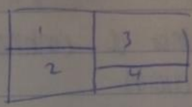
20/04/16 DA - RuleFit

- Developed by Jerry Friedman
- Combines trees and regression
- Used for descriptive or predictive purposes

- Turn a tree into a series of rules i.e. if (age < 70): left else right
- A series of indicator variables
- ~~Crea~~ Call them $r_m(x)$
- Each terminal node defines a region
- Regression of terminal nodes



$$F(x) = \underbrace{c_0 + \sum_{m=1}}_{\text{no upper limit} \rightarrow \text{choose}} \underbrace{c_m r_m(x)}_{\text{value of } r_m\text{'s}} + \underbrace{\sum_j b_j x_j}_{\text{add to capture linear effect}}$$ — ordinary variables in dataset (?) linear continuous

- $r_m(x)$ output from trees - indicator variables
- Need to determine c_i and b_j's
- Builds many trees of a certain depth
- Use these to create new indicator variables
- Uses intermediate nodes also

- P_m are the split definitions for rule $r_m(x)$
- Can include rules for all non terminal nodes also
- Counting all nodes, a J-terminal node tree generates $2*(J-1)$ rules

Combine the linear part and rules part
- Weight each of the terms
- Use some penalty function to reduce number of terms
- Lasso function
- Can use general loss function
- Post processing phase
- Could add non linear terms like x^2 or x^3

- Can be a large number of terms in rulefit function
- Many types of penalty can be used
  - Sums of absolute value of coefficients – lasso
  - Sums of squares of coefficients – ridge regression
  - Or both – elastic net
  - Big area

- $p \leq n$ number of continuous variables to be included as linear terms
- Because of outliers
- M is the number of trees
- $K = \sum_{m=1}^{M} 2 \times (J_m - 1)$ total number of rules
- J_m is number of terminal nodes for $tree_m$

Tree Size
- Important to consider.
- 2 terminal nodes only consider one factor → no interactions.
- To capture interaction need larger size trees
- For 2-way interaction → need 4 terminal nodes
- Use trees of varying size

- Number of terminal nodes is a random variable
- $t_m = 2 + fl(\gamma)$
- Where γ is drawn from an exponential dist with $Pr(\gamma) = \frac{\exp(-\gamma)(\bar{L}-2)}{(\bar{L}-2)}$
- $fl(\gamma)$ is the largest integer less than or equal to γ
- $(\bar{L} - 2)$ is the average number of terminal nodes for trees in ensemble $(\bar{L} \geq 2)$

- $\bar{L} = 2$ the entire ensemble will have just 2 terminal nodes
- For $\bar{L} = 3$ $t_m = 2 + fl(\gamma)$
- Where γ drawn from exponential distribution with probability
  $Pr(\gamma) = e^{((-\gamma)(\bar{L}-2))} / (\bar{L}-2) = e^{-\gamma}$
- Exponential Distribution with mean of 1

Linear Terms
- Windorised the linear terms
- Replace the top and lower percentiles with the next higher or lower values
- Typical value for percentile is 2.5%
- Can remove outliers
  i.e assign the top 2.5% of observed values all to be equal to the 97.5th percentile value (similar for the lower 2.5%)

Spurious Interaction
- Outliers may cause problems
- Many spurious interactions may occur especially if variables are highly correlated
- Place an incentive for fewer variables entering path
- Choose split with maximum improvement z_j splitting on x_j
- Adjust this to $k_j z_j$ where $k_j = 1$ if x_j had not been used in this branch before
- Otherwise $k_j = k (>1)$
- Discourages highly correlated variables from appearing in the same rule

Spurious Relationship - math relationship in which two events or variables have no direct causal connection, yet it may be wrongly identified that they do

When A is present, B is observed (A causes B)
When B is present, A is observed (B causes A)
OR
When C is present, Both A and B are observed (C causes both A and B)

In the last case there is a spurious relationship between A and B. In a regression model where A is regressed on B but C is actually the true causal factor of A.

Importance

- Importance of any predictor in a linear model is the absolute value of corresponding standardised predictor.
- For rules this is $RI_k = |\hat{a}_k| \cdot \sqrt{s_k(1-s_k)}$
- Where $s_k = \frac{1}{N}\sum_{i=1}^{N} r_k(x_i)$ - proportion of 1s if we have 0,1 values → the proportion
- For linear items this $I_j = |\hat{b}_j| \cdot std(l_j(x_j))$
- Where $std(l_j(x_j))$ is the standard deviation of $(l_j(x_j))$ over data
- Relative importance here - nearly controlling for other variables in $\hat{F}$

Local Measure of Importance

- Local measure of importance for each point x as the absolute change in predictor when term is removed from the ensemble
- For the tree part $RI_k(x) = |\hat{a}_k| \times |(r_k(x) - s_k)|$
- For the linear predictors $I_j = |\hat{b}_j| \cdot |l_j(x_j) - \bar{l}_j|$
- Where $\bar{l}_j$ is the mean of $l_j(x_j)$ over training set
- s_k - "the support of that rule"

- The two can be combined:
  $J_l(x) = I_l(x) + \sum_{x_l \in R_k} RI_k(x)/m_k$
- $I_l(x)$ is the importance of the linear predictor
- Second term sums over the importance of the rules divided by the total number of input variables in the rule
- Can also be averaged over any subset of the input space

Interactions

- Looks for variables which are involved in an interaction
- Compares this to what you would expect for no interactions present.
- Reference distribution is ~~computed~~ computed using a bootstrap method
- For each variable identified, look to see which variable or variables interacts with
- Can look for 2-way or 3-way interaction - depends on size of tree - if you only have a stump will get no interaction

20/04/16 DA - RULEFIT

- Charts - Show whether they interact or not.
- Red is the interaction - What you would expect to see
- Don't add up to 1 → don't have to
- Plots a single variable and shows interaction compared to other variables
- Haven't got this information from ensembles or trees.
  ↳ In ensembles, you would have to know there was an interaction

Partial Dependencies

- Can look at how response function changes with regard to important variables
- Can look at two variables at a time to understand interactions

- Plots X_1 against X_5 for when X_1 is $=-1$ and $X_1=1$
Shows $-1, 0, 1$ on x-axis for $X5$.
When there is zero height, this highlights interaction effect.

- When $X_1=1$, X_5 doesn't enter equation → no effect → has no bearing on y

- If pattern is roughly the same between charts → no interaction

Rules

Support: proportion of cases in this when $X_1=1$ and $X_2=1$ (vote 1)
Coef: coefficient in your equation: β or $\hat{a}$
X_1 not in level 1 → so no -1 in the model here
- Be careful with factors
- Categories are numbered 1,2,3
- Factor coded as $-1, 0, 1$ rules will contain reference to 1, 2 and 3

- Use defaults as much as possible
- Maybe change tree size
- Make sure results make sense
- Be careful in comparing importance measures to random forest

- ensemble method which combines the predictions of a large number of simple models
- Predictive power of these rules is combined via regularised regression
- To derive rules - large number of CART trees grown
- Along with these rules, each predictor variable is also added to ensemble - to account for any linear dimensions within the data that trees are poor at approximating

- Cross validation error - indicates how well a model will generalise to an independent dataset
  RFE for uses a k-fold cross val tech where original data is split into k subsets
  - A single subset is kept to validate a model built from remaining k-1 subsets
  - Repeated with each subset taking a turn at validation, so that the final model may be a combination of the k-models.
  - By taking an expected error value from the models, one can estimate how well a model will generalise to an independent data set.