Homework #2

DS 3001, D-term 2018

100 points total [8% of your final grade]

Due: April 5, 2018 by 11:59pm
[no submission will be accepted after April 8, 2018 at 11:59pm]

Delivery: Submit via canvas

Overview

In the first homework, you got your hands dirty with some basic data cleaning and exploratory data analysis. For this homework, you are going to extend that workflow to also include the application of a standard data mining approach: decision tree classification.

This homework is actually quite important to the defense of our planet. You see, there is strong evidence of alien beings visiting us, as you'll find in this list of UFO sightings collected by the National UFO Reporting Center. Your goal in this homework is to gain a deeper insight into these sightings, ultimately giving us the tools to accurately predict the shape of a UFO based on only a few simple features. Good luck. We are all depending on you.

Task 1: UFO Data Collection, Cleaning, and Exploratory Analysis (50 points)

Your first task is to collect, clean, and explore data from the UFO sightings database. In particular, we are especially interested in UFOs corresponding to one of **three shapes**:

- Circle
- Triangle
- Fireball

For each UFO shape, you should collect all sightings from the list of UFO sightings made **between January 1, 2005 and September 22, 2016**. In other words, ignore any sighting made after September 22, 2016 or before January 1, 2005.

You should represent each sighting by these eight features:

- Date of Sighting
- Time of Sighting
- City
- State
- Shape
- Duration
- Summary
- Posted Date (when the sighting was posted to the website)

As part of your data collection and cleaning, you should do your best to convert all Durations to seconds, whenever possible. Keep in mind a few guidelines:

- If a duration has a "<" sign, you should simply ignore the "<" sign. For example if the duration is specified as "< 1 minute", consider the duration to be "1 minute". You should subsequently convert "1 minute" to "60 seconds".
- If a duration has a range, use the upper limit as its value. For example, if the duration is listed as "5-8 minutes", you should consider the duration as "8 minutes". (Again, you will need to eventually convert minutes into seconds).
- You may encounter some other oddities in the data. Do your best to extract maximum value from the messy data; be sure to explain to us the decisions you have made in terms of data extraction and cleaning.

Based on your cleaned data, you should perform a basic exploratory data analysis to better understand what you've got. Specifically, we expect to see the following:

- A boxplot of the duration of UFO sightings of each shape (one boxplot per shape).
- A time series figure with the number of sightings per year (one line per shape).
- A bar chart for sightings by state.

Of course, those are just some basic steps you will need to take as the "data scientist" in charge of this important challenge. In addition, you should also identify some other interesting insights from the data. For example, you might:

- Normalize the sightings by state population. What do you observe? Anything interesting?
- Visualize the distributions on a map (e.g., using Tableau, Google Maps API, basemap, or D3). Do you notice anything peculiar?

These are just two suggestions. You are encouraged to explore the data based on your own intuitions, but you are required to ask and answer **at least one additional question** beyond the basic data analysis we require above (boxplots, time series, bar chart). Remember, the people of Earth depend on you to draw interesting insights from the data!

Task 2: Predicting UFO Shape (50 points)

Given your understanding of the data, your goal is now to build a decision tree classifier to predict the shape of a UFO. You have three target classes: circle, triangle, and fireball. For this task, you need only consider two simple features to represent each sighting:

- **Region of the country:** We shall divide the 50 states into the four Census Bureau-designated areas: the Northeast, Midwest, South, and West.
- **Time of Day:** We shall consider only four parts of the day. Night (00:00-05:59), Morning (06:00-11:59), Afternoon (12:00-17:59), and Evening (18:00-23:59).

That is, each sighting will simply be represented by its region and time of day. E.g., (South, Morning) or (Midwest, Evening).

Next, split your dataset to training set and test set. Training set consists of all sightings made between **January 1, 2005 and December 31, 2013**. Test set consists of all sightings made between **January 1, 2014 and September 22, 2016**.

Based on these two features with the training set, you should implement a decision tree classifier that uses Gini Impurity to determine the best feature at each step.

- You should report the classification accuracy for your decision tree using the test set.

- You should provide an illustration of the decision tree (built based on your training set). You may use a graphing toolkit (like networkx) or you may draw the tree manually.

Task 3: Improving Your Accuracy (Additional 5 points. This task is optional)

Can you improve your prediction rate (accuracy) over what you got from Task 2? You may use raw features instead of the two features or even combining all features. Or something else?

- Describe how you can achieve better result compared with Task 2's result.
- Report what result you got.
- What feature is the most important feature to distinguish shapes of UFOs?

What to turn in:

- You should turn in ONE file: a PDF report with your answers to the homework questions (hw2_yourname.pdf).
- This is an individual assignment, but you may discuss general strategies and approaches with other members of the class (refer to the syllabus for details of the homework collaboration policy). At the top of your report, please write the names of classmates you consulted and the nature of your discussion.