Assignment 1

(75 marks)

Instructions

1. This assignment is due by 11:59AM on Monday, 30th August 2021. Submissions are to be made on the helloIITK course portal.
2. No late submissions will be permitted.
3. Please submit your solutions as .zip file containing your code in .ipynb notebooks (one combined notebook with solutions to both problems), all relevant datasets, and a report (if submitting one separately from markdown comments in the notebook) named using only your roll numbers. Thus, if your roll number is 2100234, your submission will be named 2100234.ipynb.
4. Please work out problems on your own. It is ok to consult with others, or the web, but please don't copy paste code.
5. You are only permitted to use **numpy**, **pandas** and **matplotlib** libraries for your assignment.

For all your answers, present both code, explanations of your code, and explanations of your code's outputs.

Q1. For this problem, we will be working with the automobile dataset from the UCI repository. Using this dataset,

(a) train a k-nearest neighbors regression model, and report its validation set performance using root mean squared error. (15 points)

(b) find an optimal k for this model using cross-validation (10 points)

(c) Introduce L0 regularization into this setup and retrain the model (5 points) and,

(d) check whether L0 regularization improves generalization and which are the most important features identified by the model for predicting prices. Comment on your findings drawing upon real-world intuitions about car prices. (10 points)

Note: You don't have to use all the features in the dataset, if you don't want to. Also, the choice of the distance function is entirely up to you.

Q2. For this problem, we will be working with the census income dataset from the UCI repository. Using this dataset,

(a) train a decision tree classification model using information gain as the splitting criterion and using only single feature decision stumps at all non-leaf nodes and majority votes at leaf nodes, and report its validation set performance using % accuracy (15 points)

(b) use cross-validation to optimize the tree hyperparameters (10 points)

(c) Improve on the best test set performance this classifier has to offer with a better version that uses more complex splitting criteria than single-feature decision stumps (10 points)