CS771A: Introduction To Machine Learning

Users Online : 1

Assignment 1

Submitted on 30/8/2021 04:44

Instructions

- You are given an extra 10 minutes after due time to submit your assignment.
- However, please note that any submissions made after the due time are marked as late submissions.

Assignment 1

Question:

See attached pdf for details.

A1.pdf

Uploaded Files:

180032.zip

Grades:

Marks: 62
Feedback:

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 (+2 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
d. Comments also describing why it is being done (+3 marks)
6. Bonus points for using categorical data types in distance function (+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)

(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)
4. Comments
d. Comments also describing why it is being done (+3 marks)

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 (+2 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
d. Comments also describing why it is being done (+3 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)
2. Using combinations of features as decision criteria (+4 marks)
3. Comments
c. Comments describing what is being done (+2 marks)