Mini Project #2

Practice the three basic tasks of visual data analytics

- use data from mini project #1 (or other), begin with $|N| \geq 500$, $|D| \geq 10$)
- client-server system: python for processing (server), D3 for VIS (client)

Task 1: data clustering and decimation (30 points)

- implement random sampling and stratified sampling (remove 75% of data)
- the latter includes the need for k-means clustering (optimize k using elbow)

Task 2: dimension reduction on both org and 2 types of reduced data (30)

- find the intrinsic dimensionality of the data using PCA
- produce scree plot visualization and mark the intrinsic dimensionality
- show the scree plots before/after sampling to assess the bias introduced
- obtain the three attributes with highest PCA loadings

Task 3: visualization of both original and 2 types of reduced data (40 points)

- visualize the data projected into the top two PCA vectors via 2D scatterplot
- visualize the data via MDS (Euclidian & correlation distance) in 2D scatterplots
- visualize the scatterplot matrix of the three highest PCA loaded attributes

Due date: Tuesday, 3/10

Note: this is a time-intensive project – start early!!

Deliverables

You need to upload to Blackboard the following by the due date:

- 2-3 page report with illustrated description of your program's capabilities and implementation detail
  - add code snippets to show how you did things
  - discuss interesting observations you made in the data
  - constructively compare the various visualization alternatives
  - compare the effects of down sampling (2 methods) with the original data
  - make good use of visualizations
  - video file that shows all features of your software in action
  - archive file with source code

Grading

- TA will pick students at random for thorough code review sessions
- you better know your code !!!
- so, please do not just copy code beyond the D3 templates
- or even worse, videotape someone else's program

Dealing With Categorical Variables

k-means clustering is mainly defined for numerical variables

- what to do for categorical variables?

Categorical data that are ordinal can be used for clustering

- simply assign values 0,1,2,... to the ordinal levels

For nominal categorical data you can use one-hot encoding

- say you have a variable Color with three levels, green, blue, re
- make three variables Red, Green Blue and set either to 1 or 0

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1