Michael Baldwin

Josh Engelsma

Adam Terwilliger

March 15, 2016

CIS 678 – Machine Learning

Project 3

Abstract

We look to decision trees in CIS 678 – Machine Learning, as we demonstrate our expertise with a handful of unique datasets. Using Python, we were able to develop a supervised learning decision tree model that utilizes the Iterative Dichotomiser 3 (ID3) algorithm. Additionally, we explored an additional focus of splitting the dataset based on continuous data in addition to categorical. We demonstrated the validity of our approach using foundational datasets in machine learning like that of Iris and Mushrooms. To interpret our model, we utilized the Javascript library of Data-Driven Documents (D3) to create interactive web-based trees. We showcased the strength of our model by using a 50-50 training/test split on Mushrooms predicting 97% correct decisions.

Implementation details

Our program is written in Python 2.7 and D3 (Javascript). These programs were executed locally on each member's respective Macbook Pro (2012), hosting the html files on eos23.

Summary of Problem

Decision Tree models are a form of supervised learning that look to classify observations in a more interpretable way than most models through its transparency. By utilizing the ID3 greedy approach, we can find relatively simple and concise trees that demonstrate reasonable classification rates that allow for valuable insights from those in the domain. ID3 takes advantage of two main attributes, entropy and information gain, in equations (1) and (2). We can infer from these equations that entropy takes advantage of log probabilities, while information gain looks to differentiate between attributes.

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

Equation 1. Formula of Entropy (information theory)

$$IG(A, S) = H(S) - \sum_{t \in T} p(t) H(t)$$

Equation 2. Formula for Information Gain (information theory)

Results

```
[Joshuas-MacBook-Pro:src joshuaengelsma$ python decision_tree.py ../data/t-mushroom.data ../data/p-mushroom.data
Correct: 3942
Total: 4062
Percentage: 0.970457902511%
```

Figure 1. Classification Rate for Mushrooms Data

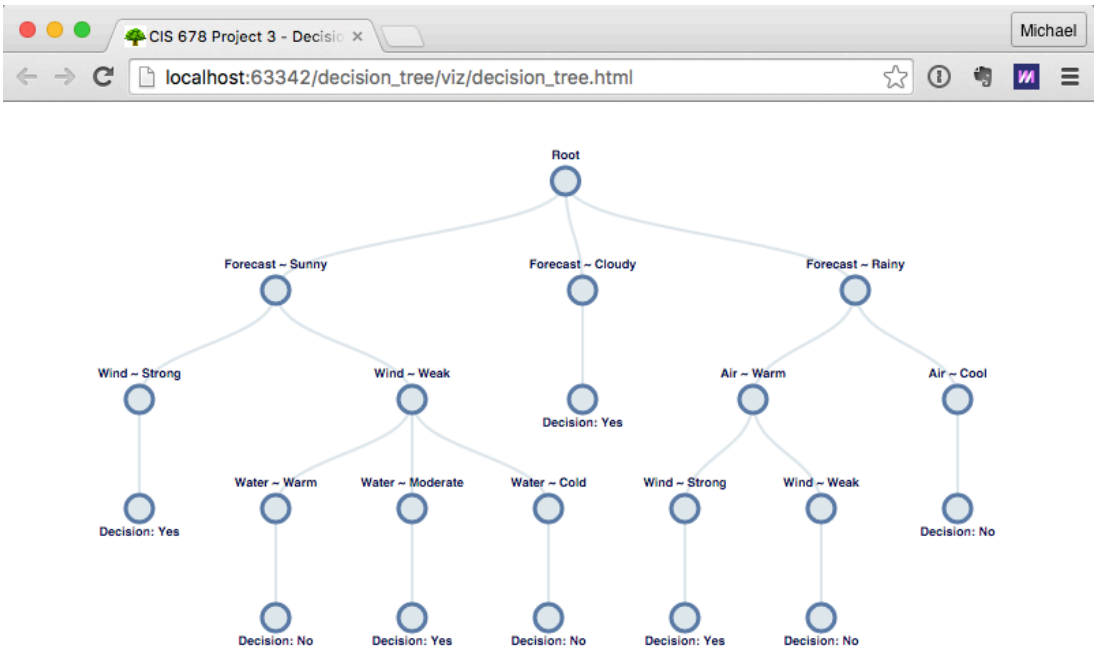


Figure 2. Decision Tree for Fishing Data

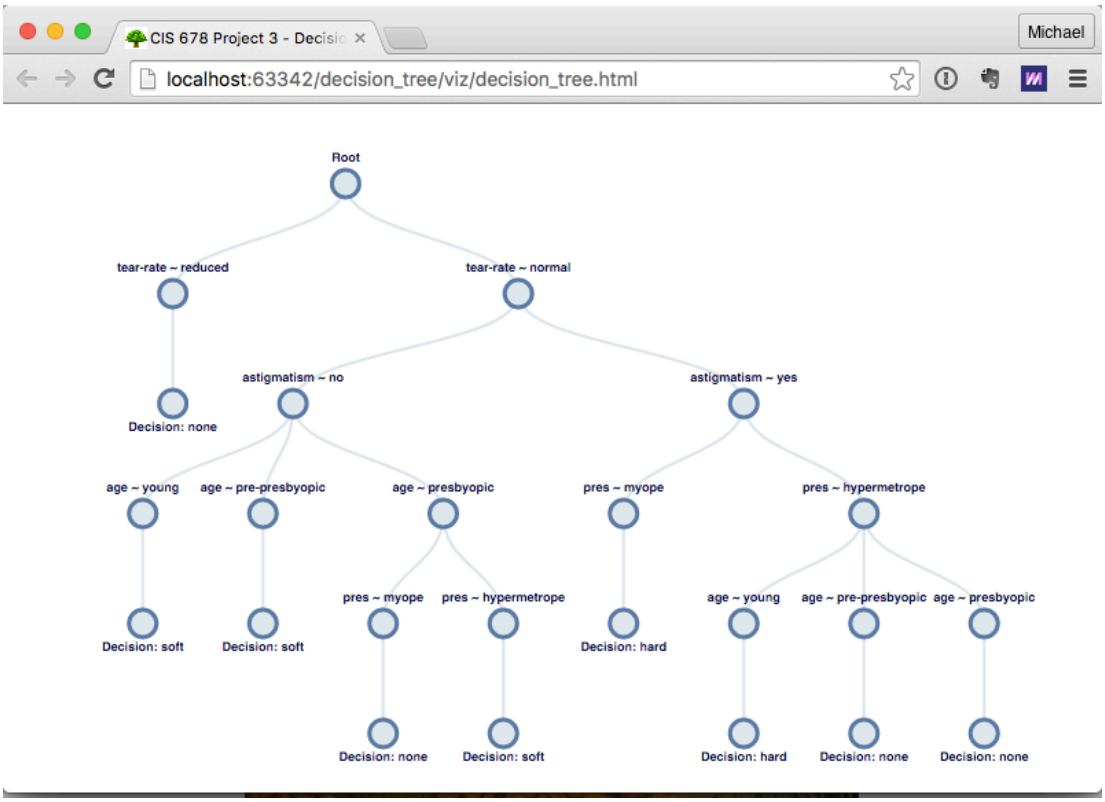


Figure 3. Decision Tree for Contact Lenses Data

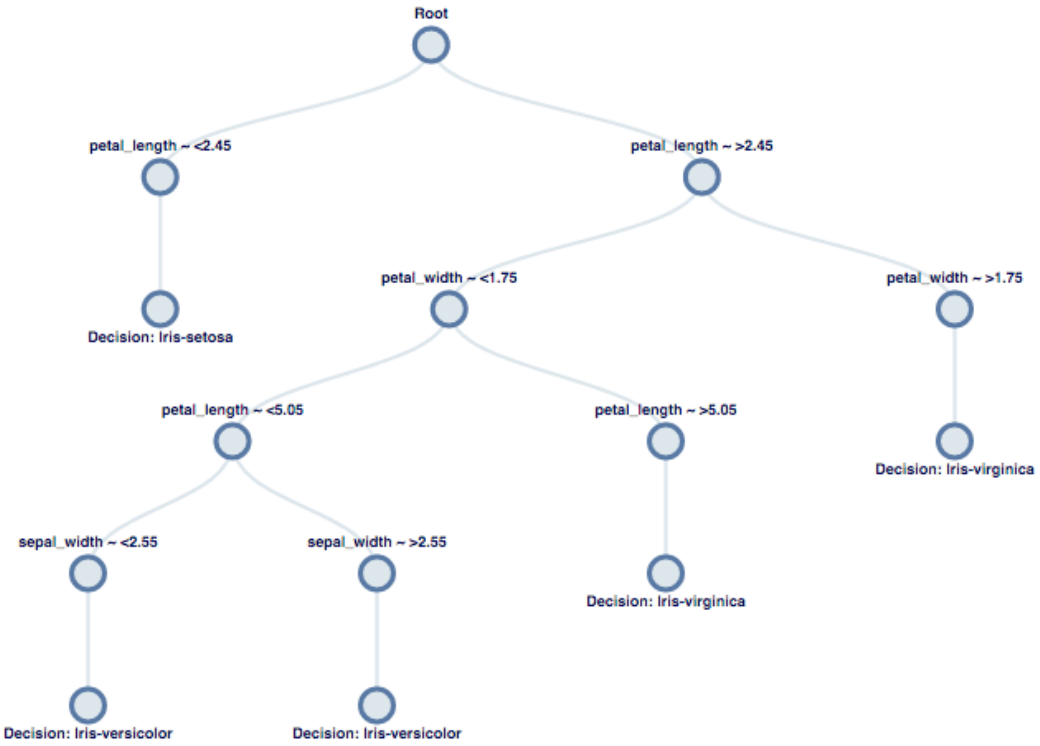


Figure 4. Decision Tree for Iris Data

We note in Figures 2-4 and Appendix 1; we have implemented four distinct supervised learning models (three categorical, one numeric) for decision trees. By using D3, we are able to host these visualizations through a web portal that allows for interactive exploration. Because the Mushrooms dataset was the only dataset with enough observations (>8000) to be split into a training set and test set, we can find results for classification in Figure 1. Encouraging points to make include a 97% classification rate with a 50/50 split, as well as, the depth of the Mushrooms tree is only 4/5 branches deep when starting with 22 possible attributes. We have demonstrated the benefits of decision trees with improve interpretability (few numbers of branches), scalability (runtime less than one second), and predictability (high classification rate).

Future Work

If time allowed, we would consider alternative splitting methods and pruning our trees. Additionally, we would look to a bigger training set with a combination of continuous and categorical variables to test our model. Our final suggestion is more of slight deviation in that we could consider variable selection or reduction techniques like that of principal component analysis to complement our decision tree models.

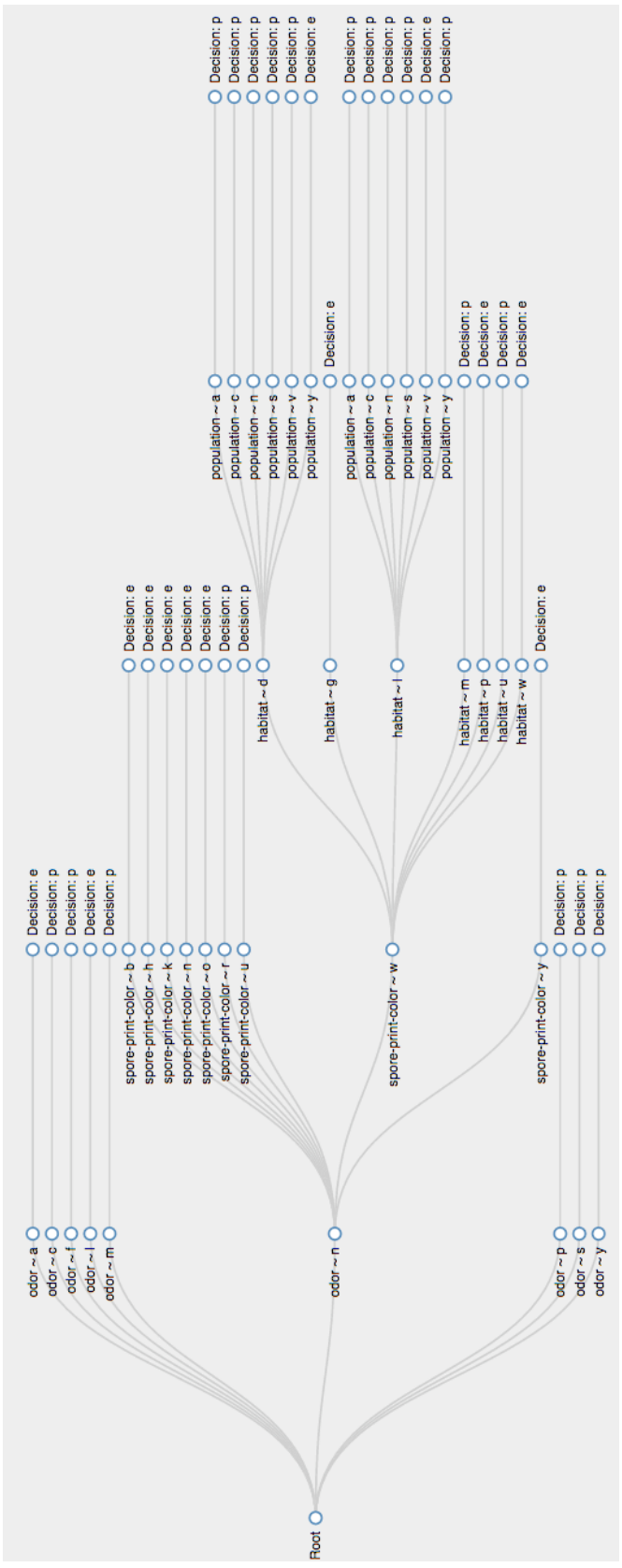
Credits

```
# Tree Visualization

- [JavaScript InfoVis
Toolkit](http://philogb.github.io/jit/index.html)
- [JSON Structure](http://stackoverflow.com/questions/14484613/load-
local-json-file-into-variable)
- [Export Python Dictionary to
JSON](http://stackoverflow.com/questions/12309269/how-do-i-write-json-
data-to-a-file-in-python)
- [D3 JS Tree Diagram](http://www.d3noob.org/2014/01/tree-diagrams-in-
d3js_11.html)
- [D3 Tree Layout](https://github.com/mbostock/d3/wiki/Tree-Layout)
- [JavaScript load local
file](http://stackoverflow.com/questions/16991341/json-parse-file-
path)

## Learning D3
- https://www.youtube.com/watch?v=x8dwXoWODZ4
- https://www.youtube.com/watch?v=OZXYk_bgQGQ

## Examples
- http://bl.ocks.org/d3noob/8375092
- http://bl.ocks.org/mbostock/4339083
- https://bl.ocks.org/mbostock/raw/4063550/flare.json
- https://bl.ocks.org/ajschumacher/65eda1df2b0dd2cf616f
```



Appendix 1. Decision Tree for Mushroom Data