University
If the above was to be mentioned in a document, NER ~~was~~ would be used to ~~figure~~ figure out that New York University is one team and would be stored as ~~one team~~ ^{team} in ~~the~~ the document - team matrix (as say, New - York - University).

→ Stanford NLP:

By the above, I assume it refers to ~~the~~ Stanford CoreNLP ~~library~~ library that was used to do lemmatization and NER ~~in~~ in pre processing.

→ Stemming:

Stemming refers to the dropping of word stem to retain the 'core' of the word during preprocessing.

for e.g. 'running' → 'run'
'hanging' → 'hang'
'studying' → 'study'

→ Sliding Windows

Use sliding windows to find common phrases by frequency count of n-grams. (Crossword)

e.g. computer science

→ Lemmatisation

Finds 'core' of a word in its normal form

e.g. 'studies' → 'study'

→ N-Grams:

Contiguous segment of ~~n~~ n - words

e.g. 2-gram : We love science

2-gram : "We love", "love science"

- Term Doc Mat:-

Matrix formed by storing counts of terms in documents by document denoting rows terms denoting columns.

Eg. D₁: We ran

- Cluster View:

Dr: I am

	We	I	am	ran
D ₁	1	0	0	1
D ₂	0	1	1	0

• Cluster View:-

- Precision :-

$$= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$= \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

- Recall = True Positive

$$= \frac{\text{True Positive}}{\text{True Positive} + \text{False Negatives}}$$

D. I

1.

No.

That also sort to

That is so because in the given table,
all the values of zip codes ~~are~~
are unique,

D. 2

2.

Entropy of discount:

$$= -\frac{2}{6} \log_2 (2/6) - \frac{4}{6} \log_2 (4/6)$$

$$\approx 0.9183$$

Entropy conditional entropy of lifestyle:

$$= \frac{3}{6} \text{ Entropy } (2/3, 1/3) + \frac{3}{6} \cdot \text{ Entropy } (0, 1)$$

$$= 0.4591$$

$$\therefore \text{IG (Lifestyle)} = 0.9183 - 0.4591 \\ = \underline{\underline{0.4592}}$$

(conditional entropy of Age):

$$= \frac{3}{6} \cdot \text{ Entropy } (2/3, 1/3) + \frac{2}{6} \cdot \text{ Entropy } (0, 1) \\ + \frac{1}{6} \cdot \text{ Entropy } (0, 1)$$

$$\approx 0.4591$$

$$\therefore \text{IG (Age)} = \underline{\underline{0.4592}}$$

D.2, D.3

Both are equally predictive since they have same IG.

~~3,~~

Let P_f be a discrete function as follows

$$P_f = \begin{cases} 30-40 \\ \dots \end{cases}$$

I'd use Bayesian ~~prob~~ ~~Bayes~~ Decile to compute probability for discount ranges for both lifestyle and Age. Then, I'd take average of both probabilities to get the final probability to decide the final discount.

Example in D.4

D. 4.

4

For David let $D_1: \cancel{6\% - 15\%}$
 $D_2: 1\% - 5\%$

PPD

For David:

$P(D_1 | \text{Age} 20-30)$

$$= \frac{\cancel{P(D_1) P(\text{Age})}}{P(20-30)} \cdot \frac{P(20-30 | D_1) P(D_1)}{P(20-30)}$$

$$= \frac{0 \times P(D_1)}{P(20-30)} = 0$$

~~Do D1 Average~~ $\Rightarrow P(D_1 | \text{Age})$

$$P(D_2 | 20-30 \text{ age}) = \frac{P(20-30 | D_2) P(D_2)}{P(20-30)}$$

$$= \frac{1 \times \frac{4}{6}}{\frac{2}{6}}$$

=

D.5

E

Drawback :-

→ ~~At the paper says~~

→ On the basis of the paper "The Best Two Independent measurements are not the Two Best" by Thomas M. Cover ~~and Bertrami~~

The premise of Forward Selection algorithm that works by greedy building of the feature set is flawed and incorrect and would not necessarily lead to the best ~~set~~ feature set.

Based on the description, the algorithm is Sequential Backward selection Algorithm

Pseudo-code :-

- 1 Start with the entire feature set $y = \{x_1, x_2, \dots, x_n\}$
- 2 Select the next ^{worst} feature x_k

$$x^* = \underset{x \in Y_k}{\operatorname{arg\,min}} [J(y_k) - J(y_{k-1} \setminus x)].$$

3. Update $\gamma_{k+1} = \gamma_k - x^-$; $k = k+1$

4. Go to 2.

Drawback :-

- Drawback is similar to the forward selection algorithm. That is, the greedy approach of to drop or remove the least significant feature is not necessarily the best move.

E.

I.

→ Cosine Similarity:

Compute similarity by ~~by~~ using the counts of data items in various categories to form a vector and using the vectors to compute cosine similarity.

Say A and B are vectors.

$$\text{Similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

→ Jaccard Similarity / Index:

Compute similarity by compute size of unions and intersections of the data items.

$$\text{Jaccard}(X, Y) = \frac{|\text{Intersection}(X \cap Y)|}{|\text{Union}(X \cup Y)|}$$

E.3

3.

~~Process~~

k Means ++ initialization:

Choose first centroid randomly from given data points.

Then ~~choose~~ to choose the second point, find the point in dataset farthest from the first chosen centroid and, select that as the 2nd centroid.

3

The authors tested for causation by using the Granger Causality to ~~test~~ not actually test for causation between mood and DJIA values but instead to test if the Twitter mood time series had any predictive effect on DJIA values.

Granger Causality tests for causality by searching for past precedences of causal relationships (direct or ~~a~~ indirect) to do causal prediction now.

F

FAQ 4.

- Action: Surveillance by authority going to people's home (in vac, drives) to search for disease
- Person is "known" by doctors in hospitals