COSC 3337

Review on October 29, 2024

for the Oct. 31 Midterm2 Exam

Also take a closer look at the recent GHC presentations which gave a review of other material which is relevant for Midterm2.

**1) Similarity Assessment**

 Design a distance function to assess the similarity of electricity company customers; each customer is characterized by the following attributes:

1. Ssn
2. Oph (“*on-time payment history*”) which is ordinal attribute with values ‘excellent’, “very good’, ‘good’, ‘medium’, and ‘poor’.
3. Power-used (which is a real number with mean 2000, standard deviation is 1000, its maximum is 10000 and minimum 100)
4. Country\_of\_Citizenship is a nominal attribute

Assume that the attributes Oph and Power-used are of major importance and the attribute Country\_of\_Citizenship is of a minor importance when assessing the similarity between customers. Using your distance function compute the distance between the following 2 customers: c1=(111111111, ‘excellent’, 2000, ‘Peru’) and c2=(222222222, ‘good’, 2500, ‘France’)!

Let ψ be the following function: ψ(excellent)=1, ψ(very good)=3/4, ψ(good)=1/2, ψ(medium)=1/4, ψ(poor)=1.

doph(a,b)= |ψ(a)- ψ(b)|

dp-u(a,b)= |(a-b)/1000|

dccit(a,b)= If a=b then 0 else 1

Let o1=(ssn1,oph1,p-u1,ccit1) and o2=(ssn2, oph2,p-u2,ccit2) be two customers:

d(o1,o2)= (doph (oph1,oph2) + dp-u(p-u1,p-u2) + 0.2\*dccit(ccit1, ccit2))/2.2)

Example: d(c1,c2)= (1/2+1/2+0.2)/2.2=1.2/2.2=0.545

Take a look at GHC presentation October 24!

**2. K-Means and K-Medoids/PAM and Clustering in General**

1. Assume we apply K-medoids for k=2 to a dataset consisting of 4 objects numbered 1,..,4 with the following distance matrix:

0 6 5 2 🡨object1

 0 4 3

 0 1

 0 (e.g. the distance between object 2 and 4 is 3)

 The current set of representatives is {3,4} (objects 3 and 4); indicate all computations k-medoids (PAM) performs in its next iteration! Does k-medoids get a new set of representatives or does it terminate in the next iteration? [6]

RS={3,4} clusters: {3} (1,2,4} SEE=2\*\*2+3\*\*2

New Represnetative sets are created

{1,4} …. SSE=3\*\*2+1\*\*2

{2,4} {2} {1,3,4} SSE=2\*\*2+1\*\*2

{1,3} …SSE=4\*\*2+1\*\*2

{2,3} … SSE=5\*\*2+1\*\*2

The SSE decreased and therefore PAN will run for another generation for the “new” representative set {2,4}

One error: at most 3.5 points; 2 errors at most 1 point.

b) 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)}

One error at most 1.5 points; 2 errors: 0 points

Problem 2 continued

c) Compare k-means with Hierarchical clustering; what are the main differences in the way they are forming clusters and in general? [4]

K-Means creates a single clustering and HC creates a hierarchy of object sets; that is, multiple clusterings [2]

HC creates a dendrogram by merging the closest clusters[1]; K-means creates clusters by assigning the objects in a dataset to the closest centroid [1]

**3. Density Estimation**

a. How do parametric density estimation techniques find the parameters of models they try to fit to a dataset? For example, if we fit a Gaussian Model to a 1D dataset how does this approach choose the mean value μ and the standard deviation σ? [4]

The idea is to select parameters; e.g. value μ and the standard deviation σ in the case of a Gaussian 1D- distribution---which maximizes the probability of the examples in D: Maximize the sample that is: ∏d∈D P(d|μ,σ)

where P is the density function of the distribution whose parameter need to be selected.

Other solutions might deserve credit!

b) Assume a dataset O={x1,x2,x3} with data points x1=(1,2), x2=(5,7), x3=(7,7), is given; moreover, assume Manhattan distance[[2]](#footnote-2) is used as the distance function and q1=(6,6) is a query point. Compute fGauss (q1) assuming bandwidth σ=1! [4]

Remark: it is okay to use an expanded formula as your answer; e.g “e-12 + e--2.5…” as your answer; it is not necessary to report the exact value!

fGauss ((6,6)) = e-81/2 + e-4/2  + e-4/2= e-81/2+ 2\* e-2

Solutions which use the normalized 2D KDE function also deserve full credit.

**4. Hierarchical Clustering**

Take a look at GHC presentation October 24!

****

Other ways to specify dendrograms through Merge history in Midterm2:

1. {A} and {B}
2. {E} and {F}
3. {E,F} and {D}
4. {D,E,F} and {C}
5. {A,B} and {C,D,E,F}

**5. Outlier Detection**

 How does model-based outlier detection work? Limit your answer to at most 4 sentences! [4]

Basic points they should mention: a. fit a model M to the dataset [2] b. using the model D compute the density of each point o in the dataset and use this density as the outlier score [2].

1. If there are any ties, break them whatever way you want! [↑](#footnote-ref-1)
2. d((x1,y1),(x2,y2))= |x1-x2| + |y1-y2| [↑](#footnote-ref-2)