MIDTERM EXAM

ANAV PRASAD ap7152

A. True Or False

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

B.

PCA

1.

Pseudo Code :-

```
public double[][] PCA(double[][] M, int n, int m){
    double[][] MT = computeTranspose(M);
        // MT → n x m matrix
    double[][] MTM = computeMatrixProduct(MT, M);
        // MTM = MT x M → m x m matrix
    double[][] temp;
    temp = MTM - λ

    double[] lambdas = computeEigenvalues(MTM);
        // lambdas → size m
    double[][] eigenvectors = computeEigenvectors(MTM, lambdas
        // eigenvectors → m x m
```

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

B.1, B.2

Discard the columns whose corresponding eigen values are too low comparatively.
Let the reduced matrix of eigen vectors be stored in double [][] reducedMat.
// reducedMat → m × k for some k
double [][] answer = computeMatrixProduct(M, reducedMat);
// answer → n × k
return answer;
}

2.

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

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

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

B.1

$$\therefore \|M^TM - \lambda I\| = \left\| \begin{bmatrix} 30-\lambda & 28 \\ 28 & 39-\lambda \end{bmatrix} \right\|$$

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

$$\Rightarrow \lambda = 58, 2$$

$$\Rightarrow \lambda^2 - 69\lambda + 30\times39 - 31^2 = 0$$

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

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

$$\Rightarrow \lambda = \frac{69 \pm \sqrt{4761 - 4\times209}}{2}$$

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

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

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

$$\therefore \lambda_1 \cong 65.8, \quad \lambda_2 \cong 3.2$$

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

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

$$\therefore \lambda_1 = 65.8$$

$$\lambda_2 = 3.2$$

B. 2

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

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

Let $e_1 = \begin{bmatrix} x \\ y \end{bmatrix}$

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

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

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

$\therefore$ Normalised: $e_1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$

$\therefore$ Answer :-

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 4 \\ 4 & 3 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 3/\sqrt{2} \\ 3/\sqrt{2} \\ 7/\sqrt{2} \\ 7/\sqrt{2} \end{bmatrix}$$

$\therefore$ 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, nonetheless, it arises from the women's low rating of science fiction movies.

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

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

→ U 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

$$\begin{bmatrix} .13 & .02 \\ .41 & .07 \\ .55 & .09 \\ .68 & .11 \\ .15 & -.59 \\ .07 & -.73 \\ .07 & -.29 \end{bmatrix}_{7\times 2} \quad \begin{bmatrix} 12.4 & 0 \\ 0 & 9.5 \end{bmatrix}_{2\times 2} \quad \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \end{bmatrix}_{5\times 2}$$

U' ↓ Σ' ↓ $V^{T'}$

B.4, B.5

→ So people's interest in concepts would be represented by $U'\Sigma'$.

→ Movies & and concept to relation would be given by $\Sigma' V^T$

→ New people - movie relation would be given by $U'\Sigma' V^T$

5.

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

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

∴ X likes romance movies.

C

1.

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

Removing stop words:

D_1 : sky blue clear

D_2 : sun sky bright

	blue	bright	clear	sky	sun
D_1	1	0	1	1	0
D_2	0	1	0	1	1

TFIDF:

After TFIDF:

	blue	bright	clear	sky	sun
D_1	1/3	0	1/3	0	0
D_2	0	1/3	0	0	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~~ sky

~~D_3~~

	blue	bright	clear	sky	sun
D_3	0	0	0	1	1

After TF IDF:

	blue	bright	clear	sky	sun
D_3	0	0	0	0	$1/2 \log(3/2)$

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

$\therefore D_1, D_3$ similarity

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

$$= 0$$

$\therefore D_2, D_3$ similarity:

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

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

Q.2

∴ D_3 is most similar to D_2 by cosine similarity.

Approach :→

Firstly stop words were removed i.e. '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 ($=0$) and D_2, D_3 ($=1/\sqrt{2}$) 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 terms in when making the document-term matrix.

For example : New York University

If the above was to be mentioned in a document, NER would be used to figure out that New York University is one term and would be stored as one term in the document-term matrix (as, say, New_York_University).

→ Stanford NLP:

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

→ Stemming :-

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

for eg. 'running' → 'run'
'hanging' → 'hang'
'studies' → 'studi'

→ Sliding Window

Use sliding windows to find common phrases (not found by NER) by frequency count of n-grams. (n=2 or 3)

Eg. computer science

→ Lemmatization :-

Finds 'core' of a word in its normal form.

Eg. 'studies' → 'study'

→ N-Grams :-

Contiguous segment of n-words

Eg. 2-gram : We love science
2-grams : "We love", "love science"

- Term Doc Mat :-
  Matrix formed by storing counts of terms in documents by documents denoting terms denoted by columns.

  Eg D_1: We ran.

  D_2: I ran.

	We	I	ran	ran
D_1	1	0	0	1
D_2	0	1	1	0

- Precision :

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

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

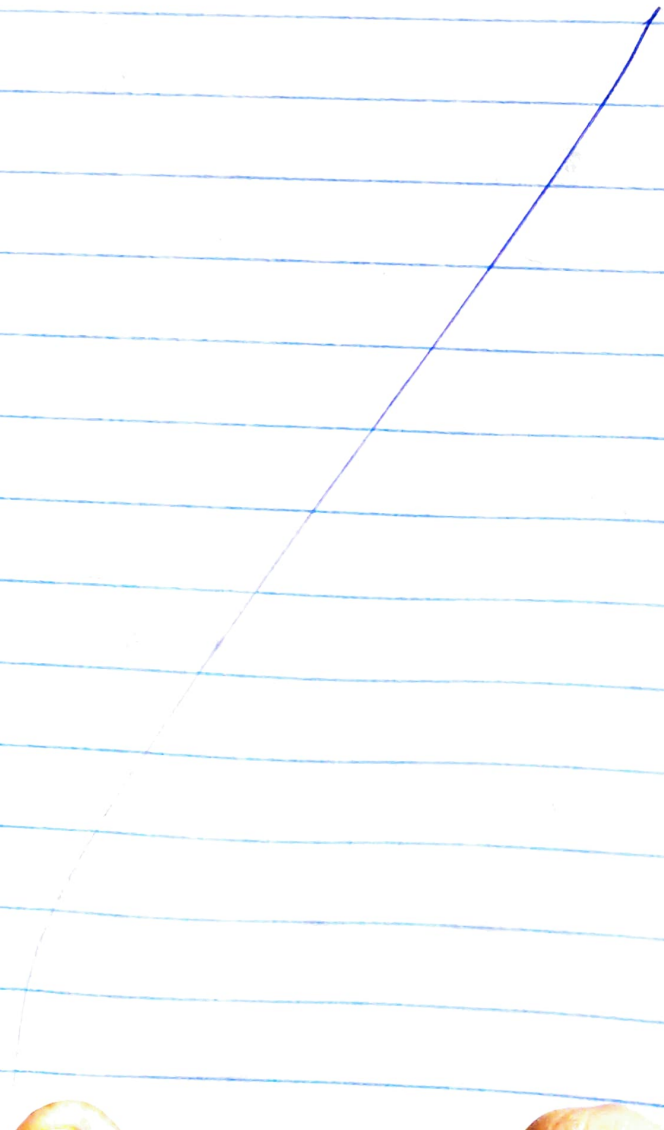
D.1

1.

No.

~~That is so because~~

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



D. 2

2.

Entropy of discount:

$= -2/6 \log_2(2/6) - 4/6 \log_2(4/6)$

$\simeq 0.9183$

Conditional entropy of Lifestyle:

$= 3/6$ Entropy $(2/3, 1/3) + 3/6 \cdot$ Entropy $(0, 1)$

$= 0.4591$

$\therefore$ IG (Lifestyle) $= 0.9183 - 0.4591$

$= \underline{0.4592}$

Conditional entropy of Age:-

$= 3/6 \cdot$ Entropy $(2/3, 1/3) + 2/6 \cdot$ Entropy $(0, 1)$
$+ 1/6 \cdot$ Entropy $(0, 1)$

$= 0.4591$

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

D.2 , D.3

Both are equally predictive since they have same IG.

3.

Let P_i be a discrete function as follows

$$P_i = \begin{cases} 30-40 \end{cases}$$

I'd use conditional probability Bayes Rule 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 Q D.4

Q. 4.

4

Let D_1: 6% - 15%

D_2: 1% - 5%

For David:

$P(D_1 \mid 20\text{-}30 \text{ age})$

$$= \frac{P(20\text{-}30 \mid D_1)\, P(D_1)}{P(20\text{-}30)}$$

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

$P(D_2 \mid 20\text{-}30 \text{ age})$

$$= \frac{P(20\text{-}30 \mid D_2)\, P(D_2)}{P(20\text{-}30)}$$

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

=

D.S

5

Drawbacks:-

1. ~~As the paper~~

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

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 feature set.

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

Pseudo-code:-

1. Start with the entire feature set $Y = X$
2. Select the next worst feature $x^- = \underset{x \in Y_k}{\operatorname{argmin}} [J(Y_k) - J(Y_k \setminus x)]$

3. Update $y_{k+1} = y_k - x^-$; $k = k+1$

4. Go to 2.

Drawback :-

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

E.

1.

→ Cosine Similarity:

Compute similarity 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

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

→ 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.

k Means ++ initialization:

Choose first centroid randomly from given data points.

Then 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 indirect) to do causal prediction now.

F

EQ 4.

→ Active: Surveillance by actively going to people's home (in val. drives) to search for diseases

→ Passive: "" by doctors in hospitals.

F.1

1

→ Firstly, replace missing values by replacing by your choice of zero values, average, etc.

~~→ Check for errors~~

→ Do dimension reduction by PCA or SVD.

F 2.

2.

2.1.

Poll predictions and predictions of who is going to win

Predictive Analytics used by government or politicians for sentiment analysis of people for general survey and opinions

2.2.

Sentiment analysis using news headlines or twitter feed to predict stock market.

2.3

Analyzing Twitter feed to possibly predict and catch out breaks of diseases (like Google Trends but done better).

F

2.3 cont.

Alternative Data sources like Twitter feed, reddit (social media in general) could possibly bring new value to solve problems in healthcare using data science