**CS 535/EE 514 – Machine Learning**

**Assignment 2**

Deadline: Oct. 2 (Monday) at 11:55 PM

### Question 1

(Suboptimality of ID3)

Consider the following training set, where X = {0,1}3 (3 binary attributes) and Y = {0,1}:

((1,1, 1), 1)

((1,0, 0), 1)

((1,1, 0), 0)

((0,0, 1), 0)

Suppose we wish to use this training set in order to build a decision tree of depth

2 (i.e., for each input we are allowed to ask two questions of the form (xi = 0?)

before deciding on the label).

**1.** Suppose we run the ID3 algorithm up to depth 2 (namely, we pick the root node and its children according to the algorithm, but instead of keeping on with the recursion, we stop and pick leaves according to the majority label in each subtree). Assume that the subroutine used to measure the quality of each feature is based on the entropy function (so we measure the information gain), and that if two features get the same score, one of them is picked arbitrarily.

Show that the training error of the resulting decision tree is at least 1/4.

**2.** Find a decision tree of depth 2 that attains zero training error.

### Question 2

[Handwritten answers only for this question]

Using the ID3 algorithm build a three-level (root, intermediate, and leaf nodes) decision tree for the balloon classification problem. There are 4 attributes: color, size (of balloon), action (on balloon) and age of person. And two class labels inflated or not. For inflated label, this condition must be true *(color=yellow and size = small) or (age=adult and act=stretch)*.

Download the dataset from [here](https://archive.ics.uci.edu/ml/machine-learning-databases/balloons/yellow-small+adult-stretch.data).

### Question 3

For this exercise we will be using python 2.7.

Please install the anaconda distribution (it comes loaded with the python interpreter)

<https://www.anaconda.com/download/>

The distribution comes preloaded with popular machine learning libraries. We are mostly interested in pandas and sklearn.

You will be using jupyter notebooks for this assignment. Jupyter notebooks are a way to use python in an interactive fashion. You can find plenty of tutorials online on how to use them.

We will be using the following dataset:

<https://archive.ics.uci.edu/ml/datasets/Car+Evaluation>

Please download and load the data using pandas.

a): Learn rules from the data using:

* At least 2 different splitting criteria
* At least 2 different stopping criteria
* Access the feature importance attribute of the model, what does it signify?

You will need to report performance metrics ( accuracy, precision recall ) using a 80:20 train-test split as well as 10-fold cross validation.

b): Implement a function that returns the best feature in a dataset using information gain in python.

c) What is feature normalization? Will it improve the results you reported above? Please explain your answer.

### Question 4

We will use the orange data mining tool for this part of the assignment:

Repeat question 3 on this dataset: <https://archive.ics.uci.edu/ml/datasets/Nursery>

Show the visualization of the full decision tree in each case.

https://orange.biolab.si/