Review for Final Exam

COSC 3337

Dec. 1, 2022

**1) Association Rule and Sequence Mining [9]**

a) Assume we have the following Transaction Database

T1: {A,B,C,D}

T2: {A,C,D,E}

T3: {C,D,E,F}

T4: {B,C,D,E}

T5: {A,D,E}

What is the support and confidence the following association rule:

IF (C and D) THEN E? [3]

Support = 3/5 [1.5]

Confidence=3/4 [1.5]

b) Assume the APRIORI algorithm identified the following five 4-item sets that satisfy a user given support threshold: **abcd, acde, acdf, acdg adfg;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [4]

acdef, acdeg, acdfg [3] One error: at most one point!

None survives pruning [1]

c) Why are association rule mining systems interested in finding rules with high support? [2]

Rules with high support are more likely to predict the occurrence of an item based on the occurrences of other items in the transaction accurately; it is hard to learn accurate rules from just a few examples.

d) Assume an association rule if smoke then cancer has a confidence of 86% and a high lift of 5.4. What does this tell you about the relationship of smoking and cancer? [2]

Con = 86% 🡪 86% people who smoke tend to get cancer; that is P(Cancer|Smoke)=0.86

Lift = 5.4 🡪 Smoking increases the probability of getting cancer by a factor of 5.4; that is, P(Cancer|Smoke)/P(Cancer)=5.4

 e) Assume the Apriori-style sequence mining algorithm introduced in the lecture is used and the algorithm generated 3-sequences listed below; what candidate 4-sequences are generated from this 3-sequence set? Which of the generated 4-sequences survive the pruning step?

Frequent 3-sequences Candidate Generation Candidates that survived pruning

Candidates that survived pruning:

<(1) (2) (3) (4)>

Candidate Generation:

<(1) (2) (3) (4)> 🡪 survived

<(1 2 3) (4)> 🡪 pruned, (1 3) (4) is infrequent

<(1) (3) (4 5)>🡪 pruned (1) (4 5) is infrequent

<(1 2) (3) (4)>🡪 pruned, (1 2) (4) is infrequent

<(2 3) (4 5)>🡪 pruned, (2) (4 5) is infrequent

<(2) (3) (4 5)>🡪pruned, (2) (4 5) is infrequent

<(1) (2) (3)>

<(1 2 3)>

<(1) (2) (4)>

<(1) (3) (4)>

<(1 2) (3)>

<(2 3) (4)>

<(2) (3) (4)>

<(3) (4 5)>

2. EM

a) What cluster models does EM use

Each cluster is described by:

a. a mean value

b. a covariance matrix

c. a cluster prior/weight (weights of the k clusters have to add up to one)

[Gaussian Mixture Models — PyPR v0.1rc3 documentation (sourceforge.net)](https://pypr.sourceforge.net/mog.html)

b) How does EM determine if a point i belongs to a cluster j

$$p\left(x\_{i}\right)=\frac{p\left(C\_{j}\right)p(C\_{j})}{\sum\_{l=1}^{k}p\left(C\_{l}\right)p(C\_{l})}$$

3. Fuzzy C-Means (FCM)

a. How is FCM different from K-means

FCM uses soft cluster memberships expressed in weight wij which can be interpreted as probability of object i belonging to cluster j; that is, objects have to belong to exactly one cluster, as it is the case with k-means. FCM uses weight based computations to determine the cenrtoid.

b. How does FCM update the weights in its iterations

Let us assume we run FCM for K=2 and the centroids are cluster 1=(1,1) and cluster 2=(2,3) and hyper parameter p is 2 and we use Manhattan distance; furthermore point i is: (1,4) in this case;

Wi1= 1/3\*\*2/(1/9+1/4)=0.309

Wi2= 1/2\*\*2)/1/9+1/4)=0.692

$$w\_{ij}=(1/dist(x\_{i},c\_{j})^{2})^{\frac{1}{p-1}}/\sum\_{q=1}^{k}(1/dist(x\_{i},c\_{q})^{2})^{\frac{1}{p-1}}$$

4) Cluster Evaluation

Compute the Silhouette for the point (2,3) for the following clustering that consists of 2 clusters:

{(0,0), (0,1), (2,3)}, {(3,3), (3,4)}; use Manhattan distance for distance computations. Interpret the result! [4]

Average Distance of (2,3) to the points in C2: (1+2)/2=1.5

Average Distance of (2,3) to points in its own cluster: (4+5)/2=4.5

s = (b – a) / max(a,b)

Silhouette((2,3))=(1.5-4.5)/4.5=-2/3

If any errors: at most 1 of 3 points for computing Silhouette((2,3)!

Interpretation: The silhouette coefficient for point (2,3) is very bad as the point should be in the other cluster [1]

b) How does external cluster evaluation work?

Clusters will be evaluated how well they are able to match existing class labels; similar to decision trees: entropy, GINI or purity will be used as evaluation measures.

5) K-means

Assume the following dataset is given: (1,1), (2,2) (4,4), (5,5), (4,6), (6,4) . K-Means is used with k=2 to cluster the dataset. Moreover, Manhattan distance is used as the distance function (formula below) to compute distances between centroids and objects in the dataset. Moreover, K-Means’s initial clusters C1 and C2 as follows:

C1: {(1,1), (3,3), (4,4), (6,6)}

C2: {(6,4), (4,6)}

Now K-means is run for a single iteration; what are the new clusters you obtain[[1]](#footnote-1) [4]

**d((x1,x2),(x1’,x2’))= |x1-x1’| + |x2-x2’| Manhattan Distance**

centroid C1= (3.5,3.5}

centroid C2= {5,5}

New Clusters

C1={(1,1), (3,3), (4,4)}

C2={(6,6},(4,6), (6,4)}

**6) Expect one more essay-style question in the final exam**

**Important: this is an essay: write complete sentences!**

**e.g What skills are important to be hired as a Data Scientist?**

(see slides that discuss this topic)

* Should know R and/or Phyton
* Should have sound software development skills
* Should have some sound knowledge of Statistics
* Should have sound knowledge of the different data analysis tasks; e.g. clustering, classification, similarity assessment
* Should be knowledgeable in data visualization
* Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.”
* The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades."
* But what’s even harder is finding people who have those skills *and* are good at communicating the story behind the data.”
1. If there are any ties, break them whatever way you want! [↑](#footnote-ref-1)