CSCI316: Description of Projects (Spring 2020)

Project 1
Covertype Dataset
https://archive.ics.uci.edu/ml/datasets/covertype
The Covertype dataset records the types of forest-covering parcels of land in Colorado, USA. Each example contains several features describing each parcel of land—like its elevation, slope, distance to water, shade, and soil type—along with the known forest type covering the land. The forest cover type is to be predicted from the rest of the categorical and numerical features, of which there are 54 in total. There are 581,012 recordings in this dataset. This is a classification task.

Project 2
Census Income Dataset
https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29
This data set contains weighted census data extracted from the 1994 and 1995 Current Population Surveys conducted by the U.S. Census Bureau. The data contains 41 demographic and employment related variables. There are 199,523 instances in the data file and 99,762 in the test file. Note that Incomes have been binned at the $50K level to present a binary classification problem.

Project 3
YearPredictionMSD Data Set
https://archive.ics.uci.edu/ml/datasets/YearPredictionMSD
This dataset is a subset of the Million Song Dataset: (http://labrosa.ee.columbia.edu/millionsong/). The dataset has been pre-processed. In particular, the numerical audio features are extracted (by using. the Echo Nest API). The first attribute is the release year. Other attributes include two group: the timbre average (12 columns) and the timbre covariance (78 columns). The task is to predict the release year of a song from audio features. Use the first 463,715 examples as the training dataset and the last 51,630 examples as the test dataset. This is a regression problem.

Project 4
Record Linkage Dataset
https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns
The dataset contains element-wise comparison of records with personal data from a record linkage setting. The task is to decide from a comparison pattern whether the underlying records belong to one person. The records represent individual data including first and family name, sex, date of birth and postal code, which were collected through iterative insertions in the course of several years. The comparison patterns in this data set are based on a sample of 100,000 records. Data pairs were classified as “match” or “non-match”. Thus “is_match” is the outcome variable. Note that “id_1” and “id_2” should not be used for prediction but could be used to construct connected components from the found matches.

Project 5
UNSW Network Intrusion Dataset
https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/
Several datasets are available for model development and model testing for IDS. This project will utilize the UNSW-NB15 dataset. The UNSW-NB15 dataset is published by Cyber Range Lab of the Australian Centre for Cyber Security. The data was collected over 15 hours by an IXIA traffic generator in 2014, then pre-processed and labelled as “normal” and various types of “attack”. Download the *training dataset* and the *test dataset* from the above link. The task is to predict whether a record represents

"normal" or "attack" (a binary classification problem). Note that the last two columns represent the targe variables, which should not be used as training features.

Project 6
Physical Unclonable Functions Data Set
https://archive.ics.uci.edu/ml/datasets/Physical+Unclonable+Functions
The dataset is generated from Physical Unclonable Functions (PUFs) for authentication purposes. Two datasets are generated from simulation: 1) 5-XOR_128bit dataset: It is generated using 5-XOR arbiters of 128bit stages PUF. It consists of 6 million rows and 129 attributes where the last attribute is the class label (1 or -1). It is divided into two sets: training set (5 million) and testing set (1 million); 2) 6-XOR_64bit dataset: It is generated using 6-XOR arbiters of 64bit stages PUF. It consists of 2.4 million rows and 65 attributes where the last attribute is the class label (1 or -1). It is divided into two sets: training set (2 million) and testing set (400K). You can choose one of the two datasets to complete the project. This is a classification task.

Project 7
Supersymmetric Particle Data Set
https://archive.ics.uci.edu/ml/datasets/SUSY
This is a classification problem to distinguish between a signal process which produces supersymmetric particles and a background process which does not. The data is produced by simulations. The first column is the class label (1 for signal, 0 for background), followed by the 18 features. The first 8 features are kinematic properties measured by the particle detectors in the accelerator. The last ten features are functions of the first 8 features; these are high-level features derived by physicists to help discriminate between the two classes. The last 500,000 examples are used as a test set.

Project 8
Character Font Images Data Set
https://archive.ics.uci.edu/ml/datasets/Character+Font+Images
This data set consists of character images from scanned and computer-generated fonts. The total number of character fonts is 153, which can be regarded as 153 classes. The data file contains .csv, comma delimited files, one for each font. Each .csv file has a header row with the data set attribute names. The first column contains the information on the font family, which is class label. The 3rd to 10th columns describe the basic information of each image, and the remaining 400 columns contain the pixel values of each image. This is a classification task.

--- END ---