CISC-481/681 Project 3: Decision Trees

Dylan Chapp, Michael Wyatt
Department of Computer and Information Sciences
University of Delaware - Newark, DE 19716
Email: {dchapp}, {mwyatt}@udel.edu

I. Introduction

II. Implementation

We implemented a system in Python to ingest data in `.csv` form and construct a decision tree from that data using the ID3 algorithm. Additionally, modules for discretization of non-categorical features, pruning, k-fold cross validation, and visualization of the decision tree were implemented. Though we always use the ID3 algorithm, we experimented with two distinct ways of representing the decision tree–in one case an object-oriented approach with explicit node objects, and in the other as a set of nested dictionaries–in order to compare their relative performance.

A. The ID3 Algorithm

The goal of the ID3 algorithm is to construct a decision tree classifier. To that end, the algorithm considers subsets of the training data–initially the entire set–and selects a feature on which to partition the subset. Selection of a feature is equivalent to creation of a node in the tree with descendent edges for each value the feature can assume. The algorithm then recurses on each branch; considering the subset of the training data that matches the value on the branch and excluding the previous feature from consideration. Below, we provide pseudocode that explicitly describes the ID3 algorithm for building a decision tree: