Assignment 1 : Half Space Classifier

The halfspace hypothesis space is the set of hypotheses that consist of a hyperplane in a d-dimensional coordinate space that classifies a feature vector $\varphi(x) \in R^{d+1}$ as either -1 or 1 based on which side of the hyperplane it lies. Here d represents the number of features of item x. A hypothesis in this space is often called a linear classifier or a perceptron.

Part 1 : Half Space classifier using Linear Programming

In this I have used the data set which is provided in file.csv which is of **Room Occupancy** can be found at link:https://www.kaggle.com/sachinsharma1123/room-occupancy.
To run the Program, first upload the data set provided with the .ipynb file or download from the above mentioned site.
Description of Dataset:

1. Experimental data used for **binary classification** (room occupancy) from Temperature,Humidity,Light and CO2.
2. This dataset contains 5 features and a target variable: Temperature, Humidity, Light, Carbon dioxide(CO2)
3. Target Variable: Occupancy
   - 1-if there is a chance of room occupancy.
   - 0-No chances of room occupancy (initially 0, set to -1 during processing)
4. This Dataset contains 2666 rows and 6 columns with 5 features and 1 target.
5. No missing values.

I have used the linprog LP Solver provided by scipy.optimize. All the required steps are already written in the .ipynb file, Different Splits are also Provided; Maximum accuracy is found as 37%.

Part 2 : Half Space classifier using Perceptron

In this I have used the data set which is provided in file.csv which is of **Room Occupancy** can be found at link:https://www.kaggle.com/sachinsharma1123/room-occupancy.
To run the Program, first upload the data set provided with the .ipynb file or download from the above mentioned site.
Description of Dataset:

1. Experimental data used for **binary classification** (room occupancy) from Temperature,Humidity,Light and CO2.
2. This dataset contains 5 features and a target variable: Temperature, Humidity, Light, Carbon dioxide(CO2)
3. Target Variable: Occupancy
   - 1-if there is a chance of room occupancy.
   - 0-No chances of room occupancy (initially 0, set to -1 during processing)
4. This Dataset contains 2666 rows and 6 columns with 5 features and 1 target.
5. No missing values.

I have coded the perceptron as per the perceptron rule. All the required steps are already written in the .ipynb file, Different Splits are also Provided; Maximum accuracy is found as 37%.

Part 3 : Half Space classifier using Logistic classification

In this I have used the data set which is provided in diabetes.csv which is of **Pima Indians Diabetes Database**.

And it can be found at link: https://www.kaggle.com/uciml/pima-indians-diabetes-database.

To run the Program, first upload the data set provided with the .ipynb file or download from the above mentioned site.

Description of Dataset:

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

1. Experimental data used for **binary classification** (diabetes) from Pregnancies, Glucose, Skin thickness, BMI, Blood Pressure, Insulin, Diabetes Pedigree Function, Age.
2. This dataset contains 8 features and a target variable.
3. Target Variable: Outcome
    - 1-if there is a chance of Diabetes.
    - 0-No chances of Diabetes.
4. This Dataset contains 768 rows and 9 columns with 8 features and 1 target.
5. No missing values.

I have coded the logistic regression using a class and also used the original to check the difference between the coded one and the original. All the required steps are already written in the .ipynb file, Different Splits are also Provided; Maximum Accuracy is Found as 61%.

Note:

1. For Nonlinear Separable Dataset I have tried the centroid clustering but it was not working so I did not add that code.
2. For Lp Solver I tried different libraries like PuLP but it does not give a feasible Solution.