Assignment # 4

T81-574: Foundations of Analytics

Washington University in St Louis — July 7, 2019

1 Mammal Classification Tree

You were given a data set "**zoo.csv**" that includes 101 animals and a list of characteristics of the animals e.g. do they have feather, do they lay eggs or not etc. Build a CART model to classify if an animal is mammal or not.

(1) calculate the entropy of of the target variable "ismammal", using the definition $H = p_1 log(p_1) + (1 - p_1) log(1 - p_1)$

(2) to build a classification tree, you need to decide the splitter for each node in a binary tree. Using the criterion that $hair > 0.5$ and split the dataset in to two branches. Calculate the entropy at each branch and the average entropy change.

(3) check the entropy changes for all the other features i.e. 'feathers', 'eggs', 'airborne', 'aquatic' and 'backbone'. Which one would you use to make the first split?

(4) build a CART model using the sklearn package and compare the model with your calculation. Is the first split the same as yours? You may use the python code provided in "CARTmammals.py"