**Experimental Setup for LDP-NB**

This document contains the explanation of the experiment codes and the instruction to repeat the experiments for LDP-NB and LDP-GNB.

**Experiment Environment:**

* Python 2.7
* **Package prerequisites:**
  + numpy
  + scipy
  + scikit-learn
  + pandas
  + matplotlib

For all experiments, the dataset is located in “**./datasets/” or “./cont-datasets”,** all datasets follow the same format. Each row corresponds to a sample, and the first column indicates the label, and the rest of columns are features.

Three files are specifically important:

* ***gldpnb.py***: This file contains the main class for the categorical LDP Naïve Bayes. It includes code that performs LDP-NB using five different encoding/perturbation mechanisms based on randomized response.
* ***Ldp-gaussian\_naive\_bayes.py***: This file contains the main class for numerical LDP Gaussian Naïve Bayes. It includes the code that performs LDP-GNB using 3 different perturbation mechanisms, two of which are based on adding Laplace noise.
* ***Helper.py***: This file contains the code for some helper functions, especially for dimensionality reduction.

To run the experiment on numerical datasets while trying different DR techniques, one should run the python file “***exp-dr-ldp-cont-gnb.py***”. One would just need to change the index value of the variable “dtID” with values from 0 to 3 that correspond to these four different datasets: ["Australian", "breast-cancer", "diabetes", “German"]. Furthermore, one can choose any of the perturbation methods by setting the variable “ldpID” to 0, 1 or 2 corresponding to “Basic Laplace – One feature submission”, “Basic Laplace – All features submission” or “Algorithm 2 – One feature submission”. Finally, one can use the variable “drMethods” to either choose no dimensionality reduction “0 - None”, or choose one of the two DR techniques: “1 – PCA” or “2 – DCA”.

For categorical datasets, one would need to use the python file “***expLNB.py***”. Two values can be changed. The dataset ID “dtID” which can take five values from 0 to 4 to denote one of the five datasets: ['sample', 'car', 'connect', 'mushroom', 'chess']. Furthermore, the encoding ID can be changed to denote one of the five encoding methods (0 to 4): ["DE", "SUE", "OUE", "SHE", "THE"]. These are Directing Encoding (DE), Symmetric Unary Encoding (SUE), Optimized Unary Encoding (OUE), Summation with Histogram Encoding (SHE), Threshold Histogram Encoding (THE).