Feedback — Week 3 Quiz

Help Center

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Tue 12 May 2015 2:48 AM PDT**. You got a score of **6.00** out of **6.00**.

Question 1

Which of the following measures can be used as external measures for clustering validation?

Your Answer		Score	Explanation
☑ F-measure	✓	0.20	
□ Silhouette coefficient	✓	0.20	
□ Beta-CV measure	✓	0.20	
☑ Purity	✓	0.20	
□ Normalized cut	✓	0.20	
Total		1.00 / 1.00	

Question Explanation

Beta-CV measure and Normalized cut are internal measures. Silhouette coefficient is a relative measure.

Question 2

The following table summarizes the clustering results of a newly designed algorithm where C_1, C_2, C_3 denote the clusters while T_1, T_2, T_3 are ground truth. Based on the table below, calculate the purity of the clustering algorithm.

C\ T	T_1	T_2	T_3	Sum
C_1	20	30	10	60
C_2	30	40	10	80

C_3	0	0	60	60
m_j	50	70	80	200

Your Answer		Score	Explanation
○ 0.6			
○ 1			
○ 0.35			
● 0.65	✓	1.00	
○ 0.667			
Total		1.00 / 1.00	

Question Explanation

Since there are three partitions, based on the definition of purity, we have that

$$Purity = \frac{60 \times \frac{30}{60} + 80 \times \frac{40}{80} + 60 \times \frac{60}{60}}{200} = 0.65.$$

Question 3

The following table summarizes the clustering results of a newly designed algorithm where C_1, C_2, C_3 denote the clusters while T_1, T_2, T_3 are ground truth. Based on the table below, calculate the maximum matching score of the clustering algorithm.

C\ T	T_1	T_2	T_3	Sum
C_1	10	40	10	60
C_2	20	10	30	60
C_3	30	0	50	80
m_j	60	50	90	200

Your Answer	Score	Explanation
○ 0.597		
○ 0.45		

○ 0.6			
● 0.55	✓	1.00	
○ 1			
Total		1.00 / 1.00	

Question Explanation

Based on the definition of maximum matching, we have the following maximum matching schemas:

(C1 – T2), (C2 – T1), (C3 – T3)

Based on this maximum matching, we have

$$match = \frac{60 \times \frac{40}{60} + 60 \times \frac{20}{60} + 80 \times \frac{50}{80}}{200} = 0.55.$$

Question 4

The following table summarizes the clustering results of a newly designed algorithm where C_1, C_2 denote the clusters while T_1, T_2 are ground truth. Which of the following statements are correct?

C\ T	T_1	T_2	Sum
C_1	8	2	10
C_2	3	7	10
m_j	11	9	20

Your Answer		Score	Explanation
□ The false positive is 41.	✓	0.12	
☑ The false positive is 37;	✓	0.12	
□ The true positive is 49.	✓	0.12	
□	✓	0.12	

The true negative is 58.		
☑ The true negative is 62.	✓	0.12
☑ The true positive is 53.	✓	0.12
☑ The false negative is 38;	✓	0.12
□ The false negative is 42.	✓	0.12
Total		1.00 / 1.00

Question Explanation

1. The true negative is $TN = \sum_{i=1}^{2}\sum_{j=1}^{2}\binom{n_{ij}}{2} = \binom{8}{2}+\binom{2}{2}+\binom{3}{2}+\binom{7}{2} = 28 + 1 + 3 + 21 = 53;$
2. The false negative is $FN = \sum_{j=1}^{2}\binom{m_j}{2} - TP = 91 - 53 = 38;$
3. The false positive is $FP = \sum_{i=1}^{2}\binom{n_i}{2} - TP = 90 - 53 = 37;$
4. The true negative is $FP = N - TN - FN - FP = \binom{20}{2} - TN - FN - FP = 62.$

Question 5

Which of the following clustering methods is suitable for clustering high-dimensional data?

Your Answer		Score	Explanation
○ DBSCAN			
● CLIQUE	✓	1.00	
○ K-means			
○ BIRCH			
Total		1.00 / 1.00	

Question Explanation

Due to the curse of dimensionality, the distance measure used in k-means, DBSCAN and BIRCH becomes meaningless. Thus, these methods cannot be applied to high-dimensional data. CLIQUE is a subspace clustering method that might find clusters in subspaces.

Question 6

Suppose we are going to divide 40 students in a class into two groups to play a game. Some of the students are friends, and thus, they want to be in the same group. Which of the following two strategies is better for grouping students according to their friendships?

Your Answer		Score	Explanation
● Take the friendships as hard constraints (Must-Link), and apply a constraint-based clustering algorithm, such as COP-k-means, to cluster students based on their other information.	✓	1.00	
○ Starting from some user specified features, for example, nationality and hobbies, apply k-means algorithm to group the students.			
Total		1.00 / 1.00	

Question Explanation

The friendships in a group of 40 students are not very large and easy to obtain. In this case, must-links as hard constraints are feasible and will make sure all friends are in the same cluster. Using the user specified features is good; however, it may still place the friends in different clusters. Therefore, must-link is a better plan.