CS6301: R For Data Scientists

LECTURE 10: CLUSTERING NON-NUMERICAL DATA

Clustering non-Numeric Data

We saw how to do text, but what if we have categorical data?

Or a combination of numerical and categorical data?

We need a way to define distance …

- Kmeans assumes Euclidian^2, so data must be numeric
- Hclust will work with any distance matrix
- dist() defines some "metrics" – ways of defining distance
- … but they apply to numeric data

Let's explore distance metrics in more detail first

dist() metrics

Suppose $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$

euclidean:

Usual distance between the two vectors (L_2 norm).

$$\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

maximum:

Maximum distance between two components of x and y (L_∞ norm).

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \max(|x_1 - y_1|, |x_2 - y_2|)$$

manhattan:

Absolute distance between the two vectors (L_1 norm).

$$\|\boldsymbol{x} - \boldsymbol{y}\|_\mathbf{1} = |x_1 - y_1| + |x_2 - y_2|$$

dist() metrics

Suppose $\boldsymbol{x} = (x_1, x_2)$ and $\boldsymbol{y} = (y_1, y_2)$

canberra:

A weighted version of the L_1 metric.

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{2} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

binary:

Intended for use with binary data. Compare the components of two vectors, compute the number of components with only ‘1’ versus the number of components with at least one ‘1’.

minkowski:

Like Euclidian, replace ‘2’ with ‘p’.

Categorical Data

If all of our data is categorical, can we replace the categories with dummy variables and then use the above metrics?

- model.matrix() will do this, but remember dummy variables have a base case – we do not get a variable for every category, and we may need this

Use acm.disjonctif() in package **ade4**

- Gives a binary variable for all factors
- But need all data to be categorical

K Modes

There is an algorithm called “Kmodes” which works with strictly categorical data …

Invented by Huang in 1997

Implemented in the **klaR** package in R

Very similar to K-Means; set number of centroids, computes distance differently

Mixed Data – Gower Metric

What if our data has a mix of numeric and categorical variables?

There is a distance metric called “Gower” that is designed to handle this

It is implemented in a couple of packages:

- vegdist() in the **vegan** package – also does other metrics
- daisy() in the **cluster** package

For each variable type, a particular distance metric that works well for that type is used and scaled to fall between 0 and 1.

Then, a linear combination using user-specified weights (most simply an average) is calculated to create the final distance matrix.

More Metrics

Jaccard is similar to binary – can be done using the vegdist() function in **vegan** package

For string data, can use an "edit distance" function

- The function in R is called adist(x,y) – give it two strings
- Can be useful for creating dissimilarity matrix for data with char values
  - But all data would need to be chars

Also possible to do a correlation dissimilarity

- The function is corDist(), and is in the **MKmisc** package
- Note correlation can be negative, so need to be careful with linkage (Ward)

Clustering: Summary

Know your data: numeric, categorical, mix?

If all numeric, can do K-means of Hclust

- Can use dist() to create distance matrix with different metrics

All categorical?

- Use ade4 to convert to binary
- Use dist(), vegdist(), daisy() to create dissimilarity matrix
- Kmeans or Hclust will work
- Can also do K-modes

Mixture?

- Use Gower metric

Links

https://www.r-bloggers.com/clustering-mixed-data-types-in-r/

https://shapeofdata.wordpress.com/2014/03/04/k-modes/

https://www.rdocumentation.org/packages/klaR/versions/0.6-12/topics/kmodes

http://dpmartin42.github.io/blogposts/r/cluster-mixed-types