Visual Analytics: Homework 4

Nishit Umesh Parekh

October 6, 2017

1 How to run

Use the command `bokeh serve hw4.py`

2 Folder Structure

There are two files for making the plots.

1. `hw4.py` contains the main code, used for making the charts using Bokeh.
2. `helper_functions.py` contains function for making and editing the data.

3 Explanation

There are a total of 4 charts in the document, divided into 2 groups.

1. The **Category-Category** plot.

   This plot displays two categorical variables in the X and Y Axis, and uses a 3rd categorical variable as the Target Class. Since many data points will lie at the same point in this chart (due to constraints of categorical data), I have shown the **count** of items, indicated by the radius of the circle at the location in the plot. Thus, a bigger circle means there are more data points with those values of the categorical variables. The radius value is scaled, with 0 representing no data points and 1 representing all data points.

2. The **Precision-Recall for Categorical Data** plot.

   This plot shows the precision-recall chart, based on the decision tree trained on the data points represented in the plot above. A green square indicates positive connotation (for the TT or FF case), and the red squares represent the FT and TF cases. Just like the above plot, size of the square indicates the proportion of items in each section.

3. The **Produce-Meat** plot.

   This plot displays two real-valued variables, indicating the total amount of consumption (Frequency x Quantity) of various types of food consumed per week. The target variable is a categorical variable, like the case above. Since the variables are real valued, the data points are represented in their raw form, with individual points.

4. The **Precision-Recall for Real-Valued Data** plot.

   This plot shows the precision and recall values for the above plot, much like the other similar precision-recall plot explained above.

4 Inference

4.1 What to look for

To get an easy-to-understand statement like the one described in the homework prompt, there are two necessary conditions:

1. The precision-recall chart should have maximum size for the green squares, and minimum size for the red squares. This ensures that the decision tree is apt in dividing the classes into homogenous parts.

2. There should be few partitions of the data in the scatter plot. This ensures that the resulting inference is not too complicated.

4.2 What I found

Unfortunately, I could not find noteworthy exampled of such correlations as explained above. This may be due to the choice of variables I have taken (which are not generally correlated, if seen from a biological point of view), and also because there are very few data points to get a real inference.

Nevertheless, given below are some setting for which I could make a few weak inferences:

1. (Plot 1: Axes: (Type) Handedness, (Boolean) Diabetes; Target: (Boolean) Rash)— This one shows some correlation, implying that: **You are less likely to have a Rash, if you're right handed or have no diabetes**

2. (Plot 2: Axes BROCCOLI , HAMBURGER; Target: (Boolean) Cancer)— This one shows a correlation, explained as: **Eating a minimum of 3 hamburgers, and within 2-12 broccoli in a week, can lead to a lower probability of Cancer**.