AYASDI

4400 Bohannon Drive, Suite 200
Menlo Park, CA 94025
650.704.3395 t
ayasdi.com

Ayasdi Data Scientist Challenge One

The following questions refer to the data set defined in the tab-delimited file **dataset_challenge_one.tsv**. The first line in the file contains column headers. Each line after the header line represents a sample from an experiment. There are 1553 columns in the file that correspond with variables measured for each sample, starting with column Variable_1 and ending with column Variable_1553. There is also a column called **class** which has values of 0 and 1 for each sample row.

For this challenge, we're not looking for perfect answers - instead we want to see how you approach a new data problem and think about appropriate statistics and visualizations. Also, if you feel like you can't answer a question, or if you feel like the question is unclear, please explain why.

Coding must be done in python. Include all code you write to answer the questions. It would be helpful if we can run your code and reproduce your results. Use of libraries for standard methods (e.g. PCA, classifiers) is strongly encouraged.

Note: The use of ipython notebooks to submit your results is discouraged. Your code should be modular (preferably object-oriented), maintainable and follow PEP-8 guidelines. Additionally, any solution that does not depend on the **pandas** python package will be strongly preferred.

1) Explore the data. Provide summary statistics and at least three visualizations for the variable columns (one at a time, or in combination). In a brief paragraph, summarize the distributions for variable values and explain your choices for visualization. Are there any anomalous distributions for variables? How did you determine this?

2) Present a Principal Components Analysis (PCA) plot for the samples. It should contain a scatterplot of the sample points with the axes PC1 vs. PC2. Indicate on the plot which samples have class = 1 and which have class = 0.

3a) Calculate a statistic for every variable that describes its relationship with the class column. Don't list them all, but for the variable column with the most significant statistic, provide a visualization that shows its relationship with class. Include a brief paragraph describing your choices of statistic and visualization.

3b) Calculate a statistic for every variable that describes its relationship with PC1 (i.e. the first principal component). For the variable with the most significant statistic, provide a visualization that shows its relationship with PC1. Include a brief paragraph describing your choices of statistic and visualization.

4) Create a classifier model predicting class of each sample using some or all of the variables in the dataset. Use cross-validation to calculate the effectiveness of your classifier. Provide a short paragraph detailing your rationale for picking a classifier method, selecting a subset of variables for the model (if you did this), followed by a summary of your classifier's performance.