COSC 425: Intro to Machine Learning
Lab 1: Python Data Manipulation and Visualization and Decision Trees

Due: September 9, 2022, 11:59 PM

Introduction

In this lab, you will be practicing using Python, jupyter notebooks, numpy, pandas, and matplotlib to process, manipulate, and visualize data. You will also use the sklearn implementation of decision trees to explore the parameters associated with decision trees and the application of decision trees.

You will create a Jupyter notebook for each part (parts 1 and 2) that you will submit where you will put the code you wrote to answer each of the questions described below. You will submit a separate written report in PDF form in which you include the answers to the questions and/or the plots required by the question.

The dataset we will be using in this component of the project is the Life Expectancy (WHO) dataset. (URL: `https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who`)

You should use pandas to load these data files into Python, pandas and numpy to manipulate the data when making calculations, and matplotlib to create the plots.

Part 1: Python Data Manipulation and Processing

Question 1.1 (5 points): Plot the boxplot for life expectancy in developing and developed countries in 2015. Label y-axis and change xticks to appropriate labels for the boxes (Developing and Developed). You should have two boxes in the plot: one for developing and one for developed.

Question 1.2 (5 points): Create a scatter plot between life expectancy and GDP in 2015. Label axes appropriately.

Question 1.3 (10 points): Create a stacked histogram for BMI in 2015, where the stacked bars are on status (developed and developing). Don't forget to include a legend.

Question 1.4 (10 points): Create a solid line plot showing the average (across all nations) life expectancy changing over time, with the standard deviation above and below shown with filled between plot (use fill_between plotting function with alpha = 0.3). Additionally, plot the maximum life expectancy for each year and the minimum life expectancy for each year as dotted line plots over time. Include legend for mean, minimum, and maximum, and don't forget axes labels.

Question 1.5 (15 points): Extract Hepatitis B, Polio, Measles, BMI, Diphtheria, HIV/AIDS for 2015 and remove the rows with NaN elements (use dropna()). Create the correlation coefficient matrix (use np.corrcoef()) and create a heatmap showing the correlations, using pcolor() and matplotlib. Label the rows and columns with the appropriate diseases. Include a colobar and a label on the colorbar. Which of these two are the most heavily correlated?

Extra Credit: For up to 5 points of extra credit: Define a question about the data that you would like to answer and include a plot and calculations to illustrate the answer to the question.

Part 2: Decision Trees

We are going to create a dataset to do a decision tree classification from this dataset. Our features are going to be the following:

- Hepatitis_B
- Polio
- Measles
- BMI
- Diphtheria
- HIV/AIDS
- Adult_Mortality

Our labels are going to be the Status category (Developed or Developing) for a binary classification problem.

First, load in the data, and drop all rows that have NaN values. Then create X and y numpy arrays, where X is a matrix so that all of the rows are the different rows from the dataframes and the columns are the features listed above, and y is the numpy array of the Status column. Once you have completed this, X and y should both have 1649 rows.

The next step is to create a train test split on the data. We're going to use a test size of 0.33 and a random state of 42.

Question 2.1 (15 points): Create a decision tree classifier with the entropy criterion and fit to the training data. What is the accuracy score of this classifier on the testing set? What is the first decision that is used to split the data (which feature does it use and what value does it split on)?

Question 2.2 (15 points): Create decision trees with max depths from 2 to 20 and calculate the training and testing accuracy for each decision tree and display in a table. Which one is best? Provide a short (1-2 sentence) explanation for why that depth is the best performing.

Question 2.3 (5 points): Show the diagram of the decision three with max depth of 2.

Question 2.4 (5 points): For the best performing tree in Question 2 (in terms of testing results), show the confusion matrices that are produced for the training set and testing set.

Question 2.5 (15 points): Create two additional decision trees, where min samples leaf set is set to 10 and min samples split set to 2 and a second where min samples leaf set is set to 10 and min samples split set to 30. Calculate the training and testing scores for each. Which of these performs better? Provide a short explanation (2-3 sentences) for why one performs better than the other.

Submission Checklist

In this lab, you will submit on Canvas:

- Jupyter notebook for part 1
- Jupyter notebook for part 2
- PDF of your report for both parts

Late Penalty: 20 points off per day late.