Machine Learning BITS F464 **10-09-2022**

Lecture 4-6: Decision Tree, k-Nearest Neighbours, Crossvalidation

Lecturer: Harikrishnan N B *Student:*

FDTA: Devale Tanmay Abhijit, Biyani Param Hemant Kumar

READ THE FOLLOWING CAREFULLY:

Honour Code for Students: I shall be honest in my efforts and will make my parents proud. Write the oath and sign it on the assignment.
Deadline for Assignment Submission: 07:00 PM, 16 September 2022. (No assignments will be accepted after the deadline. The submission of the assignment is via online. The google doc link to submit assignment will be provided soon. The file name has to be named in the following format: rollno_firstname_lastname. Question 1 carries 15 marks, Question 2 carries 15 marks - provide the code snippet wherever the coding is compulsory.

Question 1

Consider the following dataset provided in Figure -2.1 corresponding to a binary classification problem. The features/ attributes are denoted as f_1, f_2. Table -2.1 provides data information.

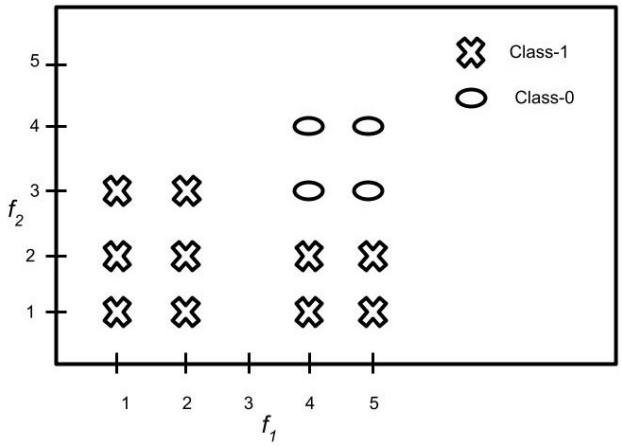

Figure -2.1: Toy Example: A binary classification problem where the circles represent the data instances belonging to class-0 and crosses represents the data instances belonging to class-1.

- Write a python code to plot the dataset (use matplotlib) provided in Figure -2.1 (Table -2.1).
    1. **Mandatory Checklist:** Make sure the fontsize of xlabel, ylabel, xticks, yticks, and legends are visible (If one takes a print out of your assignment, one should be able to clearly read the values).
    2. Make sure to use circles and crosses for data instances belonging to class-0 and 1 respectively.

Table -2.1: Dataset details for the toy example provided in Figure -2.1.

Data Instances.	(f_1, f_2)	**Class Labels/ Target Value**
x_1	(1, 1)	1
x_2	(1, 2)	1
x_3	(1, 3)	1
x_4	(2, 1)	1
x_5	(2, 2)	1
x_6	(2, 3)	1
x_7	(4, 1)	1
x_8	(4, 2)	1
x_9	(4, 3)	0
x_{10}	(4, 4)	0
x_{11}	(5, 1)	1
x_{12}	(5, 2)	1
x_{13}	(5, 3)	0
x_{14}	(5, 4)	0

3. Make sure you provide comments to the codes you have written. (The reader should be able to understand the code by reading the comments).

- Calculate the Shannon entropy of the labels (hand calculation)
- Write a computer program to compute the Shannon entropy. Verify the correctness of your code by comparing with the Shannon entropy of the labels obtained by hand calculation (Table -2.1).
- Write a computer program to compute the gini impurity. Verify the correctness of your code by comparing with the gini impurity of the labels obtained by hand calculation (Table -2.1).
- What is your intuition on how the decision tree classifier (ID3) will split the data so as to get pure leaf nodes? Draw the decision boundary?
- Verify your intuition by writing a computer program (from scratch) to construct the decision tree for the dataset provided in Table -2.1.
- Plot the decision boundary (use matplotlib).
- Use the inbuilt sklearn decision tree package and plot the decision tree and the decision boundary for the dataset provided in Table -2.1 (you can use the inbuilt package provided in sklearn).
- Compare the results obtained using coding from scratch and coding using inbuilt decision tree package.

Question 2

Consider the Iris dataset. The dataset is available here: `https://scikit-learn.org/stable/auto_examples/datasets/plot_iris_dataset.html`.

- Write a small paragraph describing the Iris dataset.
- Identify the features/ attributes in Iris dataset?

- Identify the total number of classes in Iris dataset?
- In a table, summarize the total data instances of each class (Remember table and figure should have self contained appropriate captions.)
- Split the Iris dataset randomly into training (80%) and testing (20%) (you can use sklearn train-test split - randomseed= 42).
- In a table, provide the number of data instances used for training and testing for each class.
- Using the train data (obtained after splitting the total data into training and testing), perform three fold crossvalidation to find the best value of k in k Nearest Neighbour classifier (the k value can range from 1 to 25, and use euclidean norm to compute the distance). The algorithm to find k-nearest neighbours should be written from scratch (You can use the k-fold crossvalidation package provided in sklearn for hyperparameter tuning - `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html`).
- Plot the average macro f1-score obtained using three fold crossvalidation with respect to the different values of k considered in three fold crossvalidation.
- Identify the best value of k for which you get the peak performance in three fold crossvalidation.
- Using the best value of k, evaluate the performance of the k nearest neighbour classifier on the testdata (Remember testing should be done only once!).
- Report the test accuracy, precision, recall, f1-score and macro f1-score.

Provide the code snippet for questions where coding is compulsory! The submission of the assignment is via online. The google doc to link to submit assignment will be provided soon. The file name has to be named in the following format: rollno_firstname_lastname.

Write your name and roll no. in the front sheet of your assignment.