CS412 – Mini MP 1 – Knowing your data

The MP borrows quite considerable amount of material from a certain source. We will publish the source after the submission's due date because it contains answers for a few questions. Don't try to find the existing answers because the MP is not hard, and it is really fun and useful. If you have any questions regarding the MP, please contact Long Pham, ltpham3 at Illinois.edu.

To finish the MP, please read this document from the beginning to the end. You will find a few red sentences. They specify the tasks you need to do. If the task requires you to insert a diagram, simply replace the red text by the diagram. If it requires you to write something, please write the answers in red.

The MP requires Matlab. The software is free for grad students in UIUC Webstore. And it is also available in EWS machines on campus. If you are not able to access both sources, please let us know ASAP. Please also note that it may take you only 1-2 hours to finish the MP, so if you don't often use the heavy Matlab software, you may want to use one of the EWS machines on campus.

For students who don't want to use Matlab: the purpose of the MP is to play with software and data, not programming. Thus, you can use any software other than Matlab but you will not have some existing source code and supports from TA.

In this MP, we'll use Matlab's built-in `carbig` dataset, a dataset that contains various measured variables for about 400 automobiles from the 1970's and 1980's. We want to illustrate multivariate visualization using the values for fuel efficiency (in miles per gallon, MPG), acceleration (time from 0-60MPH in sec), engine displacement (in cubic inches), weight, and horsepower. We'll use the number of cylinders to group observations.

```
load carbig
X = [MPG,Acceleration,Displacement,Weight,Horsepower];
varNames = {'MPG'; 'Acceleration'; 'Displacement';
'Weight'; 'Horsepower'};
```

Scatterplot Matrices

Viewing slices through lower dimensional subspaces is one way to partially work around the limitation of two or three dimensions. For example, we can use the `gplotmatrix` function to display an array of all the bivariate scatterplots

between our five variables, along with a univariate histogram for each variable.

```
figure
gplotmatrix(X,[],Cylinders,['c' 'b' 'm' 'g'
'r'],[],[],false);
text([.08 .24 .43 .66 .83], repmat(-.1,1,5), varNames,
'FontSize',8);
text(repmat(-.12,1,5), [.86 .62 .41 .25 .02], varNames,
'FontSize',8, 'Rotation',90);
```

Please show us the scatter matrix by running the above code on Matlab

The points in each scatterplot are color-coded by the number of cylinders: blue for 4 cylinders, green for 6, and red for 8. There is also a handful of 5 cylinder cars, and rotary-engined cars are listed as having 3 cylinders. This array of plots makes it easy to pick out patterns in the relationships between pairs of variables. However, there may be important patterns in higher dimensions, and those are not easy to recognize in this plot.

Parallel Coordinates Plots

The scatterplot matrix only displays bivariate relationships. However, there are other alternatives that display all the variables together, allowing you to investigate higher-dimensional relationships among variables. The most straight-forward multivariate plot is the parallel coordinates plot. In this plot, the coordinate axes are all laid out horizontally, instead of using orthogonal axes as in the usual Cartesian graph. Each observation is represented in the plot as a series of connected line segments. For example, we can make a plot of all the cars with 4, 6, or 8 cylinders, and color observations by group.

```
Cyl468 = ismember(Cylinders,[4 6 8]);
parallelcoords(X(Cyl468,:), 'group',Cylinders(Cyl468), ...
               'standardize','on', 'labels',varNames)
```

Please show us the scatter matrix by running the above code on Matlab

The horizontal direction in this plot represents the coordinate axes, and the vertical direction represents the data. Each observation consists of measurements on five variables, and each measurement is represented as the height at which the corresponding line crosses each coordinate axis. Because the five variables have widely different ranges, this plot was made with standardized values, where each variable has been standardized to have zero mean and unit variance. With the color coding, the graph shows, for example, that 8 cylinder cars typically have low values for MPG and acceleration, and high values for displacement, weight, and horsepower.

Even with color coding by group, a parallel coordinates plot with a large number of observations can be difficult to read. We can also make a parallel coordinates plot where only the median and quartiles (25% and 75% points) for each group

are shown. This makes the typical differences and similarities among groups easier to distinguish. There is a parameter of function parallelcoords that allows you to create what has just been described. Please modify the above code, write it here, and show us the new parallel coordinate plot. You may need to read the documentation of function parallelcoords.

Glyph Plots

Another way to visualize multivariate data is to use "glyphs" to represent the dimensions. The function glyphplot supports two types of glyphs: stars, and Chernoff faces. Here, we consider only the Chernoff face. This glyph encodes the data values for each observation into facial features, such as the size of the face, the shape of the face, position of the eyes, etc.

```
models77 = find((Model_Year==77));

dissimilarity = pdist(zscore(X(models77,:)));

Y = mdscale(dissimilarity,2);

glyphplot(X(models77,:), 'glyph','face', 'centers',Y, ...

        'varLabels',varNames, 'obslabels',Model(models77,:));

title('1977 Model Year');
```

Please show us the Chernoff face plot by running the above code on Matlab

Here, the two most apparent features, face size and relative forehead/jaw size, encode MPG and acceleration, while the forehead and jaw shape encode displacement and weight. Width between eyes encodes horsepower. It's notable that there are few faces with wide foreheads and narrow jaws, or vice-versa, indicating positive linear correlation between the variables displacement and weight. That's also what we saw in the scatterplot matrix.

Now by looking at all the plots so far, could you identify a pair of correlated features (which has not been mentioned so far)? By your intuition, could you explain why the positive or negative correlation makes sense?

Final note: If you have any trouble with using a particular Matlab function, you may type “Help [the-function]” in Matlab to read its documentation.

Have fun!