SMAI (CSE 471)
Spring-2019
Assignment-7 (40 points)
Posted on: 03/3/2019
Due on: 10/3/2019, 11:55PM

- 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 arguement.
- 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.
- 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 should contain details of algorithm implementation, results and observations.

1. (40 points) In this assignment you will be working with an already implemented question from Assignment 2 and try to see how over-fitting can be avoided using techniques like k-fold cross validation, leave-one-out cross validation and regularisation. Problem: We are given a dataset containing various criteria important to get admissions into Master's program and probablity of getting an admit. Dataset is available at `http://preon.iiit.ac.in/~sanjoy_chowdhury/AdmissionDataset.zip` You have already implemented a model using linear regression to predict the probablity of getting the admit.

    1. Implement Lasso regression also known as L1 regularisation and plot graph between regularisation coefficient λ and error **(10 points)**

    2. Implement Ridge regression also known as L2 regularisation and plot graph between regularisation coefficient λ and error **(10 points)**

3. Analyse how the hyper-parameter λ plays a role in deciding between bias and variance.**(5 points)**

4. Analyse how the two different regularisation techniques affect regression weights in terms of their values and what are the differences between the two.**(5 points)**

5. In this part implement regression with k-fold cross validation. Analyse how behavior changes with different values of k. Also implement a variant of this which is the leave-one-out cross validation. **(10 points)**

   Note:The assignment would be evaluated by laying more focus on the report you submit. Include all the observations and graphs you obtain along-with proper reasoning wherever applicable.