Tutorial 8: Decision Trees

1. Consider the example "Play Golf" in the lecture note

ID	Outlook	Temperature	Humidity	Windy	Play Golf
1	Rainy	Hot	High	FALSE	No
2	Rainy	Hot	High	TRUE	No
3	Overcast	Hot	High	FALSE	Yes
4	Sunny	Mild	High	FALSE	Yes
5	Sunny	Cool	Normal	FALSE	Yes
6	Sunny	Cool	Normal	TRUE	No
7	Overcast	Cool	Normal	TRUE	Yes
8	Rainy	Mild	High	FALSE	No
9	Rainy	Cool	Normal	FALSE	Yes
10	Sunny	Mild	Normal	FALSE	Yes
11	Rainy	Mild	Normal	TRUE	Yes
12	Overcast	Mild	High	TRUE	Yes
13	Overcast	Hot	Normal	FALSE	Yes
14	Sunny	Mild	High	TRUE	No
15	Rainy	Mild	Normal	FALSE	?
16	Rainy	Mild	High	TRUE	?
17	Overcast	Hot	High	TRUE	?
18	Sunny	Hot	High	TRUE	?

In the lecture, we have discussed on how to choose the root node. Now for each branch of the root node, use error rates to find the following features to split the branch further until the decision tree is completed.

2. Consider a simple MLP with one hide layer of two neurons and the output layer as shown below:

- $H1=\max(0,x)$, $H2=\max(0,y)$
- $O = H1+H2-1$

Here (x, y) are points in a plane (2d space) and are labelled positive if their output (i.e., $O(x,y)$) are positive, or negative otherwise. Now use three features,
- $F1=x-1$,
- $F2= y-1$, and
- $F3=x+y-1$

which are labelled positive if their values are greater than 0 and labelled negative otherwise.

Draw a decision tree of the MLP classifier to classify the points in 2D space.

3. Training Decision Trees with Scikit-Learn

   a. Read the tutorial on training decision trees with Scikit-Learn at https://scikit-learn.org/stable/modules/tree.html, load the iris data and split the data into training and testing. Use the training set to train the decision tree and test the performance of the learnt decision tree on the testing data. To split the data, you might need to use the train_test_split command (see https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html).

   b. Read the document regarding cross-validation in the link below
      https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
      and implement an algorithm to compute the five-fold cross-validation performance of the decision tree method on the IRIS data.