IIT KHARAGPUR AI4ICPS I HUB FOUNDATION

Hands-on **A**pproach to **AI, Cohort-2, July – October 2024**

Programming Assignment 2

Due date: Friday, August 9, 2024, EOD–IST.

Important Instructions about Programming Assignments

1. Programming assignments will be evaluated automatically. **Do not** change the skeleton code provided to you.
2. Write your code **only in the designated places** in the skeleton code, and process the input data provided to you in the designated variables. **Do not alter** the input output structure in the skeleton code.
3. **Do not import** any additional libraries. **Do not use any additional files** for the processing other than those mentioned in the skeleton code from **a.(i)** to **a.(iv)**.
4. Failure to comply with these instructions may lead to you getting **zero marks** for the assignment, even if the solution is largely correct.

Question:

Objective: Pulsars are a rare type of neutron star that produces radio emissions detectable here on Earth. They are of considerable scientific interest as probes of space-time, the interstellar medium, and states of matter. Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation. Thus, a potential signal detection known as a 'candidate', is averaged over many rotations of the pulsar, as determined by the length of an observation. In the absence of additional information, each candidate could potentially describe a real pulsar. However, in practice, almost all detection is caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find. Machine learning tools are now being used to automatically label Pulsar candidates to facilitate rapid analysis. Classification systems in particular are being widely adopted, which treat the candidate data sets as binary classification problems. Here, the legitimate pulsar examples are a minority positive class, and spurious examples are the majority negative class.

i. Randomly pick 80% of the data as a training set and the rest as a test set.

ii. Normalize each feature of the dataset to have a zero mean and unit variance. Note that while normalizing the features, their mean and variance should be computed over the train split only. Once the mean and variance are computed using only the train split, you normalize the test split using the mean and variance computed over the train split. Once the mean and variance are computed using only the train split, you normalize the test split using the mean and variance computed over the train split.

iii. Note that training requires solving the dual optimization problem. To solve the dual optimization problem, you must use the python package: `cvxopt.solvers`

Write a SVM function that takes a new datapoint as input and predicts the class. In SVM, the hyperparameter `C` regulates the regularization strength, affecting the balance between a smooth decision boundary and the accurate classification of training points. Now, for a given set of hyperparameter values `C = [0.1, 1, 10, 100, 1000]`, what will be their corresponding accuracies, provided we are using the linear kernel?

Instructions:

1. **Do not import any more libraries or modify any functions given in the skeleton code.**
2. **Input for evaluating the test cases; do not change the hyperparameter `C` value.**
3. **The output will be in decimal points.**
4. **You must use `random_state=42` during the train test split.**

Dataset: The dataset contains samples of pulsar candidates collected during the High Time Resolution Universe Survey (South). It has around 17898 instances with 8 continuous attributes.

The target attribute is "Class" which can be legitimate (1) or spurious (0). Please note that the dataset may contain missing values. To handle these missing values, you should use appropriate techniques.

Data Filename: pulsar_star_dataset.csv

Dataset description: The first four attributes are simple statistics obtained from the integrated pulse profile (folded profile). This is an array of continuous variables that describe a longitude-resolved version of the signal that has been averaged in both time and frequency. The remaining four variables are similarly obtained from the DM-SNR curve. These are summarized below:

1. Mean of the integrated profile.

2. Standard deviation of the integrated profile.

3. Excess kurtosis of the integrated profile.

4. Skewness of the integrated profile.

5. Mean of the DM-SNR curve.

6. Standard deviation of the DM-SNR curve.

7. Excess kurtosis of the DM-SNR curve.

8. Skewness of the DM-SNR curve.

9. Class.

Here, DM-SNR stands for two things: Dispersion Measure (DM) and Signal-to-Noise Ratio (SNR). DM, as the name suggests, measures the dispersion or spread of pulsar's signals during their journey from pulsar to earth. SNR, on the other hand, measures the strength of a pulsar's signal relative to background noise. DM is calculated from the time delay of each signal when it arrives on earth, while SNR is calculated at the peak intensity of each signal.

Sample Test Cases

```
"input": "0.9\n",
"output": "0.97\n"

"input": "9\n",
"output": "0.975\n"


"input": "90\n",
"output": "0.975\n"


"input": "900\n",
"output": "0.98\n"

"input": "9000\n",
"output": "0.98\n"
```