**Gini Curve project**

# Significance

The main idea behind the project is from the problem of characterizing the data cluster. In many data analytic problems, the data do not have ‘label’ (analyzing this data is also called ‘unsupervised learning’). In this scenario, people often try to find unknown data group, which is called ‘clustering’, which hopefully can gain some structural understanding of the data. After clustering, people would as how to interpret the results, or characterizing the data. This project is for the characterizing purpose.

For example, in figure 1, suppose that you cluster the car data and find 4 groups as in figure 1, then you can characterize the clusters such as

- Group 4: cars having almost **equal** distribution of gas consumed per mile and top acceleration

- Group 2: cars having almost the **same** acceleration, but quite **different** gas consumed per mile.

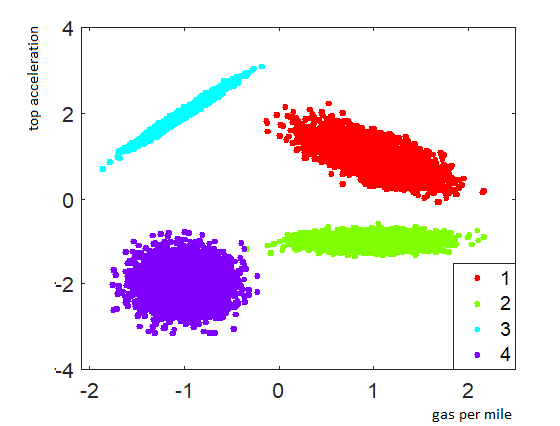


Figure 1. Toy example on car clustering based on gas per mile and top acceleration

When we say ‘equal’, ‘same’ or ‘diverged’, we want to talk about the extent of inequality. The Gini index (research it yourself) is a popular metric to measure inequality. In many scenarios, the best way to compute the Gini index is by combination of two dimensions in the data. For example, in cluster 3 (figure 1), neither using ‘gas per mile’ nor ‘top acceleration’ alone would give the highest Gini index; in this case, you would need to combine these two dimensions in some way to get the highest Gini index. Which way? Then you need to figure out the right data projection (which is called ‘rotation’ in geometry).

# Input and output

The input file contains three columns. The first two columns are the x and the y coordinate of the data points when we plot them in the 2D coordinate map. The third column is the clustering membership of each datapoint. In this problem, you don’t need to do clustering.

The output is the ‘gini index plot’, which is done as follow:

- Choose an axis z such that the z-axis and the x-axis have the angle θ.

- Project the data points onto the z-axis. This would give you a **score** to compute Gini index.

- Compute the Gini index when projecting all cluster data points onto the z-axis. Plot the Gini on the z-axis.

- If you make θ slowly increasing from 0 to 2π, then you will have a Gini index plot.

There would be other mathematical questions, especially about data transformation, that you would have to solve for on the way. But I am hiding these questions for you to figure out.

I provide two examples of input and output for you to figure out the entire plotting engine.

+ Input file: FourDisk.txt. Clustering visualization: FourDiskClusterPlot.png. Output: FourDiskGiniPlot.png

+ Input file: FourEclipse.txt. Clustering visualization: FourEclipseClusterPlot.png. Output: FourEclipseGiniPlot.png

In addition to the plotting and graphic, you are expected to build a simple web page of three components: a select file button to submit the input file (same format to the input files above), a ‘draw’ button, and the visualization of the output on the web page. Making everything run in your local web server is the basic completement.