CSCI 3223: Tashfeen's Machine Learning II

Homework 2

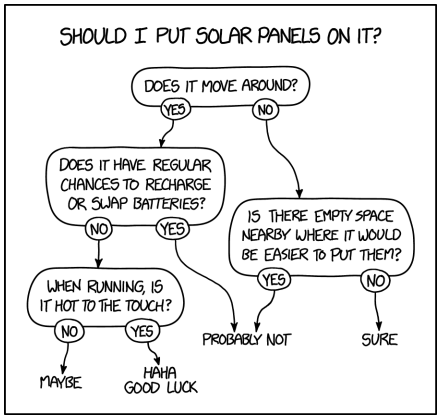

FIGURE 1. XKCD Comic 1924: https://xkcd.com/1924/

In this assignment, we'll learn about classification and (not as intuitively) regression using *decision trees*. Please do not use any other libraries than Numpy, Matplotlib and Scikit-learn.

CLASSIFICATION TREES

Figure 1 shows an example classification decision tree. Consider the data given in Table 1, we could make two possible decision trees given in Figure 2. The difference between the two trees is the first splitting question, or simply, the feature we select to split over. If we follow the Occam's razor, the tree with only one split is better. Generalising the question we need to answer when we are building a decision tree for a dataset, we ask,

What is the best feature to split the data over at a given level of the tree?

To best answer this question, we use two important concepts from information theory (the branch of mathematics most responsible for artificial intelligence),

1) Entropy $\propto \frac{1}{\text{Purity}}$
2) Information Gain

Instance	Classification	a_1	a_2
1	+	T	T
2	+	T	T
3	+	T	F
4	−	F	F
5	−	F	T
6	−	F	T

TABLE 1. Dataset S with attributes a_1 and $_2$.

Page 55 of *Machine Learning*, Tom Mitchell's Equation 3.1 defines entropy for two classes as follows:

$$\text{Entropy}(S) = -p_{\oplus} \lg p_{\oplus} - p_{\ominus} \lg p_{\ominus}$$

Note that when $p_{\oplus} = 0.5 = p_{\ominus}$, we have entropy $-0.5 \lg (0.5) - 0.5 \lg (0.5) = 1$. Where $p_{\oplus}$ and $p_{\ominus}$ are the probabilities of the positive and negative classes in the data S. Furthermore, the equation 3.4 (pg. 58) gives the definition of the information gain for a split over an attribute $A = a_i$. Observe that if A is a binary attribute then $\text{Values}(A) = \{\text{F}, \text{T}\} \ni v$ and S_v are rows where $A = v$.

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Values}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

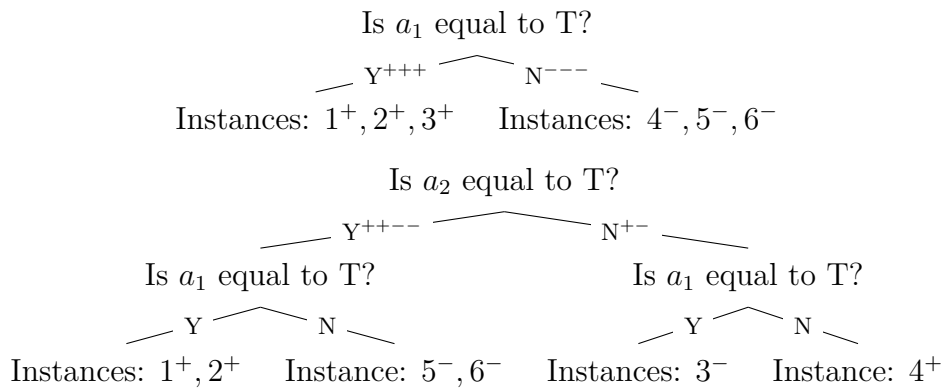

FIGURE 2. Possible decision trees for Table 1.

Question 1. Consider the data set S in table 1.

1) What is the entropy of S?
2) What is the Gain(S, a_2)? Should we split over a_1 or a_2? Justify your answer.

Question 2. Pick the columns, "pclass, sex and age" from the cleaned Titanic Dataset and remove any rows with empty values. Encode "female" as 0 and "male" as 1. Hold out 50 records corresponding to the people who survived and 50 for those who unfortunately did not. You'll be using these $50 + 50 = 100$ records for testing. Feel free to use the Python file `data.py` from the previous homework.

Use the Scikit-learn's `DecisionTreeClassifier` with the training data to build a decision tree of max depth equal to 4 (starting counting at 0). Put the code in a file called `titanictree.py`.

(a) What is the accuracy on the testing set?
(b) Plot the tree using `sklearn.tree.plot_tree`. Let `fontsize` be 10 while also setting the class variables `feature_names` and `class_names` appropriately. Also set `plt.figure(figsize=(24, 8))` and `plt.tight_layout()` before the call `plt.show()`. Give this plot.
(c) What is the probability that a first class male infant survived the shipwreck?

Question 3. Fisher's Iris data set consists of 150 rows and $4 + 1$ columns where four columns correspond to lengths and one to an Iris species. There are fifty records for each of the: Setosa, Versicolor, Virginica.

i	Sepal Length (cm)	Sepal Width (cm)	Petal Length (cm)	Petal Width (cm)	Species
1	5.1	3.5	1.4	0.2	Setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	7.0	3.2	4.7	1.4	Versicolor
⋮	⋮	⋮	⋮	⋮	⋮
101	6.3	3.3	6.0	2.5	Virginica
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	Virginica

TABLE 2. Ronald Fisher's Iris Flower Data (`loadiris.py`).

You can see the Sepal and Petal annotated in Figure 3 (right) while the left side shows the relationship between the sepal and the petal lengths. Give a two-split classification decision tree that classifies Irises into Setosa, Versicolor, Virginica where entropy in any given leaf node is at most 0.4. Give the entropy for each of your leaf nodes. Put any code for this question in a file called `iristree.py`.

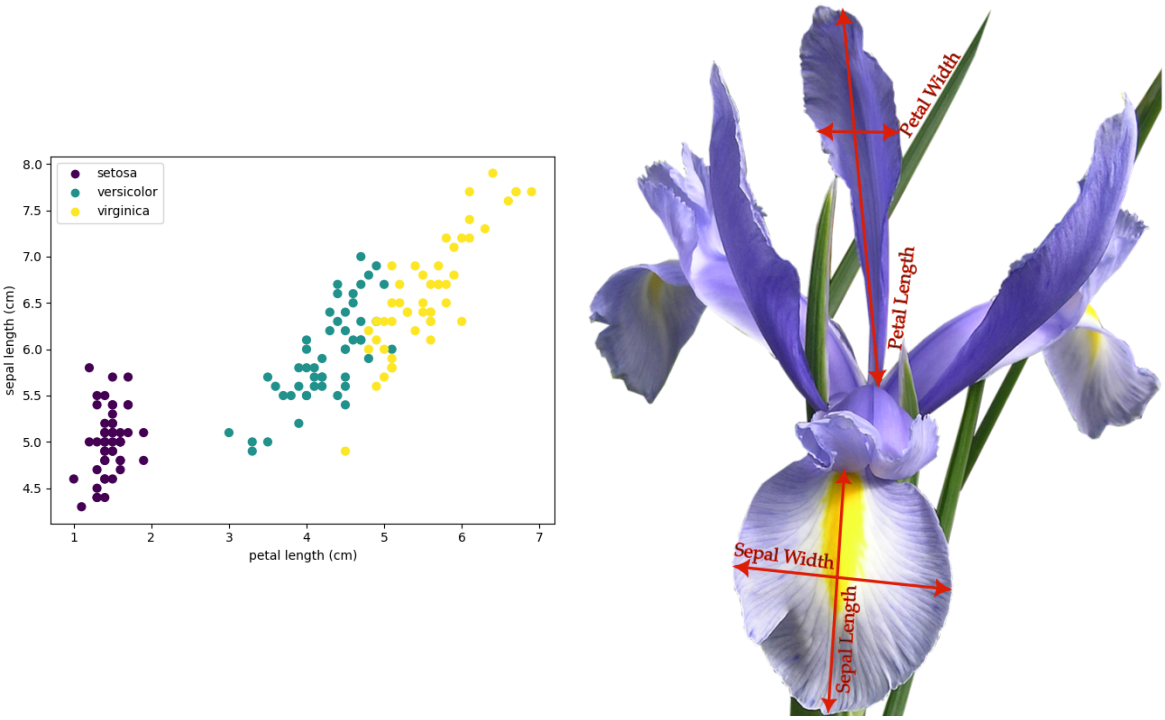

FIGURE 3. Iris' sepal length against its petal length (left). Annotated anatomy of an Iris flower (right).

REGRESSION TREES

For regression, let prediction for a leaf node ℓ_i be the mean of all the target values that fall in that leaf. As well as the mean squared error for that leaf,

$$\text{MSE}(\ell_i) = \frac{1}{|\ell_i|} \sum_{y \in \ell_i} (\hat{y} - y)^2 \quad \text{where} \quad \hat{y} = \frac{1}{|\ell_i|} \sum_{y \in \ell_i} y.$$

The error of a regression tree can be defined as the sum of errors in all the leaves.

Example 1. If we split the toy data in Table 3 over the question, "Enrollment ≥ 25," we get,

$$\begin{array}{c} \text{Enrollment} \geq 25 \\ \text{Y} \quad\quad \text{N} \\ i\text{: } 2\ 3\ 4\ 5\ 8\ 9 \quad\quad i\text{: } 0\ 1\ 6\ 7 \end{array}$$

Hence the two leaves give following number of pets,

$$\hat{y}_{\text{Y}} = \frac{2+10+5+0+3+3}{6} = \frac{23}{6} = 3.8\overline{3}$$
$$\hat{y}_{\text{N}} = \frac{0+0+0+1}{4} = \frac{1}{4} = 0.25$$

Calculating error,

$$\begin{aligned} \text{MSE}(\ell_{\text{Y}}) &= \frac{(3.8\overline{3}-2)^2 + (3.8\overline{3}-10)^2 + (3.8\overline{3}-5)^2 + (3.8\overline{3}-0)^2 + (3.8\overline{3}-3)^2 + (3.8\overline{3}-3)^2}{6} \\ &= \frac{58.8\overline{3}}{6} = 9.80\overline{5} \\ \text{MSE}(\ell_{\text{N}}) &= \frac{(0.25-0)^2 + (0.25-0)^2 + (0.25-0)^2 + (0.25-1)^2}{4} \\ &= \frac{0.75}{4} = 0.1875 \end{aligned}$$

Therefore the total error is $9.80\overline{5} + 0.1875 = 9.9930\overline{5}$.

i	Enrollment	Hour of the Day (0: AM, 1: PM)	Class Level (0: U, 1: G)	Number of Pets
0	20	0	0	0
1	12	1	1	0
2	50	0	0	2
3	100	1	0	10
4	75	1	0	5
5	30	0	1	0
6	5	0	1	0
7	12	0	1	1
8	30	1	1	3
9	25	1	1	3

TABLE 3. Toy student data to predict the number of pets.

Question 4. With each of the given root questions for the toy data in Table 3,

1) Enrollment ≥ 50
2) Hour of the Day is 1
3) Class Level is 1

Calculate the respective splits and errors for each of their leaves and give the total tree error as shown in Example 1. What is the best root split criterion to grow a regression tree?

SUBMISSION INSTRUCTIONS

1) Submit a PDF that answers any questions and contains all the plots that the assignment asks for.
2) Submit your `titanictree.py` and `iristree.py`.