YOUR QUESTIONS ANSWERED

Q: What is the difference between a deep forest and a random forest, just with more trees? What is the advantage of grouping them into multiple forests?

A: Deep forest, or gcForest, was introduced in the 2017 paper *Deep Forest: Towards an Alternative to Deep Neural Networks* by Zhi-Hua Zhou and Ji Feng. Current deep learning models are mostly built upon neural networks, that is, multiple layers of parameterized differentiable nonlinear modules that can be trained by back propagation. The deep forest algorithm builds deep models based on non-differentiable modules. Success of deep neural networks depends on three characteristics – layer-by-layer processing, in-model feature transformation, and enough model complexity.

Deep forest holds similar characteristics. This is a decision tree ensemble approach, with fewer hyper-parameters than deep neural networks, and its model complexity can be automatically determined in a data-dependent way. Inspired by the deep neural networks, deep forest has multilayer cascade structure, but each layer contains many random forests instead of neurons in deep neural networks. Deep forests use traditional forests as a subroutine, adding more depth and flexibility to cater to image and text data.

Q: How does mean/median imputation affect the validity of the tree?

A: Decision trees accommodate missing values very well compared to other modeling methods. Decision trees that split on one input at a time are more tolerant to missing data than models such as regression that combine several inputs.

The main advantage is that the tree model knows that the variable is missing, which is not the case for mean or median imputation. Imputing a large number for numeric data could be concerning for tree-based models. For example, if you split on age and the split is at 70 years, now everyone that was missing is going to be in the split with the older age group. The model will be fitted with those imputed values as well. So if they are significantly different than truly older people, the tree should make a split with true older age group and fake older age group (missing). If variability is low inside the tree node, then there is not much to worry about.

For the simplest of tree algorithms, the only observations that need to be excluded are those missing the input currently being considered to split on. They can be included when considering splitting on a different input (for example, tree algorithms that treat missing observations as a special value will use all

the observations). Trees, therefore, might be the best modeling tool for imputing missing values because of their tolerance to missing data, their acceptance of different data types, and their robustness to assumptions about the input distributions.

Q: Where can I buy Dr. Saxena's book?

A: It is available here.

Q: Where can I learn more on data manipulation/data management? I'm also interested in machine learning models.

A: Please refer to our learning path: support.sas.com/training/us/paths/.

Q: Could you suggest how to identify key interactive variables from GBM or XGB models, and possibly use the interactive variables in GLM?

A: Like any other machine learning models, correlated variables might be identified before applying machine learning models including GBM. However, correlated inputs don't pose much of a concern for GBMs.

Q: Are their problems where decision trees are a bad choice?

A: Oh yes, as with any model – if the model is a poor representation of the underlying process that generates the data, the prediction might not be useful. With trees, since they are able to approximate so many shapes, the most likely scenario for this is biased sampling, or overfitted models.

Q: Are decision trees more for exploration and not for casual inference?

A: Trees are more useful for prediction and interpretation. There is no overall hypothesis test for the tree model.

Q: How is the cutoff point for the split determined in the case of a continuous input variable?

A: The exact approach is to enumerate all possible groupings of X values and for each combination, and the splitting gain is computed. The best combination wins. An exhaustive search algorithm considers all possible partitions of all inputs at every node in the tree. When the distinct values of an interval input grow rapidly, the possible combinations are enormous. Decision trees in SAS® Viya® use an intricate strategy of split search based on how many branches are required and how many distinct values of variables.

Q: In a regression tree, can the model only predict a granular value based on the number of leaves? It can't predict values over the entire interval?

A: Yes. Regression tree models fit response surfaces that are constant over rectangular regions of the

predictor space, so they often lack the flexibility needed to capture smooth relationships between the predictor variables and the response.

Q: Is decision tree modeling a supervised learning process or an unsupervised learning process?

A: Decision trees are supervised – you have to have a target variable. However, an interesting application of trees can be to profile clusters after they've been developed.