Final Exam

COSC 3337 *Data Science I*

December 7, 2023

Your Name:

Your student id:

Problem 1 --- Outlier Detection [9]

Problem 2 --- Association Analysis [5]

Problem 3 --- Clustering [14]

Problem 4 --- Ethics for Data Science [10]

Problem 5 --- Autoencoders [11]

Problem 6 --- Similarity Assessment [9]

Problem 7 --- Parametric Density Estimation and EM [12]

70:

**Grade:**



The exam is “open books” and you have 105 minutes to complete the exam. The exam will count approx. 20% towards the course grade. The use of computers and cell phones is strictly prohibited; using calculators is fine.

Write your answers on the exam paper; if you need more space use back of exam paper!

1. **Outlier Detection [9]**

a) Assume you plan to design a system for anomaly detection which uses VAEs. Describe in 4-6 sentences the architecture of such a system and how it would solve the task at hand! [6]

no solution given

b) Assume you plan to develop a system for distance-based object outlier detection which relies on using k-nearest neighbor distances. Describe how such a system would assign outlier scores to the objects in a dataset! [3]

*Basically they have to say that they associate outlier score with each object which is the distance to the k-nearest neighbor of the object.*

**2. Association Analysis [5]**

Assume the APRIORI algorithm identified the following 8 4-item sets that satisfy a user given support threshold: **acdf, acdg, adfg, bcde, bcdf, cdef, defg;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning? [5]

Candidate 5-items that are created are [3]; if they give a different answer at most 1 point partial credit.

acdfg

bcdef

As cdfg and bdef are not frequent 4-items neither 5-item set survives pruning [2]

**3) Clustering [14]**

a) Let us assume we run Fuzzy C-means (FCM) for K=2 and the centroid for cluster 1 is (1,1) and the centroid of cluster 2 is (2,3) and hyper parameter p is 2 and we use Manhattan distance; furthermore, point i is: (2,1). Compute the wi1 and wi2 for point i! [4]

Wi1= 1/1\*\*2/(1+1/4)=0.8

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

W*i*2=1/2\*\*2)/(1+1/4)=0.2

If error at most 1 point for partial credit

b) Assume we use FCM for 4 points and k=2 and the points and their weights are as follows:

Point 1: (2,2) with w11=0.8 and w12=0.2

Point 2: (3,3) with w21=0.7 and w22=0.3

Point 3: (8,9) with w31=0.1 and w32=0.9

Point 4: (12,13) with w41=0 and w42=1

Using the methods FCM uses, compute the centroid of cluster 1; give the formula and its vector. [4]

Centroid1= (0.8\*(2,2)+ 0.7\*(3,3)+0.1\*(8,9))/1.6=

((1.6+2.1+0.8)/1.6,(1.6+2.1+0.9)/1.6)=

(4.5/1.6,4.6/1.6)=(2.81,2.88)

Can give them 2.5 points if they use the correct formula and have a major calculation error and 3 points if they had a minor calculation error; 3.5 points if their answer is (4.5/1.6,4.6/1.6). At most 1 point, if they use the wrong formula.

c) Assume we apply the CLIQUE algorithm to a numerical dataset with attributes A, B, C, D and E. What is the main difference between CLIQUE and traditional Clustering algorithms such as K-means with respect to the clusters CLIQUE finds? How does CLIQUE take advantage of the APRIORI principle? How does CLIQUE form clusters? [6]

*Find clusters in the subspace rather in the complete space A-B-C-D-E-F [2]*

*K+1 dimensional grid-cell candidates are computed from K-dimensional grid-cell which are dense (the number of points they contain is above the density threshold. [2]*

*Clusters in subspaces are formed by a growing algorithm which starts with a seed grid-cell and adds neighboring grid-cells [2]*

Other answers might deserve full or partial credit!

**4) Ethics for Data Science [10]**

Related to Video: *Statistical Imaginaries* by Danah Boyd

1. The US Census Bureau has been integrating *differential privacy* into its scientific products since 2006. This was cheered on by statisticians but displeased the public
	1. List three of the key assumptions behind differential privacy. [3]
	2. Why was reaction of the public to the revelation of differential privacy quite negative? [3]

No answer given

b. Danah Boyd refers to the following assumption “In most technical communities, it’s easy to think of statistics as objective, scientific, mathematical work.” in her video. Danah Boyd gives several examples from both inside and outside US why this assumption about census data is misleading. Describe two of them! [4]

No answer given

**5) Autoencoders [14]**

a) How are variational autoencoders different from traditional auto-encoders (VAEs)? [2]

No answer given

b) Kullback–Leibler (KL) divergences are used in the loss functions which Variational Autoencoders (VAEs) employ. What is their impact on the models VAEs learn? [4]

*The goal is to penalize latent space representations which deviate from from simple prior distribution; e.g N(0,1) or a n-dimensional distribution with 1’s in the diagonal and 0’s off the diagonal of the covariance).*

If they say that this approach is a form of regularization to obtain simple models, without mentioning priors you can give them up to 3 points for their answer.

c) The VAE learns the parameters of a distribution which then is used to generate encoded embeddings. However, this process of sampling from a distribution that is parameterized by our model is not differentiable.

1. Explain how the reparameterization trick solves this problem. [3]
2. What would you need to change if the sampling distribution is not Gaussian? [2]

No answer given

**6) Similarity Assessment [9]**

Design a distance function to assess the similarity of customers of a supermarket; each customer in a supermarket is characterized by the following attributes[[1]](#footnote-1)[7]:

1. Ssn
2. Items\_Bought (The set of items the bought last month)
3. Amount\_spend (Average amount spent per purchase in dollars and cents; it has a mean of 40.00 a standard deviation of 20, the minimum is 0.05 and the maximum is 600)
4. Age (is an ordinal attribute taking 6 values: child, young, teenager, medium, old, very\_old)

Assume that Items\_Bought and Amount\_Spend are of major importance and Age is of a minor importance when assessing the similarity of the customers.

Using the distance function you designed, compute the distance between the following 2 customers: c1=(11111111, {Coke, Oranges, Apples, Hummus}, 23.00, ‘old’)

and

c2=(222222222, {Hummus, Watermelon, Steak, Onions, Potatoes, Newspaper}, 33.00,’young’) [2]!

Let φ be a function which maps child, young, teenager, medium, old, very\_old to 0:5 and ‘|’ is assumed to be the set cardinality operator; e.g. |{1,3}|=2. Use z-scores to normalize the Amount\_Spent attribute.

Solution:

d(u,v)= (1\*(1- (|u.Items\_Bought ∩ v.Items\_Bought|/(|u.Items\_Bought ∪ v.Items\_Bought)|)) + 1\*|u.Amount\_Spend-v.Amount\_Spend|/40 + 0.2\*|φ(u.Age)-φ(v.Age|/5)/2.2

Remark: The fact that Age is less important is captured in d by using a weight of 0.2 age distances.

If they do not define a mathematical sound distance function at most 2.5 points. You can still give partial credit for correct pieces of the density function. 3 points for the first part and 2 points for part 2 and 3.

7) Parametric Density Estimation and EM [12]

a. Assume you apply EM to a 5-dimensional dataset D; list the model parameters that EM learns for D! [3]

Let k be the number of clusters; input when running EM

k cluster means μr for r=1,..,k [1]

k 5X5 covariance matrices Σr for r=1,..,k [1]

k cluster weights/priors/mixing factors πr for r=1,..,k [1]

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

b. What does the E-step of the EM algorithm compute and what formula does it use in its computation? [4]

It computes the probability that the i-th objects belongs to the j-th clusters (for each object/cluster pair) [2]

Formula used[2]:

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

c. How does EM compute the value of $p\left(C\_{j}\right)?$[3]

By plugin xi into the density functions of the j-th Gaussian model with parameters μj and Σj.

p(xi)= *Ν*(xi|μj,Σj)

with



d. Assume you apply EM to the 5-dimensional dataset D; what does the M-step compute? [2]

It computes an updated Gaussian mixture model[2] (in the form mentioned in the answer to question a).

1. E.g. (111234232, {Coke, 2%-milk, apple}, 42.42, ‘medium’) is an example of a customer description. [↑](#footnote-ref-1)