Decision Tree

School of Data and Computer Science
Sun Yat-sen University

Overview

Introduction to ID3 Algorithm

- ID3 (Iterative Dichotomiser 3) was developed in 1986 by Ross Quinlan.
- The algorithm creates a multiway tree, finding for each node (i.e. in a greedy manner) the categorical feature that will yield the largest information gain for categorical targets.
- Trees are grown to their maximum size and then a pruning step is usually applied to improve the ability of the tree to generalise to unseen data.

Entropy

Entropy $H(S)$ is a measure of the amount of uncertainty in the set S.

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

where

- S is the current dataset for which entropy is being calculated
- X is the set of classes in S
- $p(x)$ is the proportion of the number of elements in class x to the number of elements in set S.

Information Gain

Information gain $IG(A)$ is the measure of the difference in entropy from before to after the set S is split on an attribute A. In other words, how much uncertainty in S was reduced after splitting set S on attribute A.

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

where

- $H(S)$ is the entropy of set S
- T is the subsets created from splitting set S by attribute A such that $S = \cup_{t \in T} t$
- $p(t)$ is the proportion of the number of elements in t to the number of elements in set S
- $H(t)$ is the entropy of subset t.

ID3 Algorithm Execution Process

The execution process of ID3 algorithm is as following:

1. Begins with the original set S as the root node.
2. Calculate the entropy of every attribute a of the data set S.
3. Partition the set S into subsets using the attribute for which the resulting entropy after splitting is minimized; or, equivalently, information gain is maximum.
4. Make a decision tree node containing that attribute.
5. Recur on subsets using remaining attributes.

Termination Conditions of ID3 Algorithm

Recursion on a subset may stop in one of these cases:

- Every element in the subset belongs to the same class; in which case the node is turned into a leaf node and labelled with the class of the examples.
- There are no more attributes to be selected, but the examples still do not belong to the same class. In this case, the node is made a leaf node and labelled with the most common class of the examples in the subset.
- There are no examples in the subset, which happens when no example in the parent set was found to match a specific value of the selected attribute.

Shortcomings of ID3 Algorithm

- ID3 does not guarantee an optimal solution.
- ID3 can overfit the training data.
- ID3 is harder to use on continuous data.

C4.5

- C4.5 is the successor to ID3.
- C4.5 removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals.
- C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules.
- These accuracy of each rule is then evaluated to determine the order in which they should be applied.

C4.5 (cont'd)

- Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it.
- C5.0 is Quinlan's latest version release under a proprietary license.
- C5.0 uses less memory and builds smaller rulesets than C4.5 while being more accurate.

CART

- CART (Classification and Regression Trees) is very similar to C4.5
- CART differs in that it supports numerical target variables (regression) and does not compute rule sets.
- CART constructs binary trees using the feature and threshold that yield the largest information gain at each node.

Train Classification Models in Classification Learner App

- You can use Classification Learner to train models of these classifiers: decision trees, discriminant analysis, support vector machines, logistic regression, nearest neighbors, naive Bayes, and ensemble classification.
- In addition to training models, you can explore your data, select features, specify validation schemes, and evaluate results.
- You can export a model to the workspace to use the model with new data or generate MATLAB® code to learn about programmatic classification.

Train Classification Models in Classification Learner App (cont'd)

Training a model in Classification Learner consists of two parts:

- Validated Model: Train a model with a validation scheme. The app protects against overfitting by applying cross-validation.
- Full Model: Train a model on full data without validation. The app trains this model simultaneously with the validated model.

Automated Classifier Training in MATLAB

You can use Classification Learner to automatically train a selection of different classification models on your data.

- Get started by automatically training multiple models at once.
- On the **Apps** tab, in the **Machine Learning** group, click **Classification Learner**.
- Click **New Session** and select data from the workspace or from file.
- On the **Classification Learner** tab, in the **Model Type** section, click **All Quick-To-Train**.
- Click **Train**.

Manual Classifier Training in MATLAB

If you want to explore individual model types, you can train classifiers one at a time, or a train a group of the same type.

- Choose a classifier.
  - On the **Classification Learner** tab, in the **Model Type** section, click a classifier type.
  - To see all available classifier options, click the arrow on the far right of the **Model Type** section to expand the list of classifiers.
  - The nonoptimizable model options in the **Model Type** gallery are preset starting points with different settings, suitable for a range of different classification problems.
- After selecting a classifier, click **Traini**.

Adult Data Set

- The UCI dataset (`http://archive.ics.uci.edu/ml/index.php`) is the most widely used dataset for machine learning.
- For more, you can refer to `https://www.zhihu.com/question/63383992/answer/222718972`.
- Today's experiment is conducted with the **Adult Data Set** which can be found in `http://archive.ics.uci.edu/ml/datasets/Adult`.

Adult Data Set (cont'd)

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	1305515

You can also find 3 related files in the current folder, `adult.name` is the description of **Adult Data Set**, `adult.data` is the training set, and `adult.test` is the testing set. There are 14 attributes in this dataset.

Tasks

- Given the training dataset `adult.data` and the testing dataset `adult.test`, please accomplish the prediction task to determine whether a person makes over 50K a year in `adult.test` by using ID3 (or C4.5, CART) algorithm (MATLAB), and compute the accuracy.
- Your codes should output the decision tree. The decision tree can be represented with a nested dictionary structure or be depicted by using the `plotTree` function in the book `Machine Learning in Action`.

Tasks (cont'd)

- Hints (You can refer to the book `Machine Learning` written by Zhou):
  1. You can process the continuous data with **bi-partition** method.
  2. You can use prepruning or postpruning to avoid the overfitting problem.
  3. You can assign probability weights to solve the missing attributes (data) problem.

Reference

- ID3, C4.5, and CART algorithm, Wikipedia.
- Zhihua Zhou, Machine Learning.
- http://archive.ics.uci.edu/ml/datasets/Adult
- http://archive.ics.uci.edu/ml/index.php
- https://www.zhihu.com/question/63383992/answer/222718972
- https://ww2.mathworks.cn/help/stats/train-classification-models-in-classification-learner-a.html
- https://ww2.mathworks.cn/help/stats/train-classification-models-in-classification-learner-a.html

The End