CSCI 567: Machine Learning

Programming Assignment 1

Decision Trees

Due May 26, 2023, 23:59:59 PST

Overview

This assignment consists of 5 parts:

1. **K-Nearest Neighbors (KNN) [50 marks]** is a supervised learning algorithm used for both classification and regression tasks. It doesn't learn parameters or rules like other methods, but instead predicts outputs based on the similarity of inputs. In classification, a new instance is assigned the most common class among its 'K' closest instances from the training data. In regression, it's assigned the average value of the 'K' nearest instances. The choice of 'K' and proper scaling of input features are crucial for optimal results.
2. **Decision Trees (DTs) [10 marks]** are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. Complete the cells in the notebook provided.
3. **Random Forest Regression [15 marks]**: Decision trees can also be applied to regression problems, using the DecisionTreeRegressor class. Complete the cells in the notebook provided.
4. **Hyperparameter Tuning [5 marks]**: Tune the hyperparameters for the Random forest and print out the best combination of the parameters in the format present in the notebook.
5. **Retrain Random Forest Regressor & Plot Learning Curves [20 marks]:** Retrain the Random Forest Regressor using only the top 5 features and the plot Learning Curve. Additionally calculate mean and standard deviation for training set and test set scores.

Note: There is an FYI section present at the end of the notebook. The content present in that section is just for the students to have some extra information and will not be graded.

Grading and Submission

Students need to implement their code in Python. The marks distribution is given inside the notebook itself. Each of the 5 parts mentioned above has been divided into sub-parts. The maximum marks a student can earn from this assignment is 100 marks. Students will be graded on the submitted notebook. All solutions must be submitted via the DEN platform. **Any other forms of submission won't be entertained.**

Make sure your notebook has the outputs displayed before you submit the assignment to DEN.

Submit the downloaded Colab notebook which shows the results to Desire2learn. You may work in groups of 2-3 for only discussing the high-level concepts related to the homework and not writing the code. Each person should hand in their own code.

There are NO late days for assignments, and we will not accept late submissions.
We won't regrade assignments.

You are free to submit the assignment more than once on D2L, only the last submission will be considered as the final submission. So make sure you start the assignment early and submit it early to avoid any technical issues at the last moment.

Academic Honesty and Integrity

All homework material is checked vigorously for dishonesty using several methods. All detected violations of academic honesty are forwarded to the Office of Student Judicial Affairs. To be safe you are urged to err on the side of caution. Do not copy work from another student or off the web. Keep in mind that sanctions for dishonesty are reflected in *your permanent record* and can negatively impact your future success. As a general guide:

1. **Do not copy** code or written material from another student. Even single lines of code should not be copied.
2. **Do not copy** code off the web. This is easier to detect than you may think.
3. **Do not share** any custom test cases you may create to check your program's behavior in more complex scenarios than the simplistic ones considered below.
4. **Do not copy** code from past students. We keep copies of past work to check for this. Even though this problem differs from those of previous years, do not try to copy from homeworks of previous years.
5. **Do not ask on piazza** how to implement some function for this homework, or how to calculate something needed for this homework.
6. **Do not post code on piazza** asking whether or not it is correct. This is a violation of academic integrity because it biases other students who may read your post.
7. **Do not post test cases on piazza** asking for what the correct solution should be.

Do ask the professor or TAs if you are unsure about whether certain actions constitute dishonesty. It is better to be safe than sorry.