Predictive Modelling - IV

Tree Based Models

Rita P. Ribeiro

Data Mining I - 2023/2024

Summary

- Tree-based Models
  - CART: Classification and Regression Trees

Tree-based Models

Predictive Modelling: Where we at?

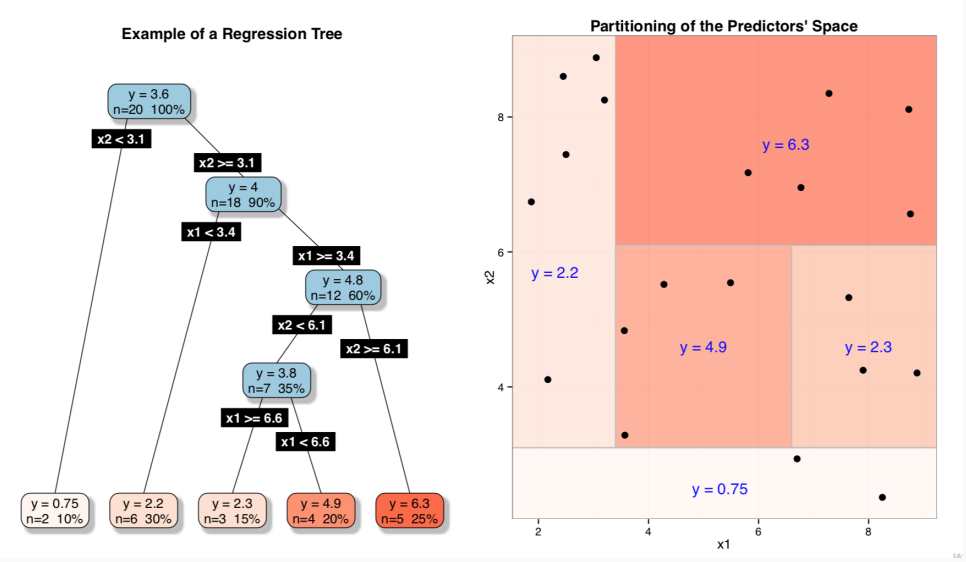
- Distance-based Approaches
  - e.g. kNN
- Probabilistic Approaches
  - e.g. Naive Bayes, Bayesian Networks
- Mathematical Formulae
  - e.g. multiple linear regression
- Logical Approaches
  - e.g. CART
- Optimization Approaches
  - e.g. SVM, ANN
- Ensemble Approaches

Tree-based Models

- Tree-based models are among the most popular models in machine learning, both for classification and regression.
- They have a recursive *divide-and-conquer* nature
- Each node in the tree is a partition of the input space and it represents either:
  - internal node with a test over the value of a predictor variable
  - leaf node that contains the value of the target variable (either class label or a numeric value)
- This partitioning is defined based on *carefully* chosen logical tests on these variables

Tree-based Models

- Each path from the root till a leaf node is a set of logical tests defining a region on the predictors space
- All the cases “falling” on a leaf will get the same prediction (either a class label or a numeric value)



The prediction for a new test case is obtained by following a path from the root till a leaf according to its predictor values.

x1	x2	y
7	8	?

Tree-based Models

Key Issues (with Recursive Partitioning)

- Which value to put on the leaves?
  - Should be the value that better represents the cases in the leaves
- How to find the best split test?
  - A test is good if it is able to split the cases of the sample in such a way that they form partitions that are “purer” than the parent node
- When to stop growing trees?
  - Too large trees tend to overfit the training data and will perform badly on new data - a question of reliability of error estimates

Tree-based Models

Some algorithms for Tree-based Models induction

- ID3 (Quinlan, 1979)
- CART (Breiman et al., 1984)
- C4.5 (Quinlan, 1993)
- etc.

Classification And Regression Trees (CART)

- Binary Recursive Partitioning trees with logical tests on each node
  - tests on numerical predictors: $x_i < \alpha, \alpha \in \Re$
  - tests on categorical predictors: $x_i \in \{v_1, \cdots, v_m\}$
- The main difference between Classification and Regression Trees lies on the used preference criterion
- This criterion has impact on:
  - The way the best test for each node is selected
  - The way the tree avoids over fitting the training sample
- Regression trees typically use the least squares error criterion
- Classification trees typically use criteria related to error rate (e.g. the Gini index, the Gain ratio, entropy, etc.)

CART Trees: Split Tests

Split Tests for Numeric Predictors

- Given a set of data D and a continuous variable A let $V_{A,D}$ be the set of values of A occurring in D
- Start by ordering the set $V_{A,D}$
- Evaluate all tests $A < x$ where x takes as values all mid-points between every successive value in the ordered set

CART Trees: Split Tests

Example for a Numeric Predictor

- Given the unsorted values of variable of temperature *Temp*

Temp	35	27	26	12	21	30	19	14	10

- Sort them

Temp	10	12	14	19	21	26	27	30	35

- Then try (i.e. evaluate) all tests in between each value:
  - $Temp < 11$, $Temp < 13$, $Temp < 16.5$, $Temp < 20$, etc.
- Choose the test in variable *Temp* with the best score.

CART Trees: Split Tests

Split Tests for Categorical Predictors

- Given a set of data D and a categorical variable *A* let $V_{A,D}$ be the set of values of A occurring in D
- Evaluate all possible combinations of subset of values in $V_{A,D}$
- There are some optimizations that reduce the computational complexity of this search

CART Trees: Split Tests

Example for a Categorical Predictor

- Given the values of variable *Weather*

Weather	*sunny*	*overcast*	*rainy*

- Try (i.e. evaluate) all subsets of these values:
  - $Weather \in \{sunny\}$
  - $Weather \in \{overcast\}$
  - $Weather \in \{rainy\}$
  - (is it necessary to test more?)
- Choose the test in variable *Weather* with the best score.

CART: Regression Trees

Which value to put on the leaves?

- The attribute chosen at each split is the one that minimizes some error estimate criterion.
- In Least Squares Regression Trees, the aim is to minimize

$$Err(t) = \frac{1}{n_t} \sum_{\langle x_i, y_i \rangle \in D_t} (y_i - k_t)^2$$

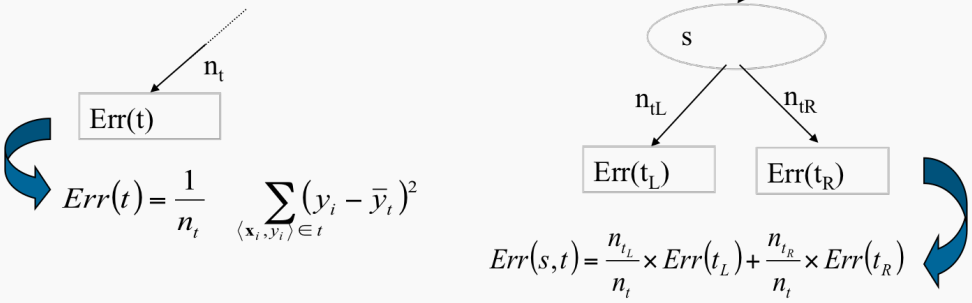
where D_t is the partition of cases in node t, n_t is the cardinality of this partition, and k_t is the constant used in the node

- the constant k that minimizes this error is the average target variable value of these cases, i.e. $\bar{y}_t$

CART: Regression Trees

How to find the best split test?

- Compare the errors in the partitions of unexpanded and expanded node *t*



- Choose the split that maximizes the reduction of error, i.e. $s^* = \text{argmax}_s \, Err(t) - Err(s,t)$

CART: Classification Trees

Which value to put on the leaves?

- The Gini index of a data set *D* where each example belongs to one of *c* classes is given by,

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$

where p_i is the probability of class *i* (e.g. estimated by the observed frequency on data)

- This index measures the impurity of a data set regarding the set of classes its examples belong to.
  - Higher the probability of a given class w.r.t to the others, more pure is the data partition.
- Each leaf node is assigned the majority class of cases that are in that partition.

CART: Classification Trees

How to find the best split test?

- Compare the impurity in the partitions of unexpanded and expanded node t

$$Gini(s,t) = \frac{n_{tL}}{n_t} \times Gini(t_L) + \frac{n_{tR}}{n_t} \times Gini(t_R)$$

- Choose the split that maximizes the reduction of impurity, i.e. $s^* = \text{argmax}_s\, Gini(t) - Gini(s,t)$

CART Split Tests Exercise

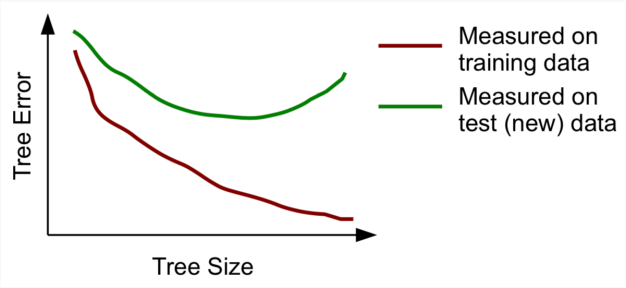
Consider the following data set for the play golf classification task.

Weather	Temperature	Humidity	Wind	Play
Rainy	71	91	Yes	No
Sunny	69	70	No	Yes
Sunny	80	90	Yes	No
Overcast	83	86	No	Yes
Rainy	70	96	No	Yes
Rainy	65	70	Yes	No
Overcast	64	65	Yes	Yes
Overcast	72	90	Yes	Yes
Sunny	75	70	Yes	Yes
Rainy	68	80	No	Yes
Overcast	81	75	No	Yes
Sunny	85	85	No	No
Sunny	72	95	No	No
Rainy	75	80	No	Yes

- Using the Gini index, what would be the best split test for:
  - *Weather* attribute?
  - *Temperature* attribute?
- Is it better to split by the *Weather* attribute or by *Wind* attribute?

CART: When to stop growing trees?

- Overall scores keep improving as we grow the tree.
- Still, as we go down in the tree the split decisions are made based on smaller and smaller sets.
- Thus, potentially less reliable decisions are made.



- It is necessary to find the “optimal” tree size, to avoid overfitting.

CART: When to stop growing trees?

Pre-prunning

- stop growing the tree if our estimate of quality indicates that is not worth continuing
- Examples:
  - minimum nr. of cases in a node;
  - minimum nr of cases in a leaf;
  - maximum depth of the tree.

CART: When to stop growing trees?

Post-prunning (most frequently used)

- grow an overly large tree and then use some statistical procedure to prune unreliable branches according to error estimates
- CART
  - grows an overly large tree
  - generates a sequence of sub-trees
  - uses cross-validation to estimate their error
    - error-complexity criterion for regression trees
    - cost-complexity criterion for classification trees
  - use the x-SE rule to select the "best" sub-tree

Tree-based Models: wrap-up

Pros

- Provide interpretable prediction models
- Handle both categorical and numerical predictor variables
- Don't require feature scaling
- Embedded variable selection and handling of missing values
- Computationally efficient

Cons

- Can be very sensitive to small variations in the data.
- Provide unstable models.

References

References

Aggarwal, Charu C. 2015. *Data Mining, the Texbook*. Springer.

Gama, João, André Carlos Ponce de Leon Ferreira de Carvalho, Katti Faceli, Ana Carolina Lorena, and Márcia Oliveira. 2015. *Extração de Conhecimento de Dados: Data Mining -3rd Edition*. Edições Sílabo.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Moreira, João, Andre Carvalho, and Tomás Horvath. 2018. *Data Analytics: A General Introduction*. Wiley.

Tan, Pang-Ning, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. 2018. *Introduction to Data Mining*. 2nd ed. Pearson.

Torgo, Luís. 2017. “Data Mining i Course.” Slides.