Decision Tree Topic Notes

A. What is a Decision Tree Algorithm?

(Wiki decision tree)

- **A machine learning algorithm used to create decision rules.**
- **Tree models represent training data by a set of binary decision rules.**

(CART (Classification and Regression Tree) Algorithms)

- **Origins in Statistics, Data Mining, Machine Learning**

[1] Breiman, Friedman, Olshen, and Stone (1984)
[2] Hastie etal (2001)
[3] Zhang etal (2001)
[4] Dudoit etal (2002)

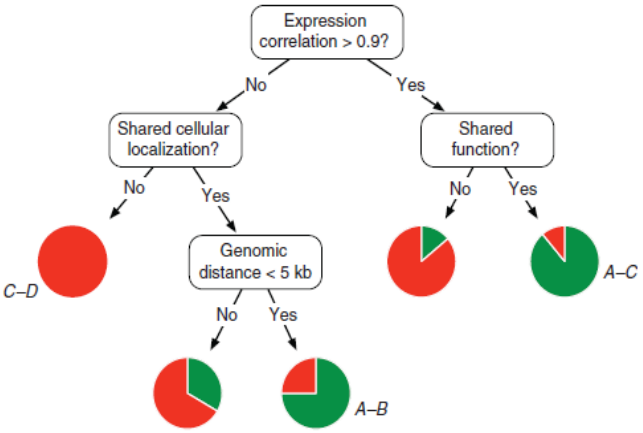
B. Decision Tree Terms

(Example Discussion) Hypothetical decision tree to study protein-protein interactions [2]

a

Gene Pair	Interact?	Expression correlation	Shared localization?	Shared function?	Genomic distance
A-B	Yes	0.77	Yes	No	1 kb
A-C	Yes	0.91	Yes	Yes	10 kb
C-D	No	0.1	No	No	1 Mb

b



- **Using training data features, the tree can be derived from binary recursive partitioning algorithms.**

- **After completing the tree, predicted values of sample data can be determined by tracing the rules of the tree model.**

- **New sample data for a gene pair are predicted to "interact" if they reach a predominately green path.**

Decision Tree Topic Notes

Part a) shows the training data for each data item.

- We are expressing whether the **response** **(Interacts)** in terms of **categorized predictor features**.
    - The **predictor** features can include both quantitative (real-valued) and categorical types.

      Real-valued features include**:** ***Expression Correlation*** *and* ***Genomic Distance***
      Categorical features include: ***Shared Localization*** and ***Shared Function***

    - The **response** value **(Interacts)** takes a categorical value (yes / no).

Part b) shows a hypothetical classification decision tree.

- Each **node** contains a yes/no question regarding the predictor feature(s) of the training data.

  Each **leaf (terminal node)** contains an outcome which is a categorical **yes/no value** and a distribution depicted by each **pie chart.**

  The pie chart distribution at each leaf node helps us to visualize the **majority rule** used to derive the outcome for each terminal node.

C. Tree Partitioning Concepts

The tree is derived by a binary recursive partitioning of the predictors to create subsets of data.
Ideally, this process stratifies the data.

Tree Construction

- Starting at the root node, CART evaluates all possible splits of all predictor variables and picks the best single split overall. The data are then partitioned according to the best split.[Berk]
- The same process is applied to all subsequent nodes until all cases can only be in one terminal node.[Berk]

Node Splitting Criteria

- Using a measure of impurity, nodes are split to minimize the weighted average of the impurity of the resulting child nodes in the tree.

  Measures used for splitting the nodes include:

  Regression Tree(s): **RSS (Residual Sum of Squares)**
  Classification Trees(s): **Deviance, Entropy, Gini Index**

Pruning

- A crucial aspect to building decision trees is limiting the complexity of the learned trees so that they do not over fit the training examples. [Kingsford]

Decision Tree Topic Notes

- One technique is to stop splitting when no question increases the purity of the subsets created.
- Alternatively, one can choose to build out the tree completely and then **prune** it back by deleting nodes.
- The pruning process collapses internal nodes into leaves to reduce the classification error on a held-out set of training examples.

Cross-Validation

- Cross-validation on an appropriate test set should be performed to assess the predictive ability of the model.

D. Summary of Decision Tree Algorithms

(Single Tree Analysis Techniques)

Algorithm creates a model that predicts the value of a response variable based on partitioning the attributes of its predictor variables.

$(X, \mathbf{Y}) = (X_1, X_2, X_3, \mathbf{Y})$

- If Y is real-valued, the tree model is called a **Regression Tree**.
- If Y is categorical, the tree model is called a **Classification Tree.**

(Multiple Tree Analysis Techniques)

Techniques for building predictive models based on other tree models include:

Bagging

- The idea of bagging is to use **bootstrapping** to build a number of different models and then average the results.

Boosting

- This technique is similar to "bagging". It combines multiple classifiers into a stronger classifier by repeatedly **reweighting** members of the training set.

Random Forests

- This technique builds trees from a random sample of the observations in the test data.

  The **training set is sampled with replacement** to produce a modified training set of equal size to the original but with some training items included more than once.

  In addition, only **a small random subset of the features** is considered during the tree building process.

Decision Tree Topic Notes

E. Applications to Computational Biology

- **Synthetic Sick and Lethal (SSL) Genetic Interaction Prediction**[Wong]
- **Computational Gene Finders**[Allen]
- **Gene Regulatory Response Prediction Using Classification**[Middendorf]
- **Cancer Risk Management Strategies**[Reynier]

F. Classification Tree Example

http://www.statmethods.net/advstats/cart.html

F. References

[1] Breiman, L., Friedman, J., Olshen, R. & Stone, C. Classification and Regression Trees (Wadsworth International Group, Belmont, CA, USA, 1984).

[2] Kingsford, C. & Salzberg, S., ***What are Decision Trees***, Nature Biology, Volume 26 Number 9 September, 2008. Pp. 1011-1013.

[3] Torgo, L., Data Mining with R: Learning with Case Studies (CRC Press), 2011.

[4] Faraway, J., Extending the Linear Model with R (Chapman & Hall CRC Press, 2006.

[5] Berk, R., Statistical Learning from a Regression Perspective, (Springer), 2008.