Machine Learning Week 3

Nikolai Klassen

19 Oktober 2017

Trees, random Forests, Bogging...

If data is splitble into various homogenous groups, the data is very easy to interpretate and causal interference are made easier than in complex data frames.

Example: Obama Clinton Divide. Homogenous groups are taken into account and their support realtive to the other homogenous groups.

How to basic algorithm:

1. Start with all variables
2. Find the variable that best separets the outcome.
3. Divide the data into two groups: "leaves" on the node.
4. Within each split, find the best variable that separets the outcome
5. Continue until the groups are too small or sufficiently "pure"

Measures of impurity:

a) Misclassification Error:

$p_{m,k} = \frac{1}{N_m} \sum_{inLeaf m} 1(y_i = k)$

with 0 = perfect purity and 0.5 = no purity.

b) Gini Index - not to be confused with the Gini Coefficient:

$\sum_{i=1}^{n} X_i$

c) Deviance/Information Gain:

$\sum_{i=1}^{n} X_i$