Overview of Ensembles

Myra O'Regan

October 31, 2017

Ensembles

$$F(x) = c_0 + \sum_{m=1}^{M} c_m T_m(x) \tag{1}$$

$T_m(x)$ are called basis functions and can be anything

- Neural nets
- Logistic regressions
- Trees
- Discriminant analysis
- or a mixture of all the above

What does theory tell us

- Theory suggests that we should create the $T_m(x)'s$ to be as different as possible
- Correlation between the predicted values should be as low as possible
- How do we this?

Focus on Trees

- What can we change from tree to tree?
- What aspects can we play with?
- Raw data
- Size of tree
- Number of trees
- Variables used to create splits
- Dependency on previous trees?
- How to combine results?

Raw Data

- Sample data
- Bootstrap approach
- Usually uses the same size dataset
- But we could sample a certain %
- What should that % be?
- Give more weight to certain points
- Input data could be output from previous tree

Consequences of taking a sample

- Much quicker
- Subsample not used for every tree
- Can be used as a test set instead of crossvalidation approach
- Maybe good for small datasets

Dependency on previous tree

- Suppose we had a Y variable and a number of $X_i's$
- We find the best X_i
- Compute residuals $R_i =$Y - $b_i * X_i$
- Then we use $R_i's$ as the Y to look for the next variable
- Implicit is the concept of a loss function in determining the b's
- We aim to minimise the $(Y_i - b_i * X_i)^2$
- We could have chosen another one e.g. minimise absolute values

Size of Tree

- Grow a very big tree
- Grow a very small tree
- Have to think about interactions
- To capture 2 way interactions need a depth of 3
- For 3 way need a depth of 4
- You may overfit data with very large trees

How do we combine results

$$F(x) = c_0 + \sum_{m=1}^{M} c_m T_m(x) \tag{2}$$

- Weight each tree equally
- Majority voting
- Weighted version with weights determined simply
- Calculate the c_i as we go along
- Calculate the $c_i's$ at the end
- Use stacking - a different approach

How about doing something with splits?

- Use all variables
- Use a subset of variables
- Use Random splits

So what now?

- Bagging
- Boosting
- Gradient Boosting
- Random Forest
- RuleFit

What else should we be thinking about?

- Formula for prediction for each type of Ensemble
- What other type of output would you like
- How good a model it is
    - A nice picture
    - Information about the structure of the model
    - Variable importance
    - Presence of Interactions
    - Relationship of variable to target variable

One final word

- All the techniques will involve setting some parameters
- We may have to conduct a grid search to find the best value for each parameter