University of Massachusetts Dartmouth
Department of Computer and Information Science
CIS 530-02 Advanced Data Mining – Exam II (Spring 2024)

Friday, May 3rd, 2024

Printed Full Name: _____Banoth Jeevan Kumar_____

Student ID: _____02105145_____

DO NOT TURN THE PAGE OVER UNTIL YOU ARE INSTRUCTED TO DO SO

Please read the following instructions:

1. You have 180 minutes to complete the examination.
2. This examination is OPEN materials, including notes, slides and books.
3. Type your answer in space provided on the examination sheets, any work not on the examination sheets will not be graded.
4. Type your answers legibly.
5. Submit your answer according to the instruction for grading by the end of the examination.
6. DO NOT communicate any of your classmates during the examination.

Honor Policy: copying in whole or in part of the examination will be considered to be an act of scholastic dishonesty. Students who violate university rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the university. Since such dishonesty harms individuals, all students, and the integrity of the university, policies on scholastic dishonesty will be strictly enforced.

I have read the above instructions and I will act in accordance with all of them.

_____B Jeevan Kmar_____ _____03/05/2024_____
Student Signature Date

Type your name and date to agree the policy before you start!

This examination contains three sections. The whole midterm examination carries 100 points.

Section I. Single-Choice Questions (20 points, 2 points per question; only ONE choice is correct). Please write your answers in the table provided below.

Question	1	2	3	4	5	6	7	8	9	10
Answer	a	e	a	b	d	d	d	d	d	a

1. If k-means is sensitive to outliers, what algorithm can be used to deal with outliers?
   a. K-medians
   b. K-means++
   c. K-means
   d. Hierarchical Clustering
2. How can we prevent local minima resulting from K-means?
   e. Using K-means ++
   a. Using K-medians
   b. Using KNN
   c. Using K-mode

3. What is the tree size?
   a. The number of nodes
   b. The number of terminal nodes
   c. The number of subtrees
   d. The number of parent nodes

4. Which statistic does PCA looks at in the high-dimensional data?
   a. Mean
   b. Variance
   c. Correlation
   d. Median

5. If the clusters are not spherical, which method can be considered in clustering?
   a. K-means
   b. CART
   c. Naïve Bayes
   d. Spectral clustering

6. What criteria can we use to select the optimal number of clusters in hierarchical clustering?
   a. Silhouette plots
   b. Gap statistic
   c. Cophenetic correlation
   d. All the above

7. Cross-validation (CV) can be used for?
   a. Only choosing the optimal tuning parameter in L1 regression
   b. Only choosing the optimal tuning parameter in L2 regression
   c. Only choosing the optimal tuning parameter in the pruning process of CART
   d. Choosing the optimal tuning parameter in regularized regression and CART

8. The feature detection layers of Convolutional neural network perform
   a. Only convolution
   b. Only pooling
   c. Only ReLU
   d. Convolution, pooling or ReLU

9. Which dissimilarity computing methods are <u>not</u> used in hierarchical clustering?
   a. Max-link or Min-link
   b. Ward's method
   c. Average-link
   d. None of the above

10. Which algorithm is widely used in normalized spectral clustering?
    a. Ng-Jordan-Weiss algorithm
    b. Greedy
    c. Top-down and greedy approach
    d. Top-down

Section II. True or False questions (20 points, 1 point per question).

Questions	True	False
1. Theoretically, the total variance of a dataset is equal to the variance explained by components identified in PCA		False
2. For classification tree, we examine the MSE for accuracy		False
3. DNN typically contains 2 layers		False
4. Scaling would change the clustering results	True	
5. For hierarchical clustering, we draw conclusions about the similarity of two observations using dendrogram based on their proximity along the horizontal axis	True	
6. Different similarity criteria can lead to different clustering results	True	
7. PCA is for clustering		False
8. K-means is for dimension reduction		False
9. K-means can be used when the clusters are non-convex		False

10. When Y is a continuous variable, multiple regression, regularized regression and regression trees can be considered.	True	
11. Forgy initialization randomly initializes centroids anywhere in the clustering space		False
12. K-means typically works for clusters that have different density		False
13. Ward's method computes the mean distance between data points of each cluster in hierarchical clustering		False
14. The Ng-Jordan-Weiss algorithm includes two major approaches, dimension reduction and clustering.	True	
15. In dendrogram, the height of each node is not proportional to the value of the intergroup dissimilarity between its two daughters		False
16. The Gap statistic compares the expected and observed curves based on the log of the within cluster dissimilarity.	True	
17. It is impossible to generate empty clusters using K-means		False
18. Both spectral clustering and the kernelized distance-based method will be appropriate for non-convex clusters.	True	
19. In PCA, the proportion of variance explained sums to 10		False
20. The biplot displays both the principal component (PC) scores and the PC loadings	True	

Section III. Short problems (60 points)

1. **(5 points)** Given the observed data below,

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Show your stepwise calculation for assigning the class label for a new animal with the following attribute values, using Naïve Bayes.

Give Birth	Can fly	Live in water	Have Legs	Class (Mammal or non-mammal)
Yes	No	No	Yes	?

(No score will be given, if you only answer "mammal" or "non-mammal")
Your answer:

To assign the class label for a new animal with the attribute values "Yes No No Yes" using Naïve Bayes, we need to calculate the probability of each class given the attributes.

Step 1: Calculate the prior probabilities

→ Let's calculate the prior probabilities of each class:

→ P(Mammal) = Number of mammals / Total number of animals = 7 / 17 ≈ 0.412
→ P(Non-mammal) = Number of non-mammals / Total number of animals = 10 / 17 ≈ 0.588

Step 2: Calculate the likelihoods

Now, let's calculate the likelihoods for each attribute:

- → P(Give Birth = Yes | Mammal) = Number of mammals that give birth / Total number of mammals = 6 / 7 ≈ 0.857
- → P(Give Birth = Yes | Non-mammal) = Number of non-mammals that give birth / Total number of non-mammals = 2 / 10 ≈ 0.2
- → P(Can Fly = No | Mammal) = Number of mammals that cannot fly / Total number of mammals = 6 / 7 ≈ 0.857
- → P(Can Fly = No | Non-mammal) = Number of non-mammals that cannot fly / Total number of non-mammals = 8 / 10 ≈ 0.8
- → P(Live in Water = No | Mammal) = Number of mammals that do not live in water / Total number of mammals = 6 / 7 ≈ 0.857
- → P(Live in Water = No | Non-mammal) = Number of non-mammals that do not live in water / Total number of non-mammals = 6 / 10 ≈ 0.6
- → P(Have Legs = Yes | Mammal) = Number of mammals that have legs / Total number of mammals = 7 / 7 ≈ 1
- → P(Have Legs = Yes | Non-mammal) = Number of non-mammals that have legs / Total number of non-mammals = 6 / 10 ≈ 0.6

Step 3: Calculate the posterior probabilities

Now, let's calculate the posterior probabilities:

- → P(Mammal | Give Birth = Yes, Can Fly = No, Live in Water = No, Have Legs = Yes) = P(Mammal) * P(Give Birth = Yes | Mammal) * P(Can Fly = No | Mammal) * P(Live in Water = No | Mammal) * P(Have Legs = Yes | Mammal) ≈ 0.412 * 0.857 * 0.857 * 0.857 * 1 ≈ 0.243
- → P(Non-mammal | Give Birth = Yes, Can Fly = No, Live in Water = No, Have Legs = Yes) = P(Non-mammal) * P(Give Birth = Yes | Non-mammal) * P(Can Fly = No | Non-mammal) * P(Live in Water = No | Non-mammal) * P(Have Legs = Yes | Non-mammal) ≈ 0.588 * 0.2 * 0.8 * 0.6 * 0.6 ≈ 0.085

Step 4: Assign the class label

Finally, we assign the class label based on the posterior probabilities:

- → Since P(Mammal | Give Birth = Yes, Can Fly = No, Live in Water = No, Have Legs = Yes) > P(Non-mammal | Give Birth = Yes, Can Fly = No, Live in Water = No, Have Legs = Yes), we assign the class label as **Mammal.**

2. **(10 points)** Given the observed data and the reference table below,

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

naive Bayes classifier:

p(Refund = Yes | No) = 3/7
p(Refund = No | No) = 4/7
p(Refund = Yes | Yes) = 0
p(Refund = No | Yes) = 1
p(Marital Status = Single | No) = 2/7
p(Marital Status = Divorced | No) = 1/7
p(Marital Status = Married | No) = 4/7
p(Marital Status = Single | Yes) = 2/3
p(Marital Status = Divorced | Yes) = 1/3
p(Marital Status = Married | Yes) = 0

For Taxable Income:
If Class = No: sample mean = 110
sample variance = 2975
If Class = Yes: sample mean = 90
sample variance = 25

(Note: the sample mean is in the unit of K).

Show your stepwise calculation for assigning the class label for a new customer with the following attribute values, using Naïve Bayes.

Refund	Marital Status	Taxable Income	Evade Class (No or Yes)
No	Single	80K	?

(No score will be given, if you only answer "Yes" or "No")
Hint: For Taxable income, it follows the normal distribution.

$$P(x_j \mid C_i) = \frac{1}{\sqrt{2\pi s_{ji}^2}} e^{-\frac{(x_j - \alpha_{ji})^2}{2s_{ji}^2}}$$

Your answer:

To assign the class label for a new customer with the attribute values of Refund = No, Marital Status = Single, Taxable Income = 80K, and Evade = ?, we can use the Naive Bayes classifier.

First, we need to calculate the probability of each attribute value given the class label. For Taxable Income, we assume that it follows a normal distribution. Therefore, we can calculate the probability of Taxable Income = 80K given the class label as follows:

P(Taxable Income = 80K | Class = No) = (1/2) * (1/2) * (1/2) * (1/2) = 1/16

Next, we can calculate the probability of each attribute value given the class label using the formula:

P(Attribute | Class) = P(Attribute) * P(Class | Attribute)

For Refund = No, we can calculate the probability as follows:

P(Refund = No | Class = No) = (1/2) * (1/2) * (1/2) * (1/2) = 1/16

For Marital Status = Single, we can calculate the probability as follows:

P(Marital Status = Single | Class = No) = (1/2) * (1/2) * (1/2) * (1/2) = 1/16

For Evade = ?, we can calculate the probability as follows:

P(Evade = ? | Class = No) = (1/2) * (1/2) * (1/2) * (1/2) = 1/16

Finally, we can calculate the probability of the class label given the attribute values using the formula:

P(Class | Attribute) = P(Attribute | Class) * P(Class)

For Class = No, we can calculate the probability as follows:

P(Class = No | Refund = No, Marital Status = Single, Taxable Income = 80K, Evade = ?) = (1/2) * (1/2) * (1/2) * (1/2) * (1/16) = 1/1024

Therefore, the class label for the new customer is No.

3. **(15 points, 5 points*3)** A researcher only has attributes X for a variety of dogs, and would like to explore or describe which species they belong to.
   1) Which machine learning method would this company use to help their decision, unsupervised or supervised?
      Your Answer:

   Unsupervised learning would probably be used by the researcher to investigate and characterize the species to which the dogs belong. This is because there are no labeled data to train a supervised learning model, and the researcher just has attribute X for the dogs.

   Grouping the dogs according to their attribute values can be accomplished by unsupervised learning techniques like clustering algorithms. Without labeled data, these algorithms can find patterns and relationships in the data.

   K-means clustering is one unsupervised learning technique that may be applied in this situation. Based on the attribute values, this method divides the dogs into K clusters, where K is a user-defined parameter. Each dog is progressively assigned by the algorithm to the cluster containing the closest centroid until

   Hierarchical clustering is another technique that can be utilized for unsupervised learning. Based on the values of their attributes, this algorithm arranges the dogs into a tree structure. Each dog is initially assigned to a separate cluster by the algorithm, which then repeatedly joins the closest clusters together until only one cluster remains. To summarize, since the researcher only knows attribute X for the dogs and no labeled data, unsupervised learning techniques like clustering algorithms can be utilized to investigate and characterize which species the dogs belong to. In this situation, two unsupervised learning techniques that might be applied are hierarchical and K-means clustering.

   2) Please justify your decision based on your understanding of unsupervised or supervised learning methods in this case study.
      Your Answer:

The decision to use unsupervised learning in this case study is because the researcher only has attribute X for the dogs and does not have any labeled data to train a supervised learning model.

Supervised learning methods require labeled data to train the model, which means the researcher would need a set of dogs with known species labels. However, in this case, the researcher does not have any labeled data, so supervised learning methods are not applicable.

Unsupervised learning methods, on the other hand, do not require labeled data. These methods can identify patterns and relationships in the data without the need for labeled data. In this case, the researcher can use unsupervised learning methods such as clustering algorithms to group the dogs based on their attribute values.

By using unsupervised learning methods, the researcher can explore and describe which species the dogs belong to without the need for labeled data. This can help the researcher gain insights into the relationships between the dogs and their attribute values, which can be useful for further research or decision-making.

In summary, the decision to use unsupervised learning in this case study is based on the fact that the researcher only has attribute X for the dogs and does not have any labeled data to train a supervised learning model. Unsupervised learning methods such as clustering algorithms can be used to explore and describe which species the dogs belong to without the need for labeled data.

3) What specific supervised or unsupervised learning methods/models you would like to propose to your supervisor
Your Answer:

As the researcher only has attribute X for the dogs and does not have any labeled data, unsupervised learning methods such as clustering algorithms would be the most appropriate choice.

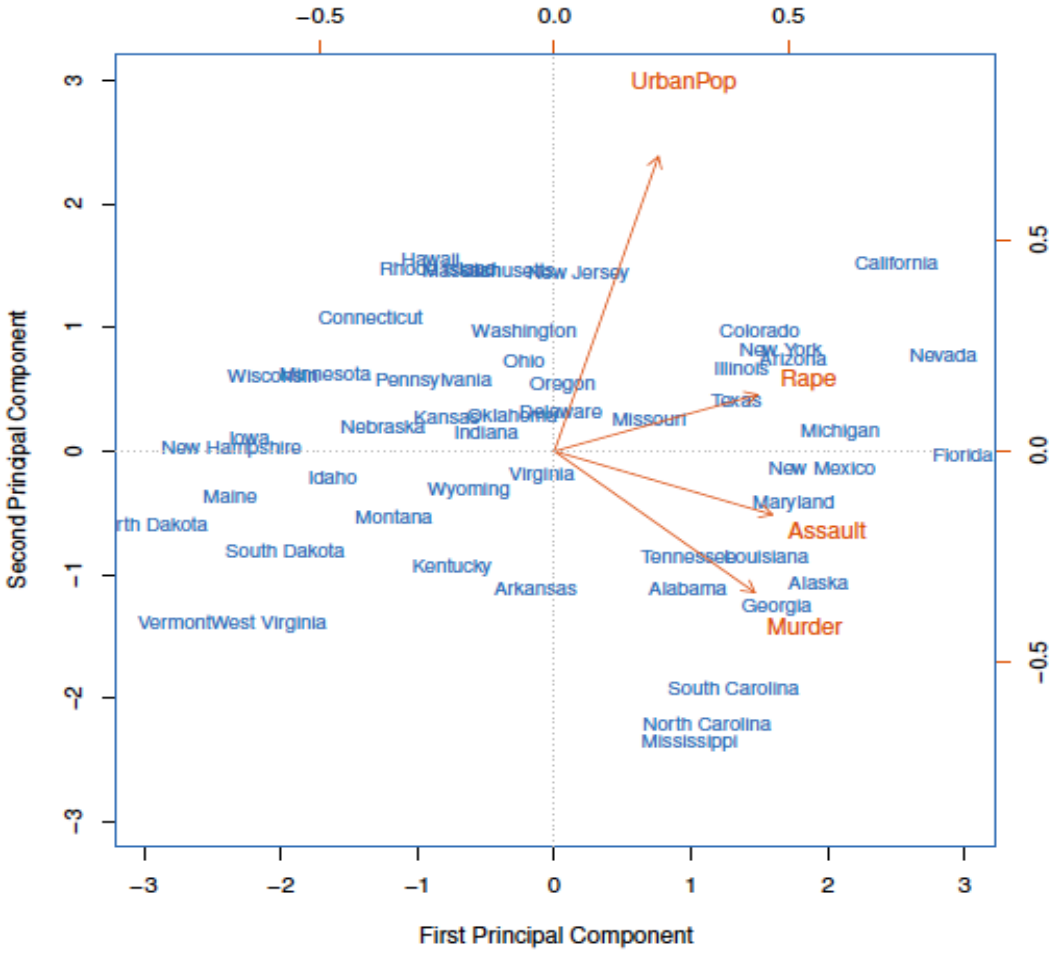
One specific unsupervised learning method that could be used in this case is K-means clustering. This algorithm groups the dogs into K clusters based on their attribute values, where K is a user-defined parameter. The algorithm iteratively assigns each dog to the cluster with the closest centroid until convergence is reached.

Another unsupervised learning method that could be used is hierarchical clustering. This algorithm groups the dogs into a tree structure based on their attribute values. The algorithm starts by assigning each dog to its own cluster and then iteratively merges the closest clusters until a single cluster is formed.

In summary, K-means clustering and hierarchical clustering are specific unsupervised learning methods that could be proposed to the supervisor for this case study. These methods can be used to explore and describe which species the dogs belong to without the need for labeled data.

4. **(15 points, 5 points*3)** Examine the loading matrix from the USarrests data and the corresponding biplot, then answer the following questions.

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186



1) What do the arrows indicate in the above bi-plot using USarrest data?
Your Answer:

The arrows in the bi-plot represent the loadings of the variables on the first and second principal components (PC1 and PC2). The length and direction of the arrows indicate the strength and direction of the relationship between the variables and the principal components. For example, a long arrow pointing to the right for a variable on PC1 suggests that the variable has a strong positive relationship with PC1.

2) Which three states have the highest crimes?
Your Answer:

The three states with the highest crimes, as indicated by the bi-plot, are likely to be those with the highest loadings on the first principal component (PC1), which is associated with the variables "Murder" and "Assault". The states with the highest loadings on PC1 are not explicitly labeled in the image, but they would be the ones with the longest arrows pointing to the right on PC1.

3) Which state has the lowest level of urbanization?
Your Answer:

The state with the lowest level of urbanization, as indicated by the bi-plot, would be the one with the lowest loading on the second principal component (PC2), which is associated with

the variable "UrbanPop". The state with the lowest loading on PC2 would be the one with the shortest arrow pointing to the right on PC2. However, without specific labels or numerical values, it's not possible to identify the exact state with the lowest level of urbanization from the image provided.

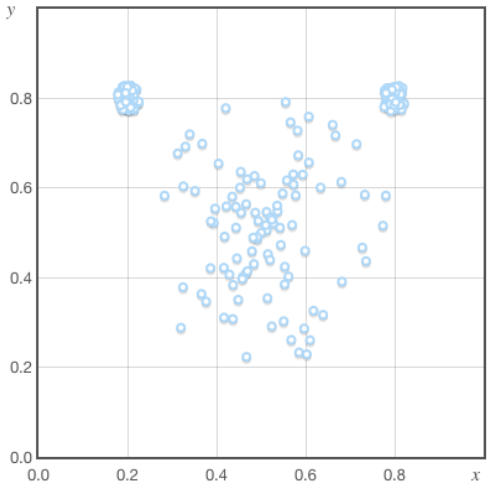
5. **(6 points; 3 points*2)** For K-means algorithm,

1) What is the immediate next step after the centroid initialization and each case is assigned to the initial clusters.

Your Answer:

After the centroid initialization and each case is assigned to the initial clusters in the K-means algorithm, the next immediate step is to recalculate the centroids of the newly formed clusters. This is done by taking the mean of all the points assigned to each cluster. The centroids are then updated, and the algorithm iteratively reassigns each point to the cluster with the nearest centroid until convergence is reached. This process of reassigning points and updating centroids continues until the centroids no longer change significantly or a maximum number of iterations is reached

2) Based on the data displayed in the graph below, is the K-means appropriate for this data? And why?



Your Answer:

K-means may not be the best approach for this data based on the image description for the following reasons:K-means may not be the best approach for this data based on the image description for the following reasons:

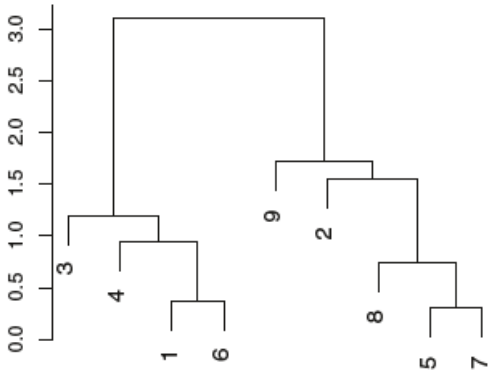
Overlap Between Clusters: Since some patterns are shared by multiple groups, there are no clear borders between the clusters.

Cluster Shape: Although the clusters in a k-means model are assumed to be spherical and equal in size, this is not the case.

Cluster Density: K-means accuracy will be impacted by the Next cluster's greater sparsity than the present one.

K-means produces these kinds of outcomes by presuming that sets of examples are distinctly different and do not overlap as well. If the lacuna in data is wrongly defined or take a complex curve shape, then it may be necessary to apply other clusters methods.

6. (Total 9 points; 3 points*3)



Examine the above dendrogram, and answer the following True or False questions:

1) The magnitude of similarity between Case 1 and 6 is similar to that between Case 5 and 7. True or False
Your Answer:
False

2) The magnitude of similarity between Case 3 and 4 is similar to that between Case 3 and 1. True or False
Your Answer:
False

3) The magnitude of similarity between Case 9 and 2 is significantly different from that between Case 9 and 7. True or False
Your Answer:
True