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)

1. Data preprocessing and normalization (+3 marks)
2. A distance function sensitive to data types is defined (+3 marks)
3. A KNN regression model is defined in the code (+3 marks)
4. Root mean squared error is calculated correctly (+3 marks)
5. Comments
   a. No comments (0 marks)
   b. Comments only as headings of sections (+1 marks)
   c. Comments describing what is being done (+2 marks)
   d. Comments also describing why it is being done (+3 marks)
6. Bonus points for using categorical data types in distance function (+5 marks)
7. Negative marks for using forbidden libraries (-5 marks)

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

1. A held out validation set is created before entering cross-validation (+2 marks)
2. Cross-validation splits are correctly selected (without replacement) (+2 marks)
3. CV is correctly implemented (+3 marks)
4. Optimal k is selected as the one that minimizes the average test set error (+3 marks)

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

1. Identifying the fact that L0 regularization simply means feature selection here (+2 marks)
2. Making any progress towards implementing feature selection in a principled manner (+3 marks)

(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)

1. Interpreting generalization as held out validation set error rather than test set error (+2 marks)
2. Reporting improved generalization (+3 marks)
3. Comments w.r.t. real world considerations (+2 marks)
4. Comments
    a. No comments (0 marks)
    b. Comments only as headings of sections (+1 marks)
    c. Comments describing what is being done (+2 marks)
    d. Comments also describing why it is being done (+3 marks)

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)

1. Data preprocessing and normalization (+3 marks)
2. Information gain calculation is correct (+3 marks)
3. A decision tree learning model is defined in the code (+5 marks)
4. Validation set accuracy is calculated correctly (+1 marks)
5. Comments
    a. No comments (0 marks)
    b. Comments only as headings of sections (+1 marks)
    c. Comments describing what is being done (+2 marks)
    d. Comments also describing why it is being done (+3 marks)
6. Negative marks for using forbidden libraries (-5 marks)

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

1. A held out validation set is created before entering cross-validation (+2 marks)

2. Cross-validation splits are correctly selected (without replacement) (+2 marks)
3. CV is correctly implemented (+3 marks)
4. Optimal hyperparameters selected as the ones that maximizes the average test set accuracy (+3 marks)

(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)

1. Using LwP based decisions (+3 marks)
2. Using combinations of features as decision criteria (+4 marks)
3. Comments
    a. No comments (0 marks)
    b. Comments only as headings of sections (+1 marks)
    c. Comments describing what is being done (+2 marks)
    d. Comments also describing why it is being done (+3 marks)