C3412 Final Exam

206 cards Deven J. Computer Science | Introduction To Computer Science

Practice all cards

True or False? When clustering, we want to put two data objects that are similar into the same cluster.	True
True or False? Cluster analysis is considered supervised learning.	False. Cluster analysis is by definition unsupervised learning.
True or False? It is impossible to cluster objects in a data stream. We must have all the data objects that we need to cluster ready before clustering can be performed.	False. Clustering algorithms can be adapted to perform clustering in a streaming fashion
True or False? Clustering analysis has a wide range of applications in tasks such as data summarization, dynamic trend detection, multimedia analysis, and biological network analysis.	True
True or False? Clustering analysis is unsupervised learning since it does not require labeled training data.	True
True or False? Clustering analysis has a wide	True

algorithms can be adapted to perform clustering in a streaming fashion.

False. This is false because, as discussed in the lecture, the objective of clustering is having similar objects in the same cluster.

True or False? K-means is an example of a distance-based clustering method.	• True
True or False? Dimensionality reduction helps make high-dimensional clustering more feasible and scalable.	True
True or False? Since cluster analysis is unsupervised learning, there's no way to incorporate user preference or guidance into the clustering process.	False
True or False? There are no clustering algorithms that can handle time-series data since we always assume that the data points are temporally independent from each other.	False
True or False? Graphs, time-series data, text, and multimedia data are all examples of data types on which cluster analysis can be performed.	True

True

False. This is false because we can also easily visualize clusters in 3D. HD-eye is an example of software used for visualizing higher dimensional clusters.

Clustering or Supervised Learning? Given the historical prices of a group of stocks, predict whether the price of a specific stock will rise or fall in the following week.	Supervised Learning. The task is more suitable to using supervised learning because there is available training data that allows us to learn classifiers for prediction.
Clustering or Supervised Learning? Find user communities in online social networks such as Facebook and Twitter	Clustering
Clustering or Supervised Learning? A real estate company wants to sell a new house; they need to determine the price for selling the house based on its condition (e.g., size, location), as well as the sales data of their previously sold houses.	Supervised Learning. The task is more suitable to using supervised learning because there is available training data that allows us to learn classifiers for prediction
Clustering or Supervised Learning? Given a large number of emails, we already know whether each email is spam or not. Now we need to determine whether a newly incoming email is spam.	Supervised Learning. The task is more suitable to using supervised learning because there is available training data that allows us to learn classifiers for prediction.

	to using supervised learning because there is available training data that allows us to learn classifiers for prediction.
latent topics discussed by those web pages.	Clustering.
Considering the k-means algorithm, after the current iteration we have three centroids (0, 1), (2, 1), and (-1, 2). Will points (0.5, 0.5) and (-0.5, 0) be assigned to the same cluster in the next iteration?	Yes
Considering the k-means algorithm, after the current iteration we have three centroids (0, 1), (2, 1), and (-1, 2). Will points (2, 3) and (-0.5, 0) be assigned to the same cluster in the next iteration?	No
Considering the k-means algorithm, after the current iteration we have three centroids (0, 1), (2, 1), and (-1, 2). Will points (2, 3) and (2, 0.5) be assigned to the same cluster in the next iteration?	Yes
True or False? The k-medoids and k-median algorithms are less sensitive to outliers than k-means.	True. Using a median or medoid will reduce the effect of a few severe outliers, making methods like kmedoids and k-medians less sensitive to outliers
True or False? The k-modes algorithm is	True. This is true since k-modes is designed for

	would be (1, 1.667) which is not an observed data point.
	False. In the k-medoids algorithm, the centroids are selected from the given data points.
True or False? The k-means algorithm is sensitive to outliers.	True. Outliers can have a significant impact on the mean of a cluster
True or False? The k-means algorithm can generate non-convex clusters.	False. k-means can only detect clusters that are linearly separable, while kernel k-means can detect some non-convex clusters.
True or False? For different initializations, the k-means algorithm will definitely give the same clustering results.	False. Different initializations may generate rather different clustering results (some could be far from optimal)
True or False? The centroids in the k-means algorithm may not be any observed data points.	True. Because the centroids are simply the average of all points in a cluster, this may not be a point in the cluster. For example, a cluster has points (1,2), (1,2), and (1,1), the centroid would be (1, 1.667) which is not an observed data point.
True or False? The k-means algorithm can directly handle non-numerical (categorical) data.	False. k-means requires numerical data to calculate the means.

	be far from optimal).
	True. Because the centroids are simply the average of all points in a cluster, this may not be a point in the cluster. For example, a cluster has points (1,2), (1,2), and (1,1), the centroid would be (1, 1.667) which is not an observed data point.
Is BIRCH hierarchical?	Yes
Is K-Means hierarchical?	No.
Is DBSCAN hierarchical?	No.
Is K-Mediod hierarchical?	No.
Is AGNES hierarchical?	Yes.
Is CHAMELEON hierarchical?	Yes.

False. BIRCH is sensitive to the insertion order of data points.

False. They can work with any similarity metric

True or False? CHAMELEON requires a graph as the input.

False. CHAMELEON constructs the kNN graph from a set of objects by measuring the distance between objects and linking each to the k nearest neighbors

True or False? BIRCH is better at capturing irregular shaped clusters than CHAMELEON.

False. Quite the opposite. By merging graphlets, CHAMELEON is able to capture complex shapes. CURE forms complex shapes by merging small disks. BIRCH tends to produce spherical clusters as limited by the diameter and radius parameters.

True or False? CHAMELEON requires a graph as the input.

False. CHAMELEON constructs the kNN graph from a set of objects by measuring the distance between objects and linking each to the k nearest neighbors.

True or False? Clusters in BIRCH are represented by a set of well-scattered representative points.

False. Representative points is the mechanism used by CURE

True or False? Each micro-cluster in BIRCH is represented by a clustering feature vector containing the number of points and the first and second moment of the points in the

True, as per the description of the algorithm.

	first step to produce the final clustering. CHAMELEON, on the other hand, uses a divisive mechanism to break the kNN graph into small graphlets, which are then merged to form the final clustering.
	False. While CHAMELEON and CURE are able to identify irregular shared clusters, BIRCH tends to produce spherical clusters.
True or False? Clustering results of BIRCH are sensitive to the insertion order of data points	True. BIRCH is sensitive to the insertion order of data points while constructing CF tree.
True or False?In the DBSCAN algorithm, given Eps and Minpts, if a point p is density-reachable from a point q, Point q is also density-reachable from p.	False. Incorrect when p is not a core point
True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is density-reachable from a point q, Point p must be directly density-reachable from q.	False. Incorrect when the chain of points include points other than p and q
True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is density-reachable from a point q, Point q is density-connected to p.	True. Correct since both p and q are density-reachable from q

True. Correct since p may not be core point

density reachable from q.

False. Incorrect since p must be density-reachable from q and the chain contains only p and q

True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is directly density-reachable from a point q, $\frac{14}{7}$ Point q is density reachable from p.

False. Incorrect since p may not be a core point

True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is directly density-reachable from a point q, ⌡ Point q is density-connected to p.

True. Correct since both p and q are density-reachable from the point q

True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is density-connected to a point q, Point p must be density-reachable from q

False. Incorrect. It is possible that neither p nor q is a core-point.

True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is density-connected to a point q, Point p may be density-reachable from q but not necessarily.

True. As long as p is density-reachable from a certain point o, from which q is also densityreachable, they are density connected. p could be but is not necessarily density reachable from q.

True or False? In the DBSCAN algorithm, given Eps and Minpts, if a point p is density-connected to a point q, Point q is directly

False. Incorrect. It is possible that neither p nor q is a core-point.

	True. Correct if neither p nor q is a core point
True or False? In the DBSCAN algorithm, suppose eps = 4cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 1cm. Points p and q may be in different clusters.	True. If (i) neither p nor q is a core point, (ii) there exists a core point o1 such that o1 is only density-reachable to p but not density reachable to q, and (iii) there exists a core point o2 such that o2 is only density-reachable to q but not density reachable to p
True or False? In the DBSCAN algorithm, suppose eps = 4cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 1cm. Points p and q must be in different clusters.	False. Incorrect since p and q might be noise/outliers and they do not belong to any clusters
True or False? In the DBSCAN algorithm, suppose eps = 4cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 1cm. Points p and q must be in the same cluster	False. Incorrect since p and q might be noise/outliers and they do not belong to any clusters
True or False? In the DBSCAN algorithm, suppose eps = 1cm and Minpts = 1. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Points p and q may be in different clusters	True. Correct. For instance, when p and q are the only two points in the dataset, since Minpts = 1, thus p and q form two clusters.
True or False? In the DBSCAN algorithm, suppose eps = 1cm and Minpts = 1. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Points p and	False. Incorrect when there exists a point o such that both p and q are density-reachable from. Note that any point in the dataset for Minpts = 1 is a core point.

False. Since p and q are both core points, it is impossible for them to be outliers, which implies p and q must belong to some clusters.

True or False? In the DBSCAN algorithm, suppose eps = 1cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Points p and q may not be in the same cluster.	True. Correct if there exists a core point that both p and q are density-reachable from o, then p and q are density-connected and belong to the same cluster. If there doesn't exist such a core point, then p and q may be noise/outliers or belong to different clusters.
True or False? In the DBSCAN algorithm, suppose eps = 1cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Points p and q may not belong to any clusters.	True. Correct since p and q may be noise/outliers such that they don't belong to any cluster.
True or False? In the DBSCAN algorithm, suppose eps = 1cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Points p and q must be in different clusters	False. Incorrect. p and q may be noise and do not belong to any clusters.
True or False? In the DBSCAN algorithm, suppose eps = 1cm and Minpts = 5. Randomly select two points p, q from the observed data points. We have dist(p, q) = 4cm. Points p and q must be in the same cluster.	False. Incorrect. p and q may be noise and do not belong to any clusters.
CLIQUE True or False? Any point that belongs to a cluster in a 1-D subspace A may not	True. Correct since the other dimension in a 2-D subspace may belong to sparse regions,

	that includes A.
	False. Incorrect since the other dimension in a 2-D subspace may belong to sparse regions, i.e., does not belong to any clusters
CLIQUE True or False? Any point that does not belong to a cluster in a 1-D subspace A may belong to some cluster (dense region) in any 2-D subspace that includes A.	False. Incorrect. For any data point, as long there is one dimension that belongs to the sparse region, the data point will not belong to any dense region in a higher dimension that includes A.
True or False? Modularity is an internal measure.	True
True or False? Conditional Entropy is an internal measure.	False
True or False? Jaccard Coefficient is an internal measure.	False
True or False? NMI is an internal measure.	False
True or False? Purity is an external measure.	True

	False.
	False.
True or False? F-Measure is a relative measure.	False.
True or False? Beta-CV is a relative measure.	False.
True or False? Modularity is a relative measure.	False.
True or False? Silhouette Coefficient is a relative measure.	True.
True or False? True Negative is a relative measure.	False.
Applications of Cluster Analysis	• An intermediate step for other tasks • Data summarization, or compression

	• Density-based • Grid-based • Probabilistic & Generative Models
	A matrix of n data points (rows) with 1 dimensions (columns)
Dissimilarity (Distance) Matrix	n data points of distance. Usually a symmetric triangle matrix.
Minkowski Distance	d(i,j) = Prt(abs(xi1-xj1)^P + abs(xi2 - xj2)^P)
A metric	A distance that exemplifies: • positivity - d(i,j) > 0 if i is not equal to j and d(i, i) = 0 • symmetry - d(i,j) = d(j,i) • Triangle Inequality - d(i,j) less then or equal to d(i, k) + d(k, j)
Manhattan Distance	d(i,j) = abs(xi1-xj1) + abs(xi2-xj2) + ...
Supremum Distance	The max difference between any component of the vectors
Symmetric Binary Variables	Roughly the same chance of either variable being true (1).

	m = # of matches p = total number of variables d(i,j) = (p-m)/p
	Have order, and can be discrete or continuous.
Cosine Similarity	cos(d1, d2) = (d1 dot d2)/(sqrt(length of d1) * sqrt(length of d2))
Variance	provides a measure of how much a value of x deviates from the mean or expected value of x.
Partitioning-Based Clustering	Discovering groupings in the data by optimizing a specific objective function and iteratively improving the quality of the partitions.
K-Partitioning Method	Partitioning of a dataset D of n objects into a set of K clusters so that an object function is optimized
Sum of Square Errors	SSE(C) = for each cluster(for each point in that cluster(abs(xi - ck)^2)) Where ck is the centroid or mediod of the cluster.
K Means Algorithms	• Select k points as initial centroids • for k clusters by assigning each point to it's nearest centroid

	False.
	True.
K-Means++ Initialization	• Select first centroid randomly • Select next centroid as the one that is farthest from the current one. • Continue until all centroids are picked.
K-Mediods Algorithm	• Select K points as initial representative objects • Assign ea point to closest mediod • Compute the cost of swapping mediods with random point O • If the cost is lower than the current mediod, switch • Repeat until converged
PAM (Partitioning Around Mediods) Algorithm	• Start with initial mediod set • Iteratively replace one mediod by one non-mediod if it improves total sum of SSE of the clustering • This is expensive.
Is K-Mediods less sensitive to outliers than K-Means?	Yes.

	feature instead.
	Yes.
	it projects the data onto high-dimensional kernel space and then performs k-means on that data. This is more computationally complex.
Agglomerative Clustering	Start with singleton clusters and continuously merge two clusters at a time to build a bottom-up hierarchy of clusters.
Divisive Clustering	Start with a huge macro-cluster and split it continuously into two groups generating a top-down hierarchy of clusters.
AGNES	• Use the single-link method • Continuously merge nodes with least dissimilarity • Eventually all nodes belong to some cluster
Single Link Similarity	Similarity between clusters is based on the similarity between their closest members. Capable of clustering non-elliptical shaped groups of objects. Sensitive to noise and outliers

	average similarity between all elements in the clusters. This is expensive to compute.
	Cluster similarity is based on the distance between centroids.
	• N is the cardinality of the cluster • C is the centroid of the cluster • (NaCa + NbCb)/Na + Nb
Wards Criterion	The increase in the value of the SSE criterion for the clustering obtained by merging them into a new cluster.
DIANA	Divisive Clustering Algorithm. Does the inverse of AGNES. More efficient than agglomerative clustering.
BIRCH	A micro-cluster approach to clustering. Uses a feature (CF) tree for multi-phase clustering. Scales linearly.
BIRCH - Phase 1	Scan the DB to build an initial in-memory CF-Tree.
CF-Tree	A multilevel compression of the data that tries to preserve the inherent clustering structure of the data.

a Cl r	The middle of the cluster. The sum of all points over the total number of points.
	The sqrt of the average distance from member objects to the centroid.
	The sqrt of the average mean square distance between all pairs of points in the cluster.

CF-Tree Parameters	• Branching Factor - Max # of children • Max DIameter of sub-clusters stored at leaf node
BIRCH Concerns	• Sensitive to insertion order of data points • Due to the fixed size of leaf nodes, clusters may not be natural • Clusters tend to be spherical given the radius and diameter measures
CURE	Clustering using well-scattered points. Incorporates features of both single and average link. Captures clusters of arbitrary shapes. More robust for outliers.


Chegg