Bits and bobs

Myra O'Regan

October 31, 2017

Growing trees using cost information

- We can decide to use cost information to grow trees as well
- c categories
- Gini without cost information

$$g(t) = \sum_{i=1}^{c} \sum_{j=1, i \neq j}^{c} p(j|t)p(i|t)$$

Growing trees using cost information

- Gini with costs

$$\sum_{i=1}^{c} \sum_{j=1, i \neq j}^{c} C(j|i)p(j|t)p(i|t)$$

- where $C(j|i)$ cost of misclassifying j as an i
- When c= 2 this reduces to

$$(C(2|1) + C(1|2))p(j|t)p(i|t)$$

- So what ??

Costs and Pruning

- Without costs r(t) $= 1 - \underset{j}{max} p(j|t)$- node level
- Remember r(t) is the misclassification rate of a node
- With costs r(t) $= \underset{i}{min} \sum_j C(j|i)p(j|t)$
- For 2 classes and assuming $C(1|1) = C(2|2) = 0$
- i $= 1$: $C(2|1)p(2|1)$ - Assign all cases to 1
- i $= 2$: $C(1|2)p(1|2)$ - Assign all cases to 2
- Choose minimum
- R(T) $= \sum_{|T|} r(t)p(t)$ tree level
- Costs may alter pruning pruning regime

Elder's Shuffling technique

1. Build a model to predict the target variable
2. Record some measure of goodness of fit e.g. Accuracy
3. Randomly shuffle the target vector to break the relationship between each target value and its vector of inputs
4. Search for a new best model and record measure of goodness of fit.
5. Repeat steps 3 and 4 a number of times and create a distribution of the measure of goodness of fit
6. Evaluate where your true result for Step 1 lie.
7. Like a p-value probability that a result as strong as this can occur by chance