Midterm
Computer Science
Fall 2015
B565

Professor Dalkilic
Sunday, October 25, 9:00 p.m.

October 21, 2015

Directions

For question 1, you'll provide as much explanation as you deem necessary to solve the problem of the client. For 2, you'll simply supply the algorithm and discussion. For 3, you'll provide the mathematics and answer. A bonus of *up to* 50pts *may* be added for presentation and readability. You cannot discuss the exam with anyone. If you think you've found something wrong with any of the problems, *fix them* and write about. No changes will be made to the exam while it's being worked on. You are free to use materials, but must *explicitly* site them if they contribute significantly to your answer. Most importantly, think creatively and positively as you work through this. Good luck!

1 Application [250pts]

Assume you work for `Filmflix.com`, an online movie rental business. The data from `Filmflix.com` is found in Table 1. Read the dialogue and answer appropriately.

1. The client says to you, "We'd like to understand our customers broadly from their tastes in movies. What do you suggest?"
   "I'll do a k-means and show you the results!

2. The client says to you, "I was reading about association rules. These rules could help us promote certain movies. What does the data suggest?"
   "Well, we can do either genre or movies–or even both. Why don't I generate a couple of rules for each individually and suggest some promotions."

3. The client says to you, "My favorite movies are 2,5,8, and 9." I'm curious, what customers is nearest to me in my choices?"
   "I suggest a *knn*. I'll find out!"

4. The client says to you, "We have a recommendation process–we ask the genres you like and we suggest the movie; but, it's not very scientific. If I like *Action* and *Drama* what are the three best recommendations for movies?"
   "We can use a Näive Bayes for this. I'll let you know."

5. The client says to you, "Right now, the genres are kind of unrelated to one another. I wonder if you could build a tree that shows how they are related from viewers' points of view?"
   "I'll do aglommerative clustering, and we'll see if we can interpret the tree."

GC

Genre	Code
Romance	r
Science Fiction	s
Horror	h
Comedy	c
Drama	d
Action	a
Documentary	o
Classic	l

MIDG

Movie ID	Genre
1	r,s
2	o,l,a
3	c,d, h
4	s, l, o, a
5	a, d, r
6	d, h, c
7	a, d, c, o
8	h, l, r
9	s, d
10	c, r

CIDM

Customer ID	Movies
CID1	1,3,5,5,10,8
CID2	4,1,2,3
CID3	7,8,1
CID4	2
CID5	4,8,10
CID6	3,9,10,1
CID7	1,2,3
CID8	5,4,9,5
CID9	10,1,2,23
CID10	2,4,3,7,9
CID11	1,10,8
CID12	3,5,1,2,
CID13	8,1,7
CID14	5,2,8

Table 1: Data from an online movie rental company. The GC table gives the genre codes, MIDG the associated genres with the 10 movies they rent online, and CIDM the movies that a customer has rented.

2 Equivocation: k, ℓ-means Algorithm [100pts]

This problem asks you to modify the k-means algorithm to k, ℓ. The ℓ is the best number of centroids that the datum matches. So, $4, 2$-means each datum is matched to the 2 closest centroids. You can assume that we're using *average* for the best representative. Complete the algorithm below for k, ℓ-means. The final result *must* be an actual partition.

ALGORITHM `k,l-means`
INPUT (data Δ, distance $d : \Delta^2 \to \mathbb{R}_{\geq 0}$, centoid number k, Closest Matches ℓ, threshold τ)
OUTPUT (Set of centoids $c_1, c_2, \ldots, c_k$)
Assume centroid is structure $c = (v \in DOM(\Delta), B \subseteq \Delta)$
$c.v$ is the centroid value and $c.B$ is the set of nearest points.
τ is a percentage change from previous centroids
For example, $\{c_1, c_2, \ldots, c_k\}$ is previous and $\{d_1, d_2, \ldots, d_k\}$ is current
Total difference is $\Sigma_i \Sigma_j d(c_i, d_j)$
$Dom(\Delta)$ denotes domain of object.
$i = 0$ ▷ Initialize iterate where superscript is iteration
for $j = 1, k$ **do** ▷ Initialize Centroids
    $c_j^i.v \leftarrow random(Dom(\Delta))$
    $c_j^i.B \leftarrow \emptyset$
end for
$f_i = \Sigma_{j=1}^k \Sigma_{\ell=1}^k d(c_j^i.v, random(Dom(\Delta))$ ▷ Bootstrap difference between past centroids and current
repeat
    `COMPLETE CODE`
until $(|f_i - f_{i-1}| < \tau(f_{i-1}))$
return $(c_1^i, c_2^i, \ldots, c_3^i)$

Modify your k-means code to allow $\ell = 2$ and rerun your analysis on the breast cancer data[1]. Discuss your results with respect to your earlier results.

Survey

Q1: Do you own your home?	Q2: Do you own your car?
Yes	No
No	Yes
Maybe	No
⋮	⋮

3 Connections [75pts]

After examining the results of the survey, you find that there are only three kinds of responses in the same proportions: (Yes, No), (No, Yes), and (Maybe, No).

1. Are the questions Q1 and Q2 statistically dependent?
2. Are the questions Q1 and Q2 statistically correlated?

References

[1] William H. Wolberg and O.L. Mangasarian. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. 87:9193–9196, 1990.