Trees and Boosting

1 Decision Trees

Pros: Do not require centring or rescaling. Non-parametric.

Cons: Overfit unless regularised with parameters, strong dependence on train set.

Scaling: $O(n * m * \log_2(m))$ with m-training samples and n-features

Parameters: Gini Purity: computed at each i-th node as:

$$G_i = 1 - \sum_{k=1}^{n} \left(\frac{\text{(Class k samples on node)}}{\text{(Class k samples from train)}} \right)^2$$

Node Split: Using CART cost function, maximises purity G_i for the two following subsets, with a weight on their size.

Extras: Using Entropy instead of Gini gives more balanced trees.

Regression Tree: Minimises on the Mean Squared Error not Gini purity.

Random Forest: Use Decision Tree with Bagging, sampling randomly over the set and averaging the votes. More random and diverse trees provide higher bias but lower variance.

Extra-Trees: As Random Forest but using random threshold for each feature at each node (no need of finding the best one). Faster, lower variance, higher bias.

Feature Importance: The most important features are near to the root of the tree. Fast estimator for feature importance.

2 Boosting

Adaptive Boosting: Train each classifier and weights the worst answers for the next tree. Not possible to run in parallel. Prediction is chosen on majority voting.

Parameters: Learning rate parameter decides how strong should the weights be. Low LRP will slow down learning and be more conservative.

Gradient Boosting: Train each classifier on the residual of the previous one. Derive final classifier by summing all the steps.

Parameters: The learning rate scales the contribution of each tree. If you set it to a low value, you will need more trees in the ensemble to fit the train set, but the predictions will generalise better.

Early stopping: Find the number of trees that give minimal validation error. Too many trees will overfit, too little will under-fit.

Stacking: Finally we can use multiple classifiers to output predictions on a CV set and use another classifier to train over the target values using the predictions as they would be features