UW Math 480 Final Project

Ayla Lampard, Jason Uanon

May 19, 2013

1 Introduction

1.1 Motivation

Figure 1 shows an example of a decision tree.

1.2 Data

We will be using the Adult Data Set from the UCI Machine Learning Repository [1]. This data is freely-available online and comes from the U.S. Census Bureau from 1994. It contains over 32,000 training instances and 16,000 test instances (although it does contain missing values, denoted with "?". The number of instances with missing values, however, is low enough that we could simply not consider them).

Using this data, we will create a decision tree that will predict whether a person's income exceeds $50,000 per year. The data itself contains 14 attributes, which are:

- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

2 Problem Setup

2.1 Definitions

Entropy measures the amount of disorder in a random variable [2]. Let X be a random variable with

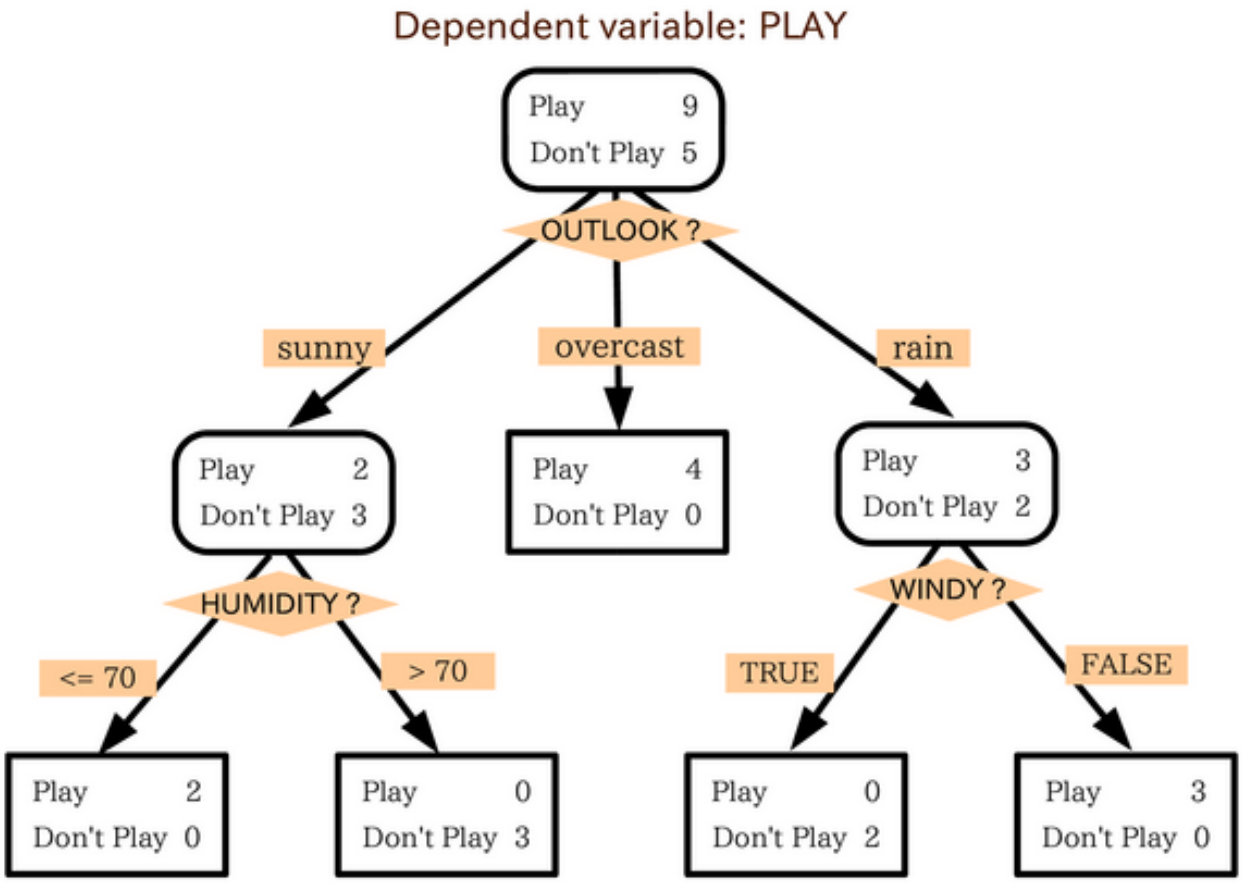

Figure 1: Example of a decision tree

$p(x)$ the probability that $X = x$. Mathematically, it can be expressed as

$$\begin{aligned} H(X) &= \sum_{i=1}^{n} p(x_i) \log_2 \left(\frac{1}{p(x_i)} \right) \\ &= -\sum_{i=1}^{n} p(x_i) \log_2 \left(p(x_i) \right) \end{aligned}$$

The **conditional entropy** of a random variable X (with events x_i) conditioned on a random variable Y (with events y_j) is

$$H(X|Y) = -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j|x_i) \log_2 \left(p(y_j|x_i) \right)$$

The **information gain** of a random variable X conditioned on a random variable Y is

$$IG(X) = H(Y) - H(Y|X)$$

2.2 The Algorithm

We will be using the ID3 Decision Tree algorithm using information gain as the splitting criteria.

- Start from the empty decision tree
- Select the next best attribute i that maximizes information gain (i.e., maximizing $IG(X_i)$)
- Recursively build the children of the root node

2.3 Implementation

We will be using Python 2.7.3 for the implementation of the decision tree algorithms. As part of the project, we will also create a Cython version using Cython 0.15.1 to enhance its performance, and compare the relative speeds of the two implementations.

References

[1] Bache, K. and Lichman, M. (2013). *UCI Machine Learning Repository: Adult Data Set*. [http://archive.ics.uci.edu/ml/datasets/Adult]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Segaran, Toby. *Programming Collective Intelligence*. O'Reilly, California, 2007.

[3]