Data Science | Lancaster University

Decision Trees

An introduction to regression and classification trees

July 5, 2018

Aaron Lowther

Resources

- Most notes are taken from Hastie et al. (2009)

https://web.stanford.edu/~hastie/ElemStatLearn/

Tree based methods

Data Science | Lancaster University

Overview...

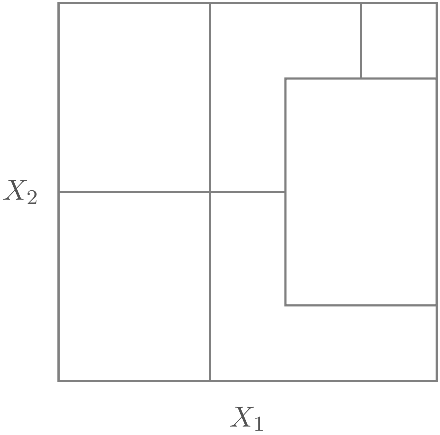
- Two specific types of *decision trees*:
    - Regression Trees: For a continuous response,
    - Classification Trees: Discrete response.
- Iteratively partition the covariate space into a series of (hyper) rectangles.
- Use a simple function in each rectangle to predict a response.
- We are going to assume we have data,

$$\{(y_t, x_{t,1}, \ldots, x_{t,P}) \text{ for } t = 1, \ldots, T\},$$

for $y_t \in \mathbb{R}$ (for now).

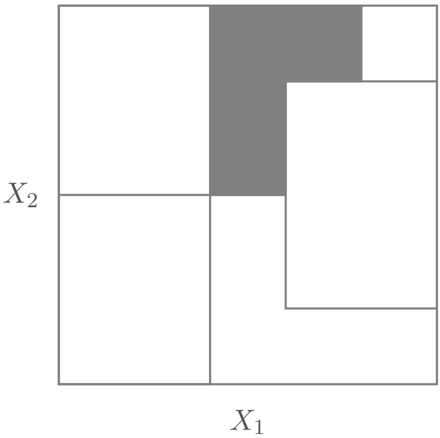
Partitioning

The covariate space



Partitioning

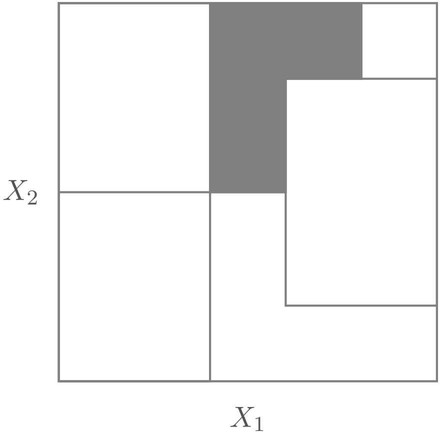
The covariate space



But not like this!

Partitioning

The covariate space



But not like this!

We have a grey area, it's hard to explain...

Partitioning

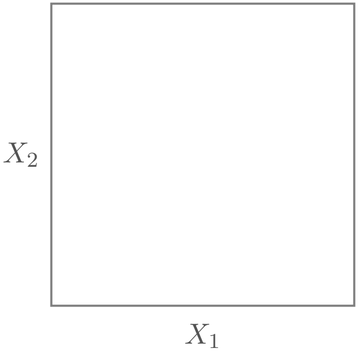
Recursive binary partitioning

Data Science | Lancaster University

- Instead of arbitrary partitioning
- We stick to recursive binary partiting
- The advantages are:
  - The partitions are easy to explain
  - There is a simple tree representation of the partitions

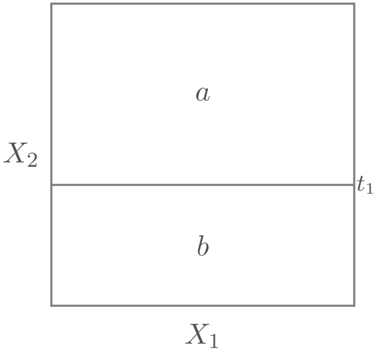
Partitioning

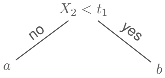
Recursive binary partitioning: An example



Partitioning

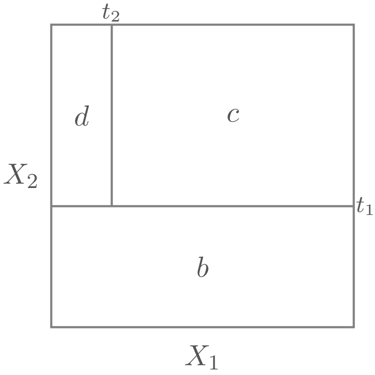
Recursive binary partitioning: An example

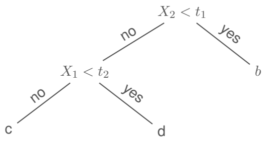




Partitioning

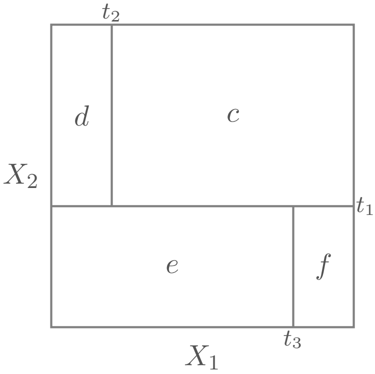
Recursive binary partitioning: An example

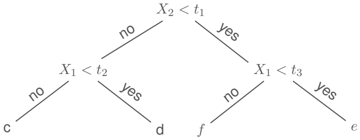




Partitioning

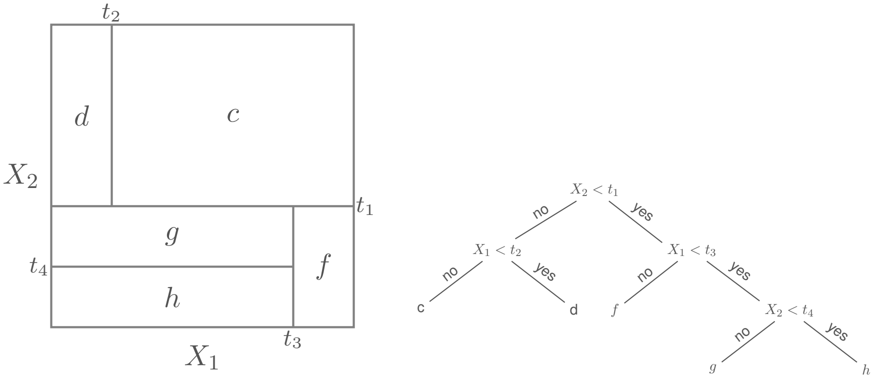
Recursive binary partitioning: An example





Partitioning

Recursive binary partitioning: An example



Making the tree

Data Science | Lancaster University

- We need to determine (automatically):
  - Which covariates to split
  - Where to split them

- With the aim of developing a model of the form,

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \mathbb{I}(\boldsymbol{x} \in \mathcal{R}_m),$$

  - M: Number of regions,
  - $\mathcal{R}_m$: Region m,
  - c_m: Predicted value of the reponse for $\boldsymbol{x} \in \mathcal{R}_m$
  - $\mathcal{R} = \{\mathcal{R}_1, \ldots, \mathcal{R}_M\}$

Making the tree

Determining the values of c_m

- In the linear regression setting we use the least squares criterion,

$$SS(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2$$

  to determine the values of $\boldsymbol{\beta}$ for the model,

$$y_t = \sum_{p=1}^{P} x_{t,p}\beta_p + e_t.$$

- We can adopt the same *least squares* criteria here, minimizing

$$\mathcal{RT}(\boldsymbol{c}; \mathcal{R}) = \sum_{t=1}^{T} (y_t - f(\boldsymbol{c}, \boldsymbol{x}_t))^2$$

Making the tree

Determining the values of c_m

- We can adopt the same criteria here, minimizing,

$$\mathcal{RT}(\boldsymbol{c};\mathcal{R}) = \sum_{t=1}^{T} (y_t - f(\boldsymbol{c}, \boldsymbol{x}_t))^2 \, .$$

- Then the values of $\boldsymbol{c}$ that minimise the function,

$$c_m = \frac{1}{N_m} \sum_{t=1}^{T} y_t \mathbb{I}_{\boldsymbol{x}_t \in \mathcal{R}_m} ,$$

  are just the mean of the response values in the region (given regions $\mathcal{R}_m$).

Making the tree

Determining the regions

- Finding $(\boldsymbol{c}, \mathcal{R})$ that minimises,

$$\mathcal{RM}(\boldsymbol{c}, \mathcal{R}),$$

  in generel, is very hard.

- Early splits will effect subsequent splits...

- But a greedy approach helps simplify the matter

Making the tree

Determining the regions

- To start the search we search each split value for each covariate, finding (p, s) such that,

$$\min_{p,s} \left[\min_{c_1} \sum_{x_i \in \mathcal{R}_1(p,s)} (y_t - c_1)^2 + \min_{c_2} \sum_{x_i \in \mathcal{R}_2(p,s)} (y_t - c_2)^2 \right],$$

  is obtained.

- $\mathcal{R}_1(p, s) = \{\boldsymbol{x} | x_p \leq s\}$, $\mathcal{R}_2(p, s) = \{\boldsymbol{x} | x_p > s\}$

- Apply this splitting approach to subsequent regions.

Summary so far

So now we have a method to build a tree,

- Finding the order to split the covariates,
- Finding the split values for the covariates.

But when do we stop?

When do we stop?

- We could continue growing the tree providing the reduction in sums of squares is greater than some threshold.
- The preffered method is to grow a large tree, and then prune it.
- Cost complexity pruning is used

Cost-complexity pruning

Let,

- T_0: The largest tree obtained,
- $|T|$: The number of terminal nodes in T,
- N_m: Number of $\boldsymbol{x}_t \in \mathcal{R}_m$,
- $Q_m(T) = \frac{1}{N_m} \sum_{\boldsymbol{x}_t \in \mathcal{R}_m} (y_t - \hat{c}_m)^2$

For each α we find the tree that minimises the cost complexity criteria $C_\alpha(T)$,

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T|,$$

by collapsing the tree by its weakest (internal) node.

Discrete response variables

When the data is discrete, taking values $1, 2, \ldots, K$ we need to adapt the:

- Slitting criteria.
- Pruning criteria.

For regression we used the *squared error node impurity measure*,

$$Q_m(T) = \frac{1}{N_m} \sum_{\boldsymbol{x}_t \in \mathcal{R}_m} (y_t - \hat{c}_m).$$

Classification

For categorical predictors, three available node impurity measures are:

- Misclassification error,

$$\frac{1}{N_m}\sum_{t\in\mathcal{R}_m}\mathcal{I}((y_t \neq k(m))) = 1-\hat{p}_{mk(m)}.$$

- Gini Index,

$$\sum_{k\neq k'}\hat{p}_{mk}\hat{mk'} = \sum_{k=1}^{K}\hat{p}_{mk}(1-\hat{p}_{mk}).$$

- Cross-entroy/deviance,

$$-\sum_{k=1}^{K}\hat{p}_{mk}\log\hat{p}_{mk}.$$

Some notes

Data Science | Lancaster University

- Cross-entropy and Gini index are more sentaive to changes in the node probabilities than the misclassification rate.
- The cross-entropy or Gini index should be used when growing the tree.
- Any of the three methods can be used for cost complexity pruning.

Other issues

The textbook covers other issues and extensions that include,

- Categorical Predictors.
- Loss matrix, when misclassifying some observations is more serious in a given class.
- Missing predictor values: A number of approaches, but could use a missing category.
- Linear combinations of splits.

Bibliography

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition.

July 5, 2018
Aaron Lowther