CIS 472/572, Spring 2019
Homework 1 (Written): Decision Trees
DUE DATE: Submit via Gradescope by Thursday, April 11th at
11:00pm.
Your answers should be typewritten, except for figures which may
be hand-drawn.

1. Answer Exercise 3.1 from Chapter 3 of Mitchell's machine learning book.
   This chapter can be accessed from the following link:
   http://www.cs.princeton.edu/courses/archive/spr07/cos424/papers/mitchell-
   dectrees.pdf

2. Consider the samples in the Play-tennis dataset from Table 3.2 in Mitchell's
   textbook (linked above). If you calculate the information-gain for all of
   the attributes of this set, you will observe that the attribute "Outlook"
   has the largest information-gain, which is equal to 0.246. Therefore, the
   attribute "Outlook" is the best heuristic choice for the root node.

   (a) List the labels of the new tree branches below the root node.

   (b) Which partition of the data will be assigned to each branch by ID3?
       Please list the sample IDs that will be assigned to each branch.

   (c) Calculate the information gain for the remaining attributes in each
       branch, and determine which attribute will be chosen as the root of
       the sub-tree in each branch.

3. Suppose a bank makes loan decisions using two decision trees, one that
   uses attributes related to credit history and one that uses other demo-
   graphic attributes. Each decision tree separately classifies a loan appli-
   cant as "High Risk" or "Low Risk." The bank only offers a loan when
   both decision trees predict "Low Risk."

   (a) Describe an algorithm for converting this pair of decision trees into
       a single decision tree that makes the same predictions (that is, it
       predicts non-risky only when both of the original decision trees would
       have predicted non-risky).

   (b) Let n_1 and n_2 be the number of leaves in the first and second decision
       trees, respectively. Provide an upper bound on n, the number of
       leaves in the single equivalent decision tree, expressed as a function
       of n_1 and n_2.