Machine Learning Basics II

Agenda

- Why should we combine classifiers ?
- Voting Classifer
- Ensemble learning
- Ensemble: Parallel Learners
- Bagging
- Out of bag errors
- Random Forest
- Variable Importance
- Ensemble: Boosted Learners
- AdaBoost
- XGBoost
- TPOT: Genetic Evolution

Why should we combine classifiers ?

Voting Classifier

How to combine the classifiers?

- (weighted) Majority voting
  - Class label output
  - Select the class most voted for

$$\sum_{t=1}^{T} d_{t,J} = \max_{j=1}^{C} \sum_{t=1}^{T} d_{t,j}.$$

$$\sum_{t=1}^{T} w_t d_{t,J} = \max_{j=1}^{C} \sum_{t=1}^{T} w_t d_{t,j}$$

- Mean rule
  - Continuous output
  - Support for class w_j is average of classifier output

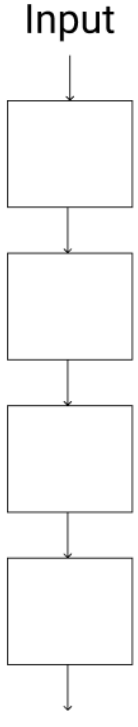
$$\mu_j(\mathbf{x}) = \frac{1}{T} \sum_{t=1}^{T} d_{t,j}(\mathbf{x})$$

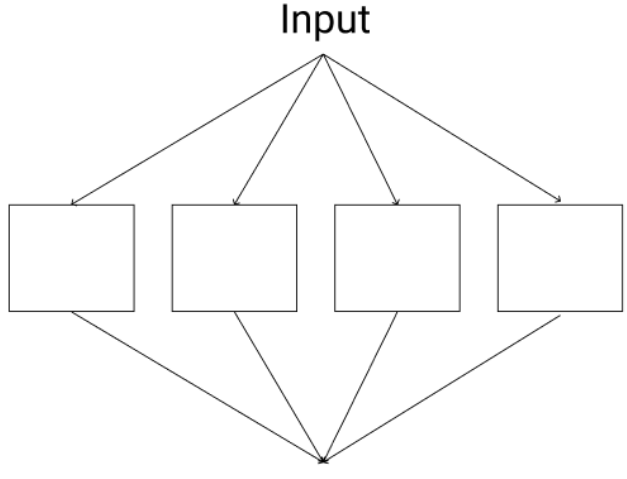
- Product rule
  - Continuous output
  - Product of classifier output

$$\mu_j(\mathbf{x}) = \frac{1}{T} \prod_{t=1}^{T} d_{t,j}(\mathbf{x})$$

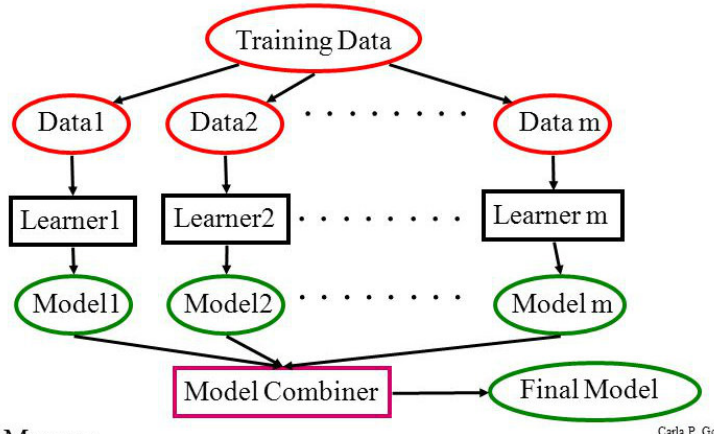
www.danielroggen.net droggen@gmail.com

Ensemble learning



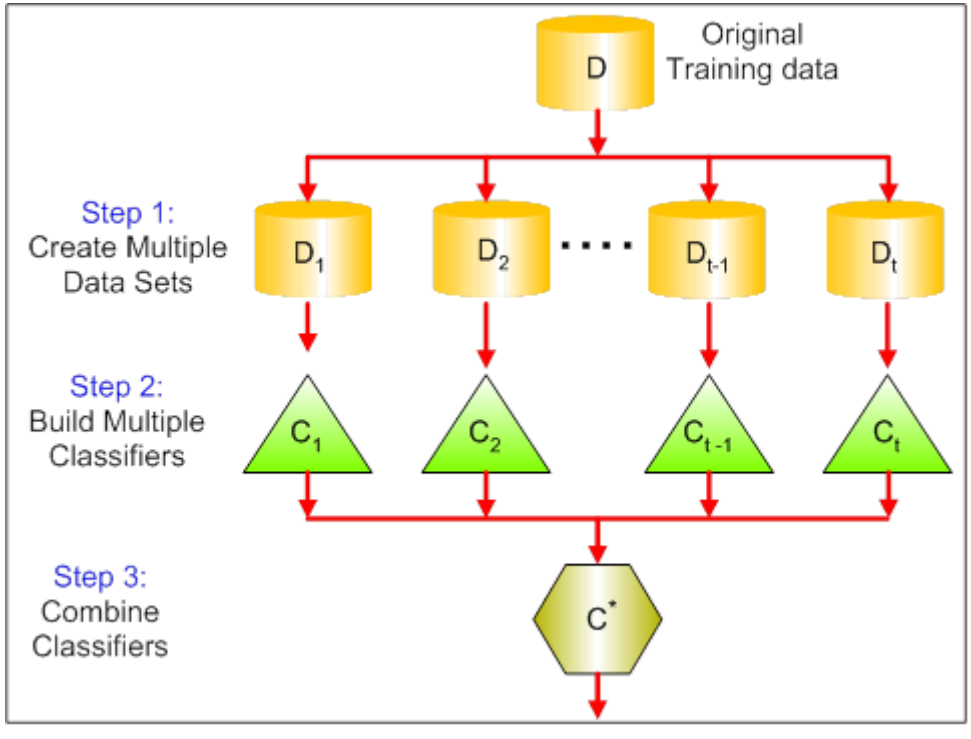


Ensemble learning: Parallel Learning



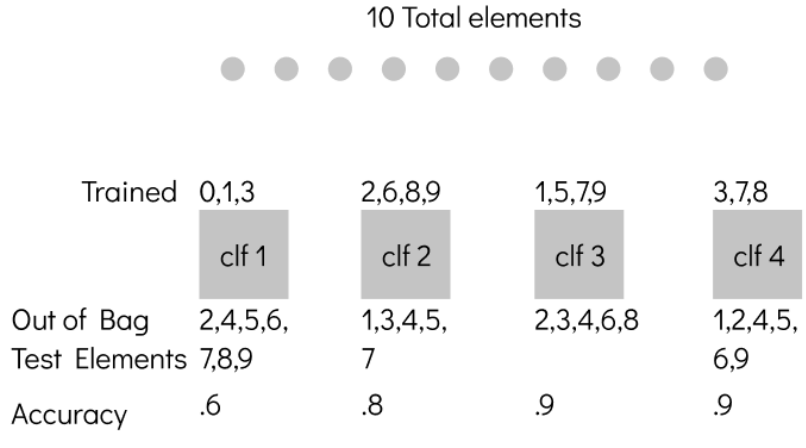
Source: Ray Mooney

Bagging


D
Original
Training data
Step 1:
Create Multiple
Data Sets
D1
D2
Dt-1
Dt
Step 2:
Build Multiple
Classifiers
C1
C2
Ct-1
Ct
Step 3:
Combine
Classifiers
C*

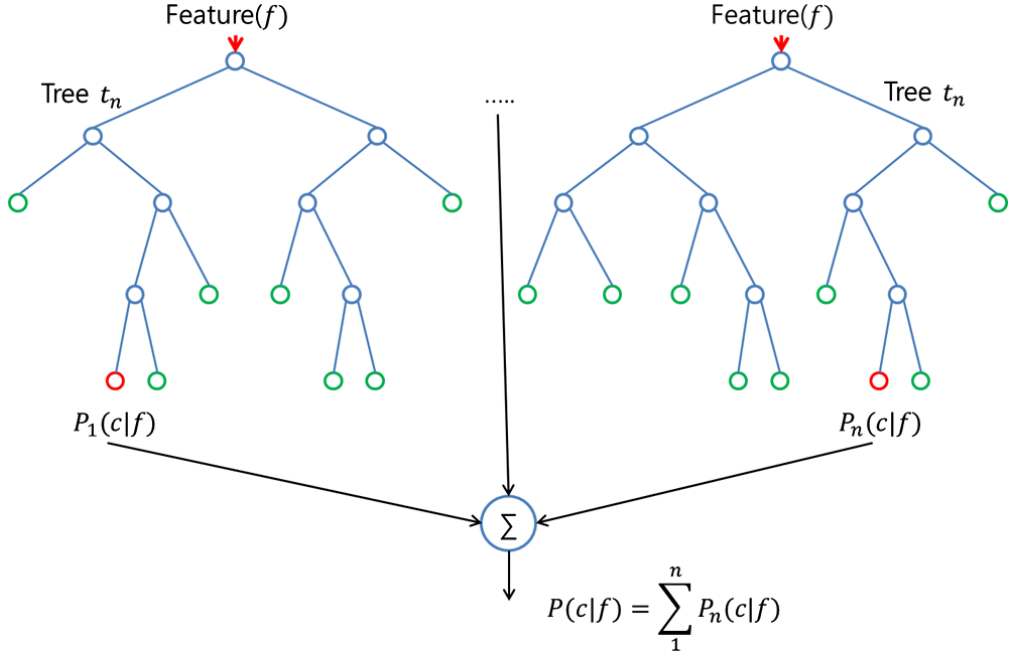
Out of Bag Errors

The out-of-bag estimate is as accurate as using a test set of the same size as the training set.



Out of Bag Error Rate = Mean of Out of Bag Accuracies
= 0.8

Random Forest



Random Forest: Parameters

The main parameters to consider tuning are:

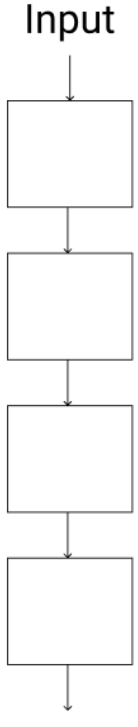
- n_estimators : The number of trees in the forest.
- criterion : The function to measure the quality of a split. Supported criteria are â□□giniâ□□ for the Gini impurity and â□□entropyâ□□ for the information gain. Note: this parameter is tree-specific.
- max_features : The number of features to consider when looking for the best split
- max_depth : The maximum depth of the tree.

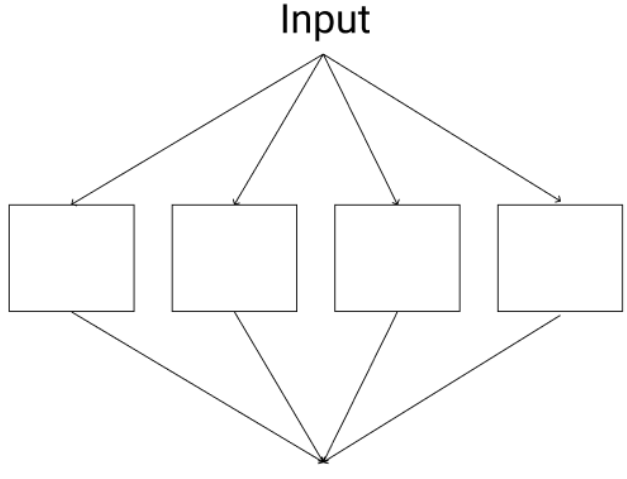
Variable Importance

The importance of each variable can be found by using 2 techniques:

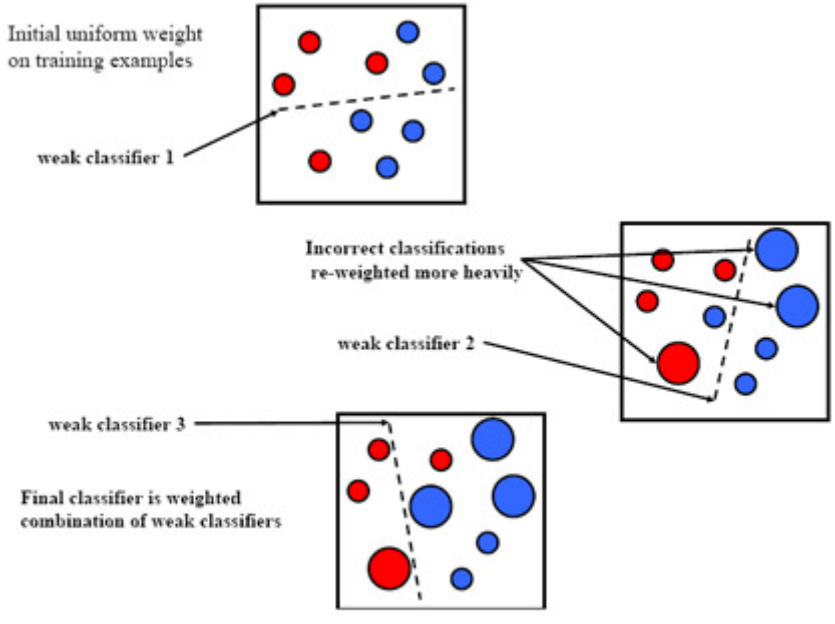
- Gini Importance: It is the average of the entropy gain in each tree (bias towards features with lots of categories).
- Permutation Importance: It is the effect on the OOB if the values of the feature are randomly permuted.

Ensemble: Boosted Learners





Adaboost



$$H(x) = sign(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \alpha_3 h_3(x))$$

Adaboost (II)

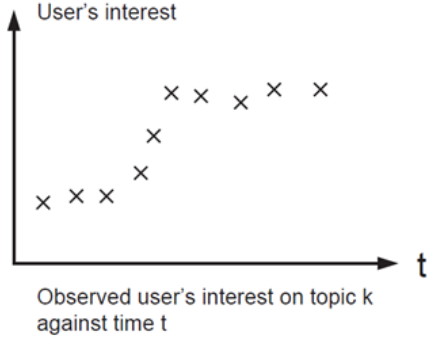
- Start same weight for all points: $\alpha_i = 1/N$
- For t = 1,…,T
  - Learn $f_t(\mathbf{x})$ with data weights α_i
  - Compute coefficient $\hat{w}_t$
  - Recompute weights α_i
- Final model predicts by:

$$\hat{y} = sign\left(\sum_{t=1}^{T} \hat{w}_t f_t(\mathbf{x})\right)$$

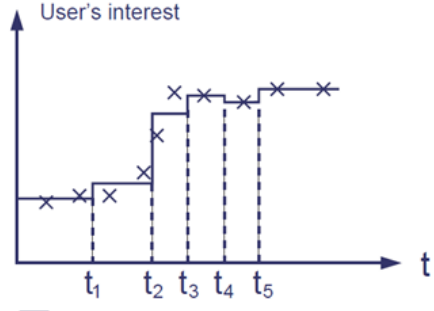
$$\hat{w}_t = \frac{1}{2}\ln\left(\frac{1 - weighted\_error(f_t)}{weighted\_error(f_t)}\right)$$

$$\alpha_i \leftarrow \begin{cases} \alpha_i e^{-\hat{w}_t}, & \text{if } f_t(\mathbf{x}_i) = y_i \\ \alpha_i e^{\hat{w}_t}, & \text{if } f_t(\mathbf{x}_i) \neq y_i \end{cases}$$

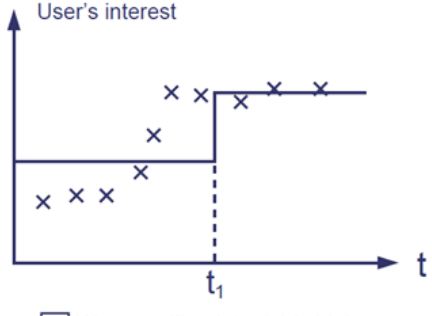
XGboost: Complexity



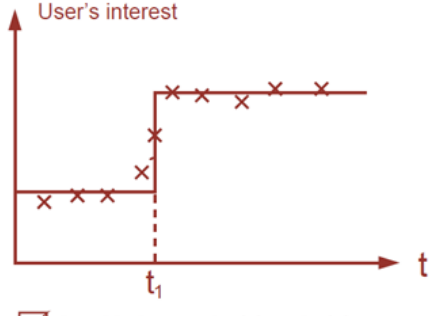
Observed user's interest on topic k against time t



☒ Too many splits, $\Omega(f)$ is high

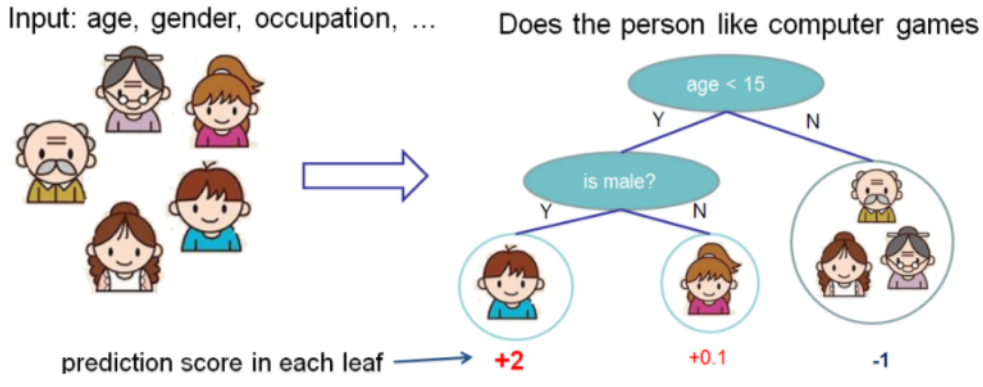


☒ Wrong split point, $L(f)$ is high

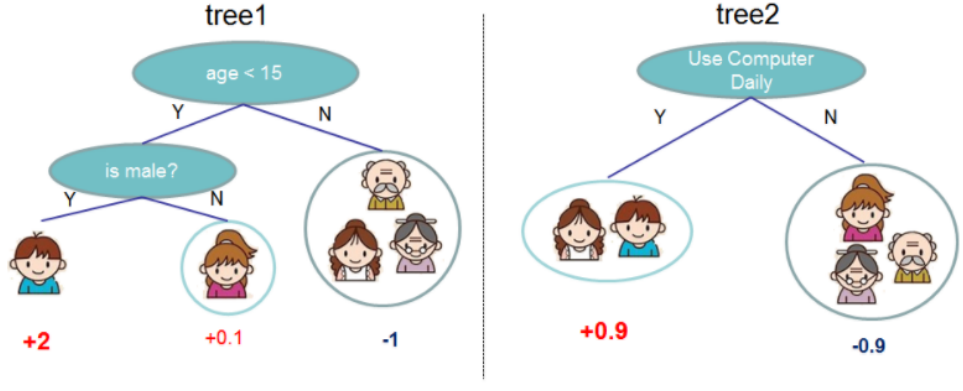


☑ Good balance of $\Omega(f)$ and $L(f)$

XGboost: Does a person like video games ?



XGboost: Trees



XGboost: Optimization Function

$$Obj(\Theta) = L(\Theta) + \Omega(\Theta)$$

Training Loss measures how well model fit on training data

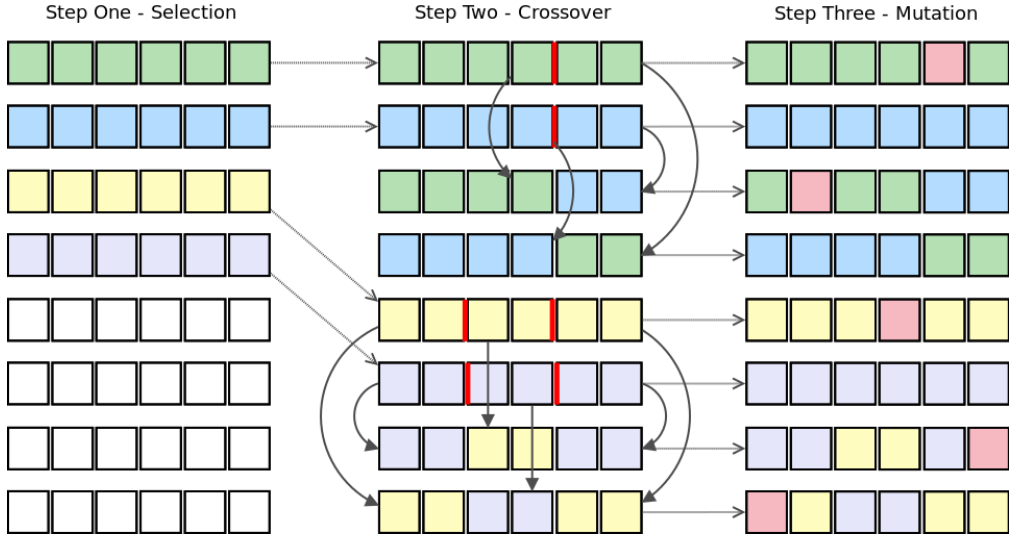
Regularization, measures complexity of model

XGboost: Parameters

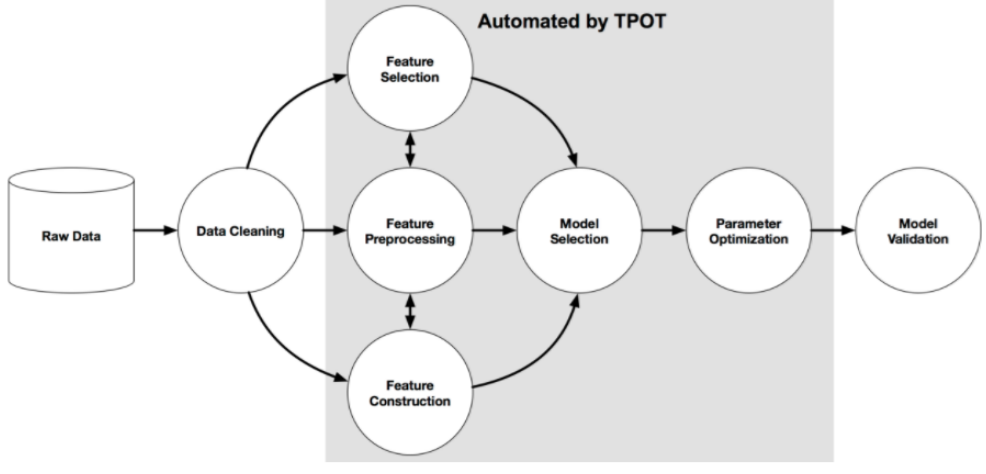
The parameters to consider tuning are:

- The number and size of trees (n_estimators and max_depth).
- The learning rate and number of trees (learning_rate and n_estimators).
- The row and column subsampling rates (subsample, colsample_bytree and colsample_bylevel).

TPOT: Genetic Algorythm



TPOT


Automated by TPOT
Feature Selection
Raw Data
Data Cleaning
Feature Preprocessing
Model Selection
Parameter Optimization
Model Validation
Feature Construction