Machine Learning

Random Forests

Jeff Abrahamson

9, 16 décembre 2016   13 janvier, 9 mars 2017

Decision Trees

Decision Trees

is sex male?
yes / no

is age > 9.5?   survived
                       0.73 36%

died   is sibsp > 2.5?
0.17 61%
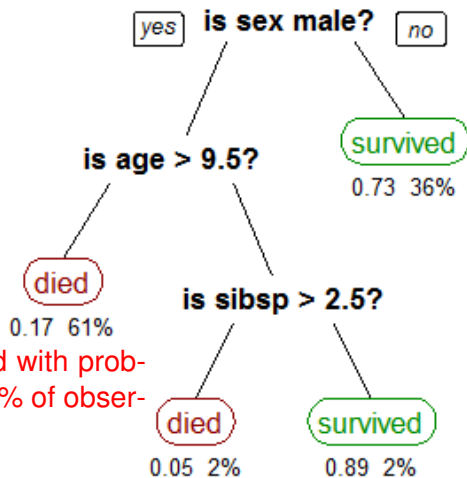
E.g., passengers died with probability .17 which is 61% of observations

died   survived
0.05 2%   0.89 2%

Stephen Milborrow

Decision Trees

Variations

- Classification tree
- Regression tree

What can go wrong?

Decision Trees

Ensemble methods

- Bagging
- Random forest
- Boosted trees *(gradient boosted trees)*
- Rotation forest

Boostrap aggregating = bagging

Boostrap

A family of statistical methods using sampling with replacement.

Boostrap aggregating = bagging

- Increase stability
- Increase accuracy
- Reduce variance
- Avoid overfitting

A type of model averaging.

Boostrap aggregating = bagging

- Training set D of size n
- Sample D *with replacement* to create $D_1, \ldots, D_k$ of size n'
- If $n = n'$, expect $1 - 1/e \approx 63.2\%$ repeats

Boostrap aggregating = bagging

- Training set D of size n
- Sample D *with replacement* to create $D_1, \ldots, D_k$ of size n'
- If $n = n'$, expect $1 - 1/e \approx 63.2\%$ repeats

- Train k models
- Average (regression) or vote (classification)

Boostrap aggregating = bagging

Do not confuse with

- Boosting (and AdaBoost)
- Boostrap (statistics)
- Cross validation

Random subspace method

attribute bagging = feature bagging

Random subspace method

Bagging (bootstrap aggregation) = resampling to create more data sets, train models on different samples

Attribute bagging = project to create more data sets, train models on different samples

Random forests

Combine bagging with random subspace method

questions?