MSiA 421: Data Mining
Professor Malthouse
Homework 1: Due January 16

Work with your team. Submit one assignment per team, and make sure that all team members have their name on the submitted assignment and give the name of your team, e.g., "Team A." The homework is due at the **beginning** of class (1pm).

1 MSO problem

The purpose of this homework is to find good segments using K-means. The file `cable2.csv` has data from a cable TV multi-system operator (MSO) that offers subscription packages of three types of products: video (e.g., cable TV), internet data and landline phone service. The purpose of the cluster analysis is to identify distinct segments of subscribers so that the MSO can send different offers to the segments and increase their value. For example, the MSO could offer a customer who has low-speed internet an offer for high-speed internet, which would increase the cash flows each month. The clusters must therefore lead to creating an offer, which will then be sent to members of the cluster.

The unit of analysis in the data set is a subscriber household. You have 10,000 households. For this assignment you will cluster on three variables, `video`, `internet` and `phone`. You can assume that those paying more for video have more channels and services. Those paying more for internet have faster speeds. These three variables give the amount paid by the customer in the most recent month for each of the three services. You also have some demographic variables from a third-party data provider: age of head of household, household income, household size, marital status, and number of children.

1. Run relevant descriptive statistics on each of the three clustering variables. Discuss aspects that will be relevant in finding the clusters. Answer: Run a summary and histograms. One important feature is that the variation in video is substantially larger than for phone.

	n	mean	sd	median	mad	min	max	range	skew	kurtosis	se
video	10000	77.81	31.14	76.17	20.20	0.39	293.72	293.33	0.41	1.58	0.31
internet	10000	36.89	18.84	40.00	11.86	0.00	100.00	100.00	-0.57	0.33	0.19
phone	10000	8.81	11.30	0.00	0.00	0.00	84.95	84.95	1.09	0.83	0.11
age	9780	52.18	18.00	52.00	20.76	18.00	99.00	81.00	0.16	-0.72	0.18
income	9780	4.63	2.53	5.00	2.97	1.00	9.00	8.00	0.10	-1.07	0.03

HHsize	9780	2.74	1.55	2.00	1.48	1.00	9.00	8.00	0.99	0.66	0.02
married	9780	0.52	0.50	1.00	0.00	0.00	1.00	1.00	-0.09	-1.99	0.01
children	9780	0.34	0.48	0.00	0.00	0.00	1.00	1.00	0.66	-1.57	0.00

2. Discuss the implications of standardizing the payment variables before the analysis. Do you suggest standardizing? If so, how? If not, why not? Answer: It is not a good idea to standardize because you will implicitly weight down video and weight up phone. Land-line phones are not a growth market, so I don't see a good reason to weight them up.

3. Find the 6–10 cluster solutions using K-means. Which do you suggest? Why? You may want to consider pseudo F statistics, etc., but the actionability should also be considered given the purpose of the analysis. Answer: Here are the F values: 5757, 5658, 5460, 5253, and 5154. They decrease monotonically, so they are not very helpful.

   - 6 clusters
     - 6.1 basic video only
     - 6.2 internet only
     - 6.3 high video, internet, some phone
     - 6.4 medium video, internet, some phone
     - 6.5 Very basic video only
     - 6.6 video + internet
   - 7 clusters
     - 7.1 high video, internet, some phone (6.3)
     - 7.2 video + internet (6.6)
     - 7.3 internet only (6.2)
     - 7.4 medium video, internet, some phone (6.4)
     - 7.5 moderate video, internet, high phone (6.6 + some 6.4)
     - 7.6 Very basic video only (6.5)
     - 7.7 basic video only (6.1)
   - 8 clusters
     - 8.1 Very basic video only (6.5, 7.6)
     - 8.2 high video, internet, some phone (6.3, 7.1)
     - 8.3 basic video only (6.1, 7.7)
     - 8.4 moderate video, internet, high phone (6.6 + some 6.4, 7.5)

8.5 medium video, internet, some phone (6.4, 7.4)

8.6 video + internet (6.6, 7.2)

8.7 moderate video, high internet, no phone (6.4+6.6, 7.2+7.4

8.8 internet only (6.2, 7.3)

- 9 clusters

9.1 Very basic video only (6.5, 7.6, 8.1)

9.2 moderate video, high internet, no phone (6.4+6.6, 7.2+7.4, 8.7)

9.3 basic video only (6.1, 7.7, 8.3)

9.4 internet only (6.2, 7.3, 8.8)

9.5 moderate video, internet, high phone (6.6 + some 6.4, 7.5, 8.4)

9.6 medium video, internet, some phone (6.4, 7.4, 8.5)

9.7 video + internet (6.6, 7.2, 8.6)

9.8 moderate video, very high internet, low phone (7.6+7.7)

9.9 high video, internet, some phone (6.3, 7.1, 8.2)

- 10 clusters

10.1 video + internet (6.6, 7.2, 8.6, 9.7)

10.2 basic video only (6.1, 7.7, 8.3, 9.3)

10.3 high video, no internet, little phone (from 9.3, 9.6, 9.9)

10.4 moderate video, very high internet, low phone (7.6+7.7, 9.8)

10.5 internet only (6.2, 7.3, 8.8, 9.4)

10.6 medium video, internet, some phone (6.4, 7.4, 8.5, 9.6)

10.7 moderate video, internet, high phone (6.6 + some 6.4, 7.5, 8.4, 9.5)

10.8 high video, internet, some phone (6.3, 7.1, 8.2, 9.9)

10.9 moderate video, high internet, no phone (6.4+6.6, 7.2+7.4, 8.7, 9.2)

10.10 Very basic video only (6.5, 7.6, 8.1, 9.1)

There is not a right answer. Give full credit if there is reasonable discussion. Given that we want to use this to increase the value, I like the 10 because there are many levels of video and internet. In this way I can isolate the low-price internet from the high, which gives me an objective (up-sell internet). Likewise there are many levels of video.

4. Profile the clusters from your best cluster solution on the demographic variables to get a better understanding of the clusters. Where are there large differences? Can you get a sense of the typical person in each cluster? Answer: It helps. You'll see that those with very high levels of video tend to be older, higher income, and more likely to be married. The very high internet group tends to be a bit younger.

5. What variables would you like to have? These could be added to the current set of variables in the cluster analysis, or clustered separately for further personalization. Think about what the MSO can observe. For example, it can know the amount of data that each household uploads and downloads each month, the number of phone calls made, the programs watched, and additional services (video on demand, pay-per-views, etc.). What exactly would you do with the Answer: If I'm going to cross-sell additional video, it would be good to know what a person is watching. If, for example, a segment watched a lot of sports, then there are premium sports and pay-per-view channels that I could offer. variables?

2 News Website Problem

(data to come) For this problem you will cluster data from a news website with subscribers. The purpose of the clusters is to develop a "lightly personalized" email newsletter to be sent each day. We know from other analyses that the more often someone reads at least some content, the less likely they are to churn. The purpose of the email is to get *relevant* stories in front of subscribers so that they will be more likely to find something that interests them. We have one year of clickstream data and about 3000 subscribers who joined, and possibly also canceled, their subscriptions during the year. For each day that the subscriber reads I will give you a count of page views by "section," e.g., home page, life and culture, crime, news, obituaries, opinion, sports.

We are looking for 5–10 clusters. Each cluster will receive its on newsletter each day featuring stories that might be of interest to the reader.

1. Discuss issues in finding the clusters. Which variables do you use? Should you combine certain variables? Do you need to "normalize" the variables in any way? Other transformations? Answer: The count variables are all right skewed and should be logged. If this is for a

newsletter then the content interests are most relevant. The location and source variables are not as relevant, nor is device.

2. Find the best cluster solution. Answer: See my markdown document

   - 5 clusters
     - 5.1 sports, news crime, life
     - 5.2 mostly news, crime, life
     - 5.3 light reader
     - 5.4 heavy reader
     - 5.5 Mostly sports
   - 6 clusters
     - 6.1 heavy reader (5.4)
     - 6.2 light reader (5.3)
     - 6.3 sports, news crime, life (5.1)
     - 6.4 Mostly sports (5.5)
     - 6.5 mostly news, crime, life (5.2)
     - 6.6 news, crime, life, sports, business(5.1+5.2+5.4)
   - 7 clusters
     - 7.1 news, sports, crime, obits, life (6.1+6.6)
     - 7.2 Mostly sports (5.5, 6.4)
     - 7.3 mostly news, crime, life (5.2, 6.5)
     - 7.4 light reader (5.3, 6.2)
     - 7.5 news, crime, life, sports, business(5.1+5.2+5.4, 6.6)
     - 7.6 sports, news crime, life (5.1, 6.3)
     - 7.7 heavy reader (5.4, 6.1)
   - 8 clusters
     - 8.1 Mostly sports (5.5, 6.4, 7.2)
     - 8.2 sports, some news
     - 8.3 mostly news, crime, life (5.2, 6.5, 7.3)
     - 8.4 news, sports, crime, obits, life (6.1+6.6, 7.1)
     - 8.5 news, crime, life, sports, business(5.1+5.2+5.4, 6.6, 7.5)
     - 8.6 light reader (5.3, 6.2, 7.4)
     - 8.7 sports, news crime, life (5.1, 6.3, 7.6)
     - 8.8 heavy reader (5.4, 6.1, 7.7)

Again, there's no right answer. I like how business starts to enter into the solutions around 6 clusters. I'm not sure we get much more after that that's useful