Introduction to Machine Learning Applications

Spring 2023

Ensemble modeling

Minor Gordon

gordom6@rpi.edu

Rensselaer

Ensemble Modeling

"**Ensemble modeling** is the process of running two or more related but different analytical **models** and then synthesizing the results into a single score or spread in order to improve the accuracy of predictive analytics and data mining applications."

From: http://searchbusinessanalytics.techtarget.com/definition/Ensemble-modeling

Categories of Ensemble Tree Models

- **Boosted Trees** Incrementally training each new instance to emphasize the training instances previously mis-modeled.
  - **ADABOOST (Adaptive Boosting)**
  - **XGBOOST (Gradient Boosting)**
- **Bagged Trees (Bootstrap Aggregating):** Create many different trees by repeatedly resampling with replacement.
  - **RANDOM FOREST** (one of many)

For each there are many other related models

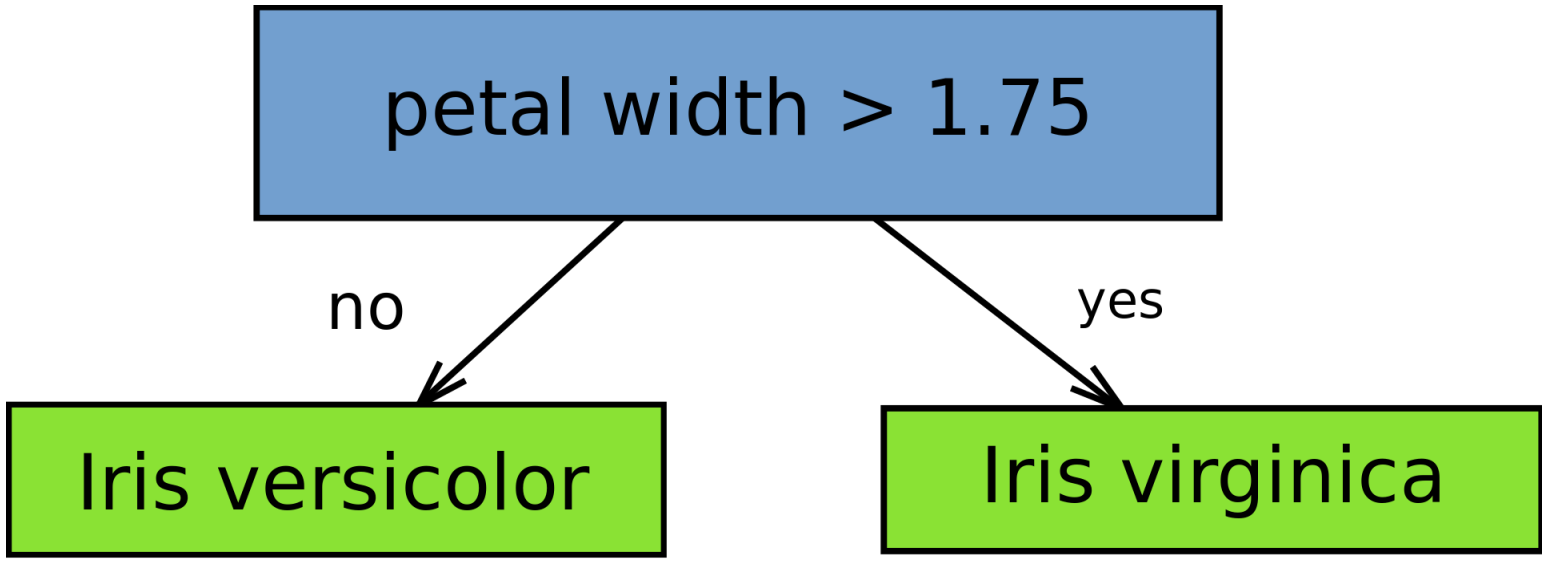
Boosting

Boosting with AdaBoost

1. Initialize training set weights
2. For T rounds (where T = # of weak classifiers to train)
   a) Train a classifier
   b) Update the weights of training set: increase if classified wrongly by the classifier, decrease if correctly
3. Form a final strong classifier as a linear combination (e.g., weighted sum) of T weak classifiers
   - For example, classifiers with 50% accuracy (random) have weight 0
   - Classifiers with < 50% accuracy have negative weight

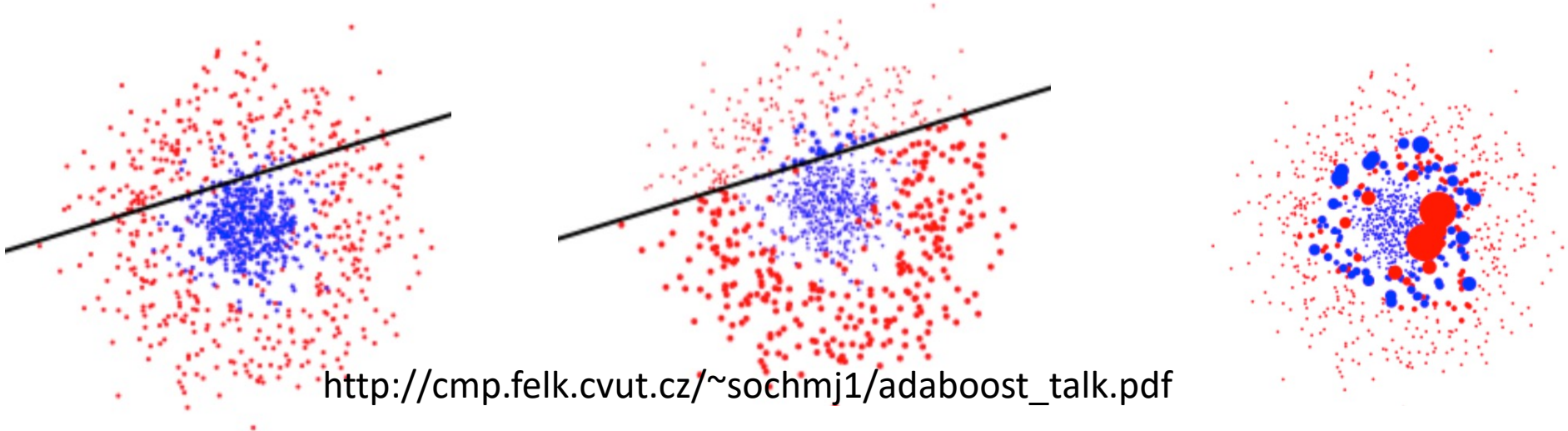
Weak classifier

- Classifier that performs poorly, but better than random guessing
- Example: decision stump



AdaBoost

- AdaBoost is an ensemble learning method for classification, regression
- AdaBoost incorporates *feature selection,* selecting only those features known to improve the predictive power of the model
- Weighting process shown below, with misclassified results weighted more.



http://cmp.felk.cvut.cz/~sochmj1/adaboost_talk.pdf

Bagging (Bootstrap Aggregating)

Bagging

1. Given a standard training set D of size n, bagging generates *m* new training sets D_i each of size n', by sampling from D uniformly and with replacement (i.e., by taking a bootstrap sample).
2. *M* models are fitted to the new training sets.
3. The M models are combined by averaging the output (for regression) or voting (for classification).

Random Forest

- Random forest is an *ensemble* learning method that combines *feature selection* and decision trees
  - Randomly select a subset of data
  - Output the class of the mode (most frequently occurring prediction) of the trees
- **Random forests** are "**bagging**" algorithms that aim to reduce the complexity of models that overfit the training data.