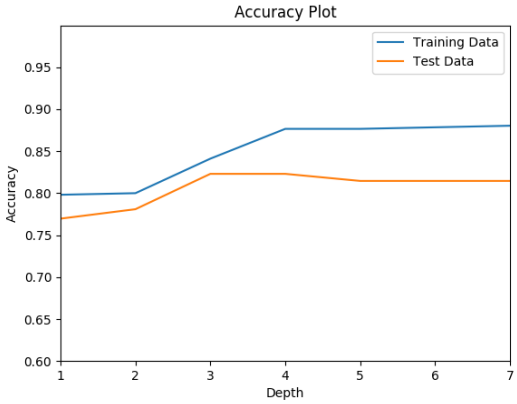
Siyuan Zou 111639762 CSE353: Homework 2 Algorithm Report



Discussion

Note: Feature used by the tree: Pclass, Sex, Age, SibSp, Parch, Fare, Embarked

Looking at the plot we can observed that around depth of 3 overfitting started to occur for the tree generated. The tree picks the attribute with the most gain to split on, thus the most discriminatory features is the first split at 'Sex'. This is the only variable that only had two categories which might have contributed to a strong information gain on the survival rate of female vs male. The least discriminatory features are located at the max depth of the tree, and the features are 'Sibsp' and 'Parch'. I believe that these two features are the least discriminatory simply because they aren't closely related to the rate of survival as much as the other feature was.

Not included features such as Name, ID, Cabin, and Ticket was removed, each due to some reasons. To start off Name and ID are two feature that can be consider as "Noise" as they provide us with no real information other than that of an identification. As ID are just 1-# of passenger, and name being name are very distinct and hard to be split upon. Cabin on the other hand could had the potential of being a strong feature, however lots of fields were empty thus we simply ignored. Ticket on the other hand was filled, but due to the formatting of some ticket exhibiting letters in them, we did not use this feature as well. Also, if we were to remove the letter of each ticket, and only consider the numbers I think that we would loose vital information and possibly overfit the data even more.

Aside from the excluded feature, one of the included shocked me at being not useful in term of providing information was the 'Embarked' feature. I thought that the location at which you might have enter the ship might have a similar relationship to that of 'Pclass' but that wasn't the case as Embarked was used in the tree at around depth 5-6.

README

python3 id3.py --dataset [/path/to/file] (default: ./data/titanic.csv)