CS F320 Foundations of Data Science

Assignment-1

Submission Time & Date: 13:00hrs on 26th Oct 2022

Instructions

- This assignment is a coding project and is expected to be done in groups. Each group can contain at most three members. Make sure that all members in the group are registered to this course.
- This assignment is expected to be done in Python using standard libraries like NumPy, Pandas and Matplotlib. You can use scipy and matplotlib for 1-A. Jupyter Notebook can be used. Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.
- All deliverable items (ex. .py files, .ipynb files, reports, images) should be put together in a single .zip file. Rename this file as A1_<id-of-first-member>_ <id-of-second-member>_ <id-of-third-member> before submission.
- Submit the zip file on CMS on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions/exemptions will be given. The demos for this assignment will be held on a later which shall be conveyed to you. All group members are expected to be present during the demo.

Assignment 1-A Prior and Posterior Distributions

Problem Statement

A survey of whether a customer likes the new update of the software or not was done by a company. Let s denote the probability of a customer liking the new update. Before the survey it was assumed that s follows a beta distribution with parameters α,β = (2,2). Out of the 50 customers surveyed, 40 of them liked the update. Plot the prior and posterior probability distribution of s. After few days another survey is conducted in which out of the 30 customers surveyed 17 of them disliked the update. Plot distribution of s after this survey. What is the prior distribution in this case and justify it with appropriate reasoning. Again, a final survey was conducted in which 70 out of the 100 people surveyed liked the update. Plot distribution of s after the final survey.

What needs to be documented

i) Describe the likelihood of s

ii) Describe in detail how did you find the posterior distribution of s

Assignment 1-B Polynomial Regression and Regularization

Problem Statement

- Aquatic toxicity caused due to manufactured chemicals and other anthropogenic and natural materials severely affects aquatic organisms at various levels of organization. The dataset consists of 2 molecular descriptors: MLOGP and GATS1i which affect the LC50 value (quantitative experimental response).
- Dataset:
  Link:https://drive.google.com/file/d/1nfA1Qet7qOR46tWCGnFR2oHwzWSV-0nO/view?usp=sharing

a) Develop a polynomial regression model (with degrees varying from 0
1, 2,. ., 9) to predict LC50 value based on the two molecular descriptors using Gradient Descent and Stochastic Gradient Descent methods. Before applying the model, shuffle the data and create a random 80-20 split to aid in training and testing. Determine the degree of the polynomial which best fits the given dataset. 437train - 109test

b) Using the same dataset, we know that a polynomial regression model of degree 1 can be built by making use of the following generalized regularized error function:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

Build different regression models by taking q as 0.5,1,2,4.

Experiment with few values of λ to obtain the optimal model for each value of q.

c) Perform a comparative study of the four optimal regularized regression models and classical regression model (best model obtained in (a)).

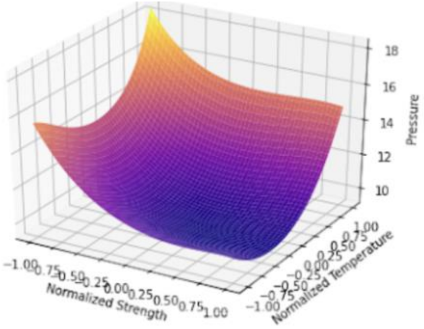
What needs to be documented

i) Give a brief description of your model, algorithms and how you implemented the regularization.

ii) Tabulate the training and testing errors obtained using polynomial regression models of various degrees and your observations on overfitting.

iii) Surface plots of the predicted polynomials (Plot of x_1, x_2 vs y(x_1, x_2) where y is the predicted polynomial.)

Eg of a surface plot



iv) The comparative analysis study of the four optimal regularized regression models and best-fit classic polynomial regression model.

Assignment 1-C Visualizing Regularization

- Aquatic toxicity caused due to manufactured chemicals and other anthropogenic and natural materials severely affects aquatic organisms at various levels of organization. The dataset consists of 2 molecular descriptors: MLOGP (x_1) and GATS1i (x_2) which affect the LC50 value (t) (quantitative experimental response).
- Dataset Link: https://drive.google.com/file/d/1nfA1Qet7qOR46tWCGnFR2oHwzWSV-0nO/view?usp=sharing
- Consider the optimization problem

$$minimise\ E(w) = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2\ subject\ to\ |w_1|^q + |w_2|^q \leq \eta$$

Where $y_n = w_1 x_{n_1} + w_2 x_{n_2}$ *and N is the total number of samples ; and*

x_{n_1} *and* x_{n_2} *represent the first and second features of the* n^{th} *sample*

- Plot the contours of the error function (unregularized) and the constraint regions for q = 0.5, 1, 2 and 4 (Refer to Fig 3.4 of textbook) and η = 1.4, 0.1, 0.035 and 0.052 respectively. Make a plot of error function contours. Also make plots of the constraint regions and error function contours, showing the tangential contour where the minima occurs. Indicate the values of w_1, w_2 at the point of intersection of the tangential contour and the constraint region for which the global minima will be obtained.

- **What needs to be documented**

  i) Error contour plot

  ii) Constraint regions and error function contours, showing the tangential contour and the point of intersection where the minima occurs

  iii) Mean Squared Errors when the hence obtained weights are used to make polynomial model for regression for each case.