Ensemble Methods

Avinash TAMBY

What Are Ensemble Methods?

- Combine a set of learners that are (or may be) weak individually to create a stronger overall model
- 2 families
  - **Bagging**: Build models on different bootstrapped subsets and average them for a final prediction
  - **Boosting**: Build a model, see what it got wrong, increase weight of incorrect classifications, repeat. Allows subsequent models to focus on more difficult cases

Overview

- Averaging Methods
  - Bagging (Bootstrap Aggregating)
  - Random Forests
  - Extra Trees (Extremely Randomized Trees)

- Boosting Methods
  - AdaBoost (Adaptive Boosting)
  - Gradient Boosting

Bootstrap **Aggregating**

- Take B repeated random subsets of samples from the training set, fit a model on each of the B samples and finally average all the predictions
- Reduces variance (especially for decision trees). Why?

Random Forest

- Similar to bagging, except each time a split in a tree is considered, a random sample of predictors is chosen is chose as split candidates.
- If there are p total predictors, a sample m is chosen where usually m = sqrt(p)
- Why do this?
- If this is one very strong predictor and many other moderate predictors, almost all trees in a bagged model will use the strong predictor in the top split; thus, all of the bagged trees will look similar; thus, predictions from the bagged trees will be highly correlated
- Choosing from a random subset for each tree decorrelates the predictions from each tree
- **When is a Random Forest a bagged decision tree?**

Extra Trees Classifier

- Similar to Random Forests except:
  - RFs compute the locally optimal feature/split combination
  - With Extra Trees (Extremely Randomized Trees), a random value is selected for the split

AdaBoost

- Fits a standard classifier (usually a decision tree)
- Finds all misclassified observations and increases their weights
- Fits a new DT on the weighted data
- Finds all misclassified observations and increases their weights
- Repeat

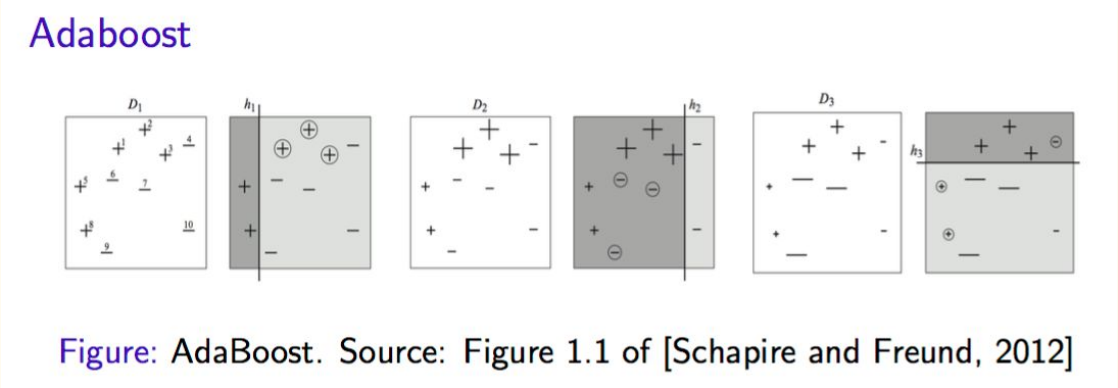

Figure: AdaBoost. Source: Figure 1.1 of [Schapire and Freund, 2012]

Gradient Boosting

- Generalized form of AdaBoost
- Find the between each observed y_i and F(x)_i which is the function to predict y_i (sort of like residuals)
- Model those pseudo-residuals
  - Residuals can be interpreted as negative gradient
- Aggregate these models and apply gradient descent algorithm to minimize loss function
- Find optimal number of trees to fit to prevent overfitting

XGBoost

- Provides a parallel tree boosting
- Installing XGBoost
  - https://www.ibm.com/developerworks/community/blogs/jfp/entry/Installing_XGBoost_on_Mac_OSX?lang=en
  -
- Using XGBoost with the Adult dataset (classify income as above or below $50k)
  - https://jessesw.com/XG-Boost/
- Guide to Parameter-Tuning
  - https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
  -

Conclusion

- Random Forests?
- XGBoost?