Review for Final Exam

COSC 6335

Nov. 30, 2023

1) Similarity Assessment

Design a distance function to assess the similarity of gradute students; each customer is characterized by the following attributes:

1. Ssn
2. qud (“*quality of undergraduate degree*”) which is ordinal attribute with values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’, ‘very poor’.
3. gpa (which is a real number with mean 2.8 standard deviation is 0.8, and maximum 4.0 and minimum 2.1)
4. gender is an nominal attribute taking values in {male, female}.

Assume that the attributes qud and gpa are of major importance and the attribute gender is of a minor importance when assessing the similarity between students. Using your distance function compute the distance between the following 2 students: c1=(111111111, ‘good’, 2.9, male) and c2=(222222222, ‘very poor’, 3.7, female)!

We convert the Oph rating values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’, ‘very poor’ to 5:0 using **φ; then we compute the distance** by taking L-1 norm and dividing by the range, 5 in this case.

Normalize gpa using Z-score and find distance by L-1 norm

dgender(a,b):= if a=b then 0 else 1

Assign weights 1 to qud, 1 to Power-used and 0.2 to Gender

Now[8]: one error: 2.5-5 two errors: 0-2 distance functions not properly defined: at most 3 points

**d(u,v) = (1\*|(u.gpa)/0.8 – (v.gpa)/0.8| + 1\*|φ(u.qud) – φ(v.qud)|/5 + 0.2\*dgender(u.gender, v.gender)) /2.2**

2 students: c1=(111111111, ‘good’, 2.9, male) and c2=(222222222, ‘very poor’, 3.7, female)!

d(c1,c2)= (1 + 3/5 + 0.2)/2.2= 1.8/22=9/11=0.82 [2]

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 centroid.

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) Miscellaneous Questions [19]**

a. What are the characteristics of overfitting when learning decision trees? What can be done to deal with overfitting? [3]

1. reduce the size of the tree [1.5]

2. increase the size of the training set by adding new training examples [1.5]

b) A confusion Matrix of a classification model for distinguishing dogs, cats and rabbits is given below:

What is the accuracy of the classification model; what is its precision for class rabbit? What is its recall for class rabbit? It is okay to represent your answers as fractions; e.g. 17/36! [3]



Accuracy=23+29+24/(total number of examples)

Precision rabbit: 24/(24+10+4)=24/38

Precision rabbit: 24/(24+13+7)

**5) Autoencoders**

a) What role do Kullback–Leibler (KL) divergences play in Variational Autoencoders (VAEs)?



i. KL-divergences measure the distance between two distributions e.g. how close are N(0.3,1.3) and N(5,13) to N(0,1); obviously,

dKL(N(0.3,1.3),N(0,1))<< dKL(N(5,13),N(0,1))

ii. KL-divergences are used in VAE loss functions to create a penalty that is proportional to how much the latent vector deviates from an assumed prior (e.g. from N(0,1) or from a covariance matrix which has 1 in the diagonal and 0 everywhere else[[1]](#footnote-1)); this accomplish some regularization of the latent space and also can be used to enforce independence of the different latent variables)

b) How can autoencoders be used for outlier detection?

Steps:

i. Learn an Autoencoder Y for your Dataset D

ii. Feed all example d∈D into Y and add the reconstruction loss (d(x,r)) in an additional column to d.

Remark: The newly added column serves as an outlier score (OLS); the larger d(x,r) is the more likely x is an outlier

c) If I decrease the dimensionality of the latent vector h what will be the consequences?

The reconstruction loss will increase.

**6) 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

7) 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[[2]](#footnote-2) [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)}

**8) 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.”

Remark: Other final exams topics have been discussed in the GHC presentations given in the last weeks.

1. No correlation between variables. [↑](#footnote-ref-1)
2. If there are any ties, break them whatever way you want! [↑](#footnote-ref-2)