

SMAI (CSE 471)  
Spring-2019  
Assignment-2 (100 points)  
Posted on: 22/1/2019  
Due on: 2/2/2019, 7:00 PM

---

- Questions can involve a mix of writing code/scripts and answering questions or analyzing results.
- Code: Your scripts should be of the form `q-x-y.py` where  $x$  is the main question,  $y$  is the sub-question. For e.g., `q-1-2.py` is Python script for sub-question 2 within question 1. If you are submitting Jupyter notebook file (.ipynb), make sure that it is properly formatted and documented with question part numbers (Part-1, Part-2 etc.).
- In case you are submitting Jupyter notebook, you MUST submit .py file as well.
- Your code should accept test file name as command line argument.
- Ensure that submitted assignment is your original work. Please do not copy any part from any source including your friends, seniors and/or the internet. If any such attempt is caught then serious action will be taken.
- Use suitable train-validation split for your training and validation (20% of data).
- Numpy, pandas/csvReader(for data processing) are allowed. Inbuilt library function are not allowed.
- Evaluation will be done based on your understanding, report and accuracy on purely unseen test data (provided at the time of assignment evaluation).
- Report your precision, recall, F1 score and accuracy on validation data in your report.
- Report should contain details of algorithm implementation, results and observations.

## 1 Question

1. (20 points) We will be working on two datasets as part of this question.
  1. **Robot1 & Robot2** The robot problem is from artificial robot domain in which robots are described by 6 different attributes. The learning task is a binary classification task. Perform modelling on both the datasets. Data is available at [http://preon.iiit.ac.in/~sanjoy\\_chowdhury/RobotDataset.zip](http://preon.iiit.ac.in/~sanjoy_chowdhury/RobotDataset.zip)

2. **Iris.csv** The data set consists of samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Data and it's description is available at [http://preon.iiit.ac.in/~sanjoy\\_chowdhury/Iris.zip](http://preon.iiit.ac.in/~sanjoy_chowdhury/Iris.zip)

**Use the above datasets to answer following questions**

1. **Implement a KNN classifier for each of the datasets.** Report precision, recall, f1 score and accuracy. Compare your result with in-built(scikit-learn) KNN function to check correctness of your algorithm.**(10 points)**
2. Use different distance measures as applicable. Plot graph to report accuracy with change in value of K. Also suggest possible reason for better performance. **(10 points)**
2. (20 points) A bank is implementing a system to identify **potential customers who have higher probability of availing loans to increase its profit.** Implement Naive Bayes classifier on this dataset to help bank achieve its goal. Report your observations and accuracy of the model. Data is available at [http://preon.iiit.ac.in/~sanjoy\\_chowdhury/LoanDataset.zip](http://preon.iiit.ac.in/~sanjoy_chowdhury/LoanDataset.zip)
3. (20 points) We are given a dataset containing various criteria important to get admissions into Master's program and probability of getting an admit. Dataset is available at [http://preon.iiit.ac.in/~sanjoy\\_chowdhury/AdmissionDataset.zip](http://preon.iiit.ac.in/~sanjoy_chowdhury/AdmissionDataset.zip)
  1. Implement a model using **linear regression** to predict the probability of getting the admit.**(10 points)**
  2. Compare the performance of Mean square error loss function vs Mean Absolute error function vs Mean absolute percentage error function and explain the reasons for the observed behaviour.**(5 points)**
  3. Analyse and report the behaviour of the coefficients(for example: sign of coefficients, value of coefficients etc.) and support it with appropriate plots as necessary**(5 points)**
4. (20 points) Use the Admission dataset as in the third question.
  1. Implement logistic regression model to predict if the student will get admit.**(10 points)**

2. Compare the performances of logistic regression model with KNN model on the Admission dataset. **(5 points)**
3. Plot a graph explaining the co-relation between threshold value vs precision and recall. Report the criteria one should use while deciding the threshold value. Explain the reason behind your choice of threshold in your model. **(5 points)**
5. (20 points) Implement logistic regression using One vs All and One vs One approaches. Use the following dataset [http://preon.iiit.ac.in/~sanjoy\\_chowdhury/wine-quality.zip](http://preon.iiit.ac.in/~sanjoy_chowdhury/wine-quality.zip) for completing the task. Report your observations and accuracy of the model.