Lecture 10 – Random Forest

Amit K. Verma, A.D. (Tony) Rollett

Revised: 21st Apr., 2021

Carnegie Mellon University

Overview

- Topics Covered

- Random Forest – Regressor
    - Decision Trees
    - Tree Pruning
    - Bagging
    - Define RF
    - Confidence Intervals
    - Feature Importance
    - Partial Dependence
    - Tree Interpreter

- Random Forest – Classifier
    - Gini Impurity
    - Metrics – Precision; Recall; Accuracy

Carnegie Mellon University

Regression; Feature Selection; Dimension Reduction

$y = \beta_0 + \beta_1 x + \varepsilon$; *assuming the functional form*

log(da/dN) $= \beta_0 + \beta_1 x + \varepsilon$; which x? -- *LR*

log(da/dN) $= \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots.. + \beta_n x_n + \varepsilon$; -- *Multiple LR*

Subset Selection; Bias-Variance Trade-off; Ridge / Lasso Regularization

<u>Principal Component Analysis</u> – Maximizes the variance of the data
Data – Semiconductor Compounds

$\{YS, UTS, Elong, RA\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots.. + \beta_n x_n + \varepsilon$; -- <u>*CCA*</u>

Sensitivity Analysis & GB Character with Properties – <u>*CCA*</u>

Carnegie Mellon University

Dataset: High Entropy Alloys

Composition (24 Elements); Phases (5 Phases); Rule of Mixtures (ROM) Density

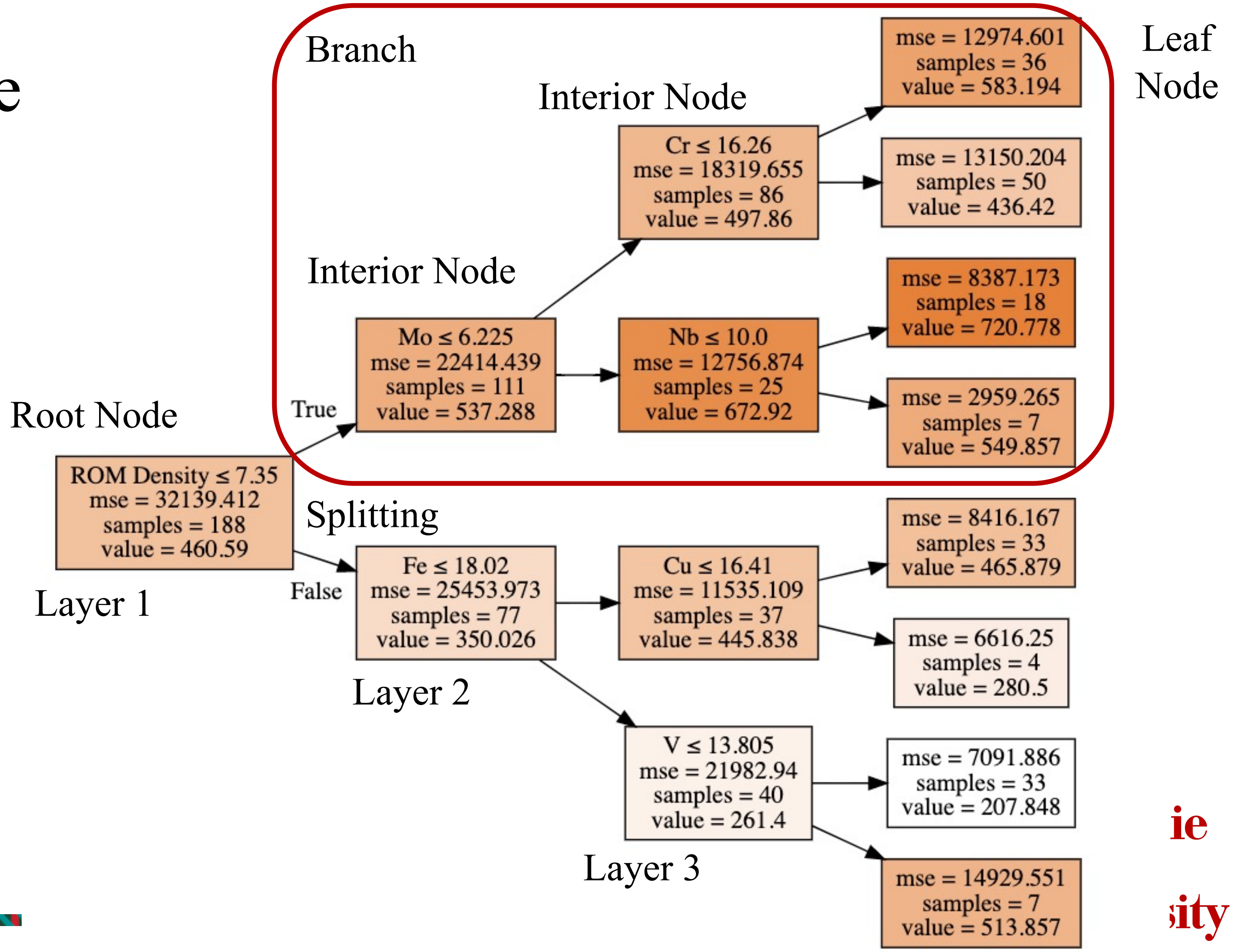
Predict: Vickers Hardness

	Al	Co	Cr	Cu	Fe	Hf	Mo	Nb	B	C	...	Zr	Zn	Y	BCC	FCC	Im	HCP	B2	ROM Density	Vickers Hardness
0	NaN	33.33	NaN	NaN	33.33	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	0	0	0	8.5	125.0
1	NaN	33.33	NaN	NaN	33.33	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	0	0	0	8.5	125.0
3	NaN	30.77	NaN	NaN	30.77	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	0	0	0	7.7	149.0
4	NaN	28.57	NaN	NaN	28.57	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	1	0	0	7.1	287.0
5	NaN	26.67	NaN	NaN	26.67	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	1	0	0	6.6	570.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
349	NaN	17.86	NaN	NaN	17.86	NaN	17.86	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	1	0	0	8.5	520.0
350	NaN	17.24	NaN	NaN	17.24	NaN	17.24	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	1	0	0	8.5	510.0
351	NaN	16.67	NaN	NaN	16.67	NaN	16.67	NaN	NaN	NaN	...	NaN	NaN	NaN	0	1	1	0	0	8.5	382.0
352	NaN	14.29	NaN	NaN	14.29	NaN	14.29	NaN	NaN	NaN	...	14.29	NaN	NaN	0	0	0	0	0	7.3	790.0
353	NaN	NaN	NaN	16.67	16.67	NaN	NaN	NaN	NaN	NaN	...	16.67	NaN	NaN	0	0	0	0	0	6.8	590.0

236 rows × 31 columns

Decision Tree

1. Top-down greedy approach: best split is made at that step
2. Splitting: regions that leads to the greatest possible reduction in RSS
3. Repeat the process



ie

ity

Decision Tree

1. Top-down greedy approach: best split is made at that step
2. Splitting: regions that leads to the greatest possible reduction in RSS
3. Repeat the process

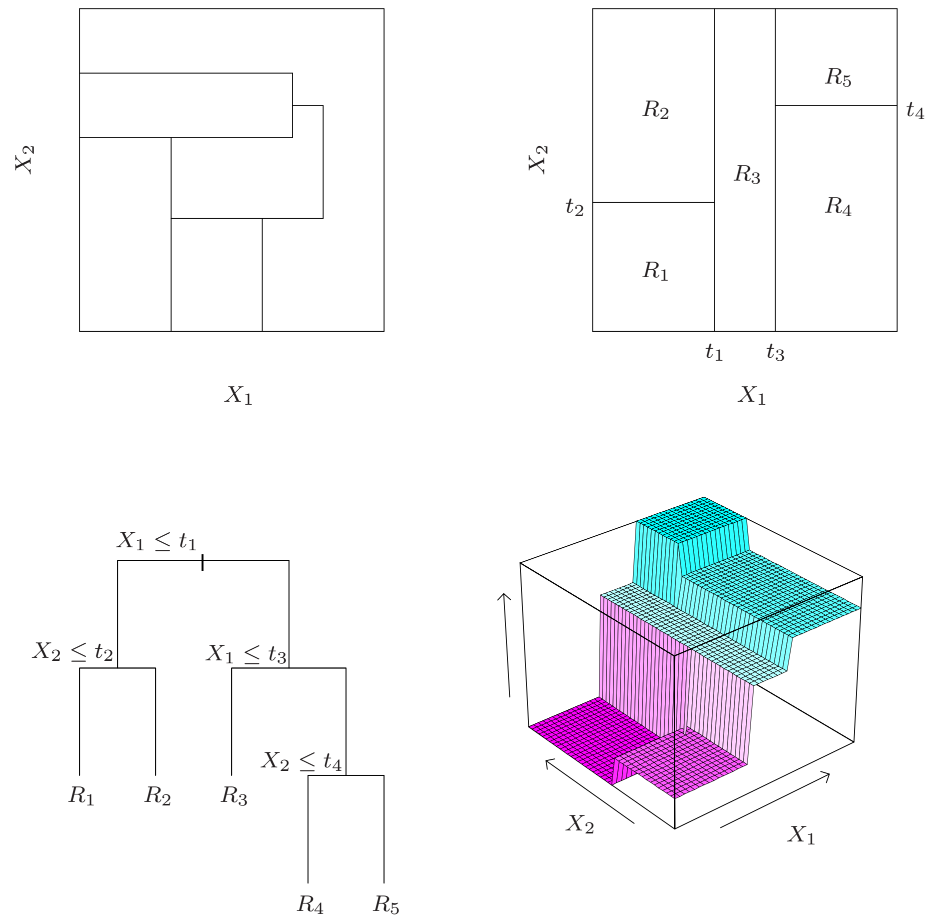
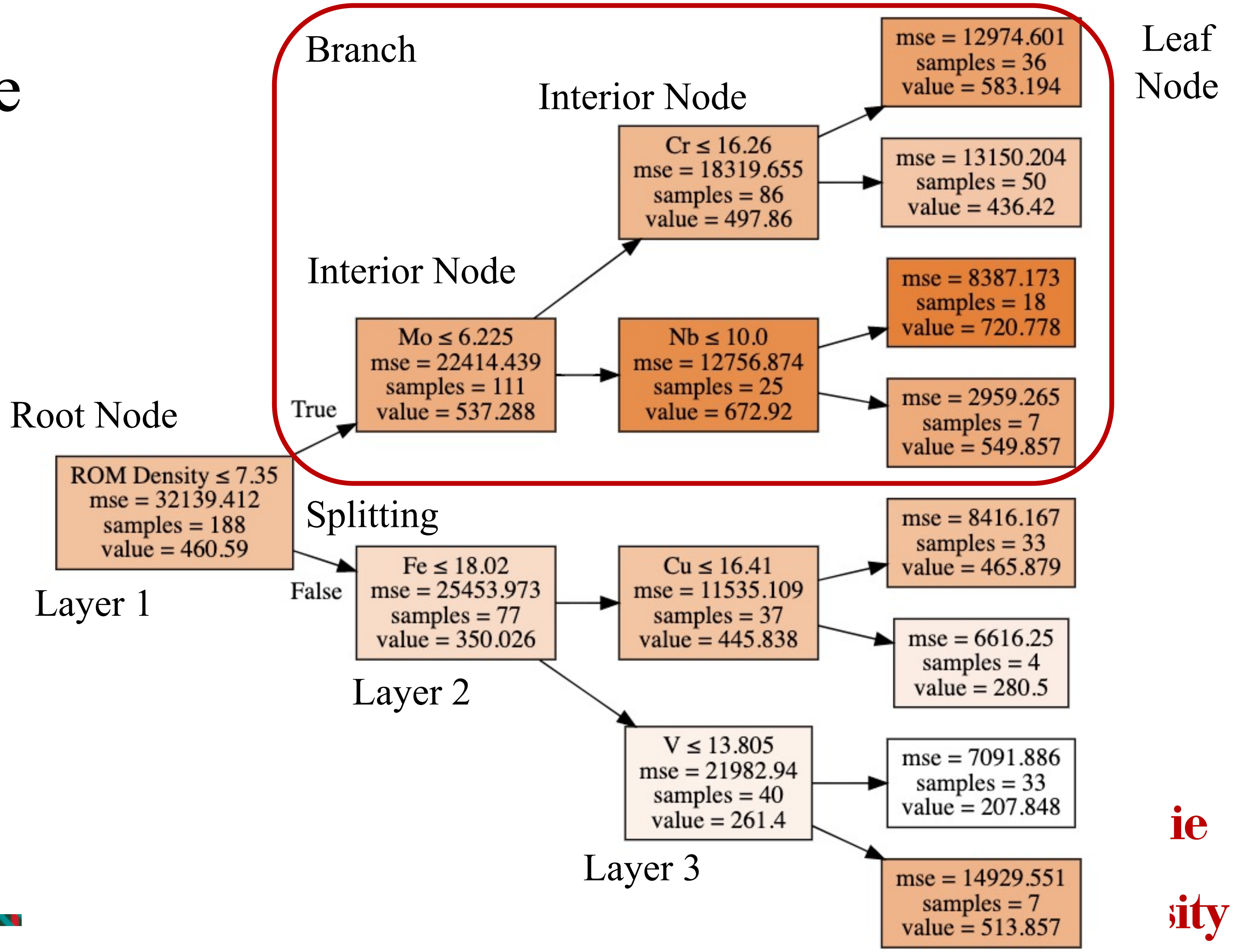


Image: (Figure 9.2) Chapter 9, Friedman et al. - 2009 - The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition

Decision Tree

1. Top-down greedy approach: best split is made at that step
2. Splitting: regions that leads to the greatest possible reduction in RSS
3. Repeat the process



ie
ity

Overfitting: Tree Pruning

- If each leaf node has only 1 sample – $R^2 \sim 1$ for the training dataset

    - minimum number of observations in a leaf node (number of rows)
    - grid search for finding the optimum value

- Suffers from high variance
    - a small change in the data can cause a large change in the final estimated tree

Bagging

- Reducing Variance: a natural way to reduce the variance and hence increase the prediction accuracy of a statistical learning method is to take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions

- Bagging: generate different bootstrapped training data sets
    - bootstrap: sampling with replacement

    - each bagged tree makes use of around 2/3 of the observations
    - remaining 1/3 of the observations are referred to as the out-of-bag (OOB) observations

- Each individual tree has high variance, but low bias, averaging these trees reduces the variance

- Reduce overfitting; reduce bias; break the bias-variance trade-off

Carnegie Mellon University

Random Forest

- Bagged Trees (greedy algorithm) + a small tweak that decorrelates the trees
    - Suppose that there is one very strong predictor in the data set, along with several other moderately strong predictors. Then in the collection of bagged trees, most or all the trees will use this strong predictor in the top split. Consequently, all the bagged trees will look quite like each other

- Each time a split in a tree is considered, a random sample of m predictors is chosen as split candidates from the full set of p predictors
    - Using a small value of m in building a random forest will typically be helpful when we have many correlated predictors

Decision Trees vs Random Forest

+ Trees yield insight into decision rules

+ Rather fast

-- Prediction of trees tend to have a high variance

+ RF has smaller prediction variance and therefore usually a better general performance

+ OOB error “for free” (no CV needed)

-- Rather slow

-- “Black Box”: Rather hard to get insights into decision rules

Carnegie Mellon University

Attributes

- No statistical assumptions
- Works with any kind of data – continuous / categorical – intrinsically multiclass
- Can express any function – regression / classification
- Works well with small to medium data, unlike neural network which requires large data
- Can handle thousands of input variables without variable selection
    - provide feature importance
- It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing

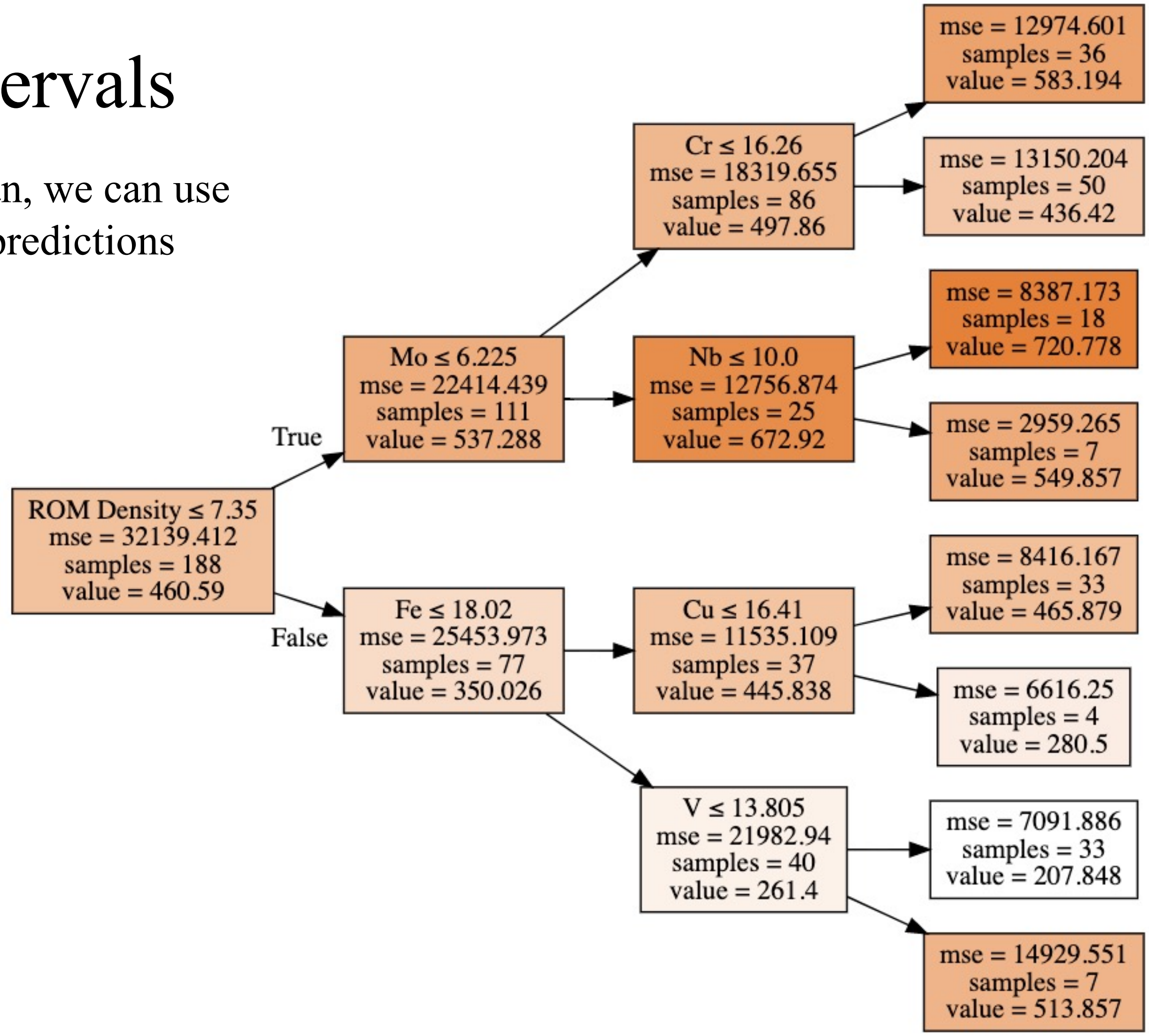
Carnegie Mellon University

Let’s look at the code!

Ref: fast.ai (http://course18.fast.ai/ml)
github: https://github.com/fastai/fastai

Confidence Intervals

- Instead of just the mean, we can use standard deviation of predictions



Feature Importance (Overall Interpretation)

1. How much each feature decreases the variance in a tree
    - For a forest, the variance decrease from each feature can be averaged and the features are ranked according to this measure
    - Biased towards preferring variables with more categories https://link.springer.com/article/10.1186/1471-2105-8-25
    - When dataset has two (or more) correlated features, then one shows up high while other as low (applies to other methods too)
        - *The effect of this phenomenon is somewhat reduced by random selection of features at each node creation*

2. Random shuffling of the variable
    - permute the values of each feature and measure how much the permutation decreases the accuracy of the model
    - The OOB data is passed along each tree to determine the "test error" (since the OOB were not used to train). See section 15.3.1 in Hastie *et al*.
    - For each variable, the values are permuted in the OOB to evaluate the sensitivity to that variable (from the increase in the test error).

Partial Dependence (Single Feature)

- Useful to isolate the effect of a feature
- Based on : https://arxiv.org/abs/1309.6392

- Fill the feature with a constant value and pass through the model

A	B	C	Y
A1	B1	C1	Y1
A2	B2	C2	Y2
A3	B3	C3	Y3

A	B	C	Y	mean
A1	B1	C1	Y11	Y(A1)
A1	B2	C2	Y21	
A1	B3	C3	Y31	
A2	B1	C1	Y12	Y(A2)
A2	B2	C2	Y22	
A2	B3	C3	Y32	
A3	B1	C1	Y13	Y(A3)
A3	B2	C2	Y23	
A3	B3	C3	Y33	

Carnegie Mellon University

Images from: https://towardsdatascience.com/introducing-pdpbox-2aa820afd312

Tree Interpreter (Local Interpretation)

- Explains a prediction for a given data point
- Gives the sorted list of bias (mean of data at starting node) and individual node contributions for a given prediction

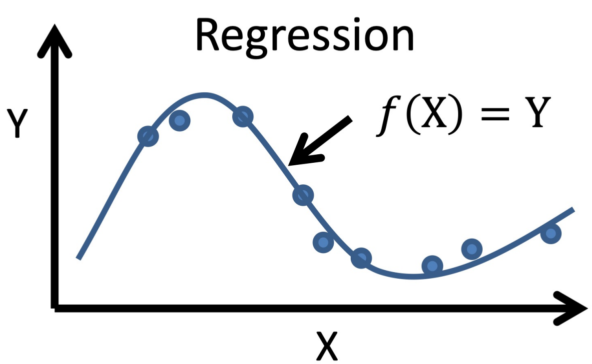
RF Classifier

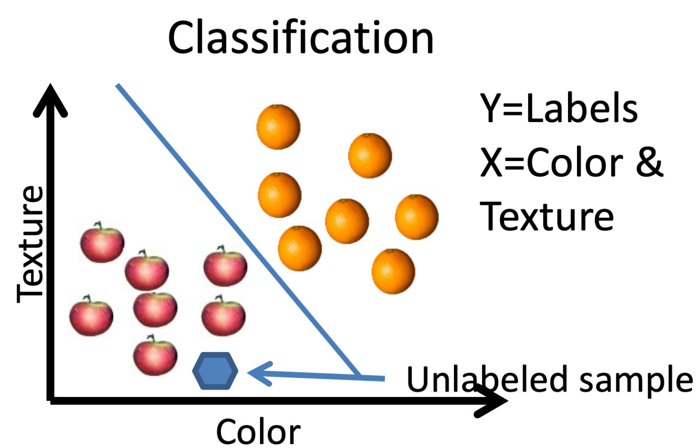
Supervised Learning

- Find the function f that maps the input data x to the output data y

$$f: x \rightarrow y$$

- y is continuous: Regression
- y is discrete: Classification
- Cross-validation to check performance & determine parameters





RF Classifier

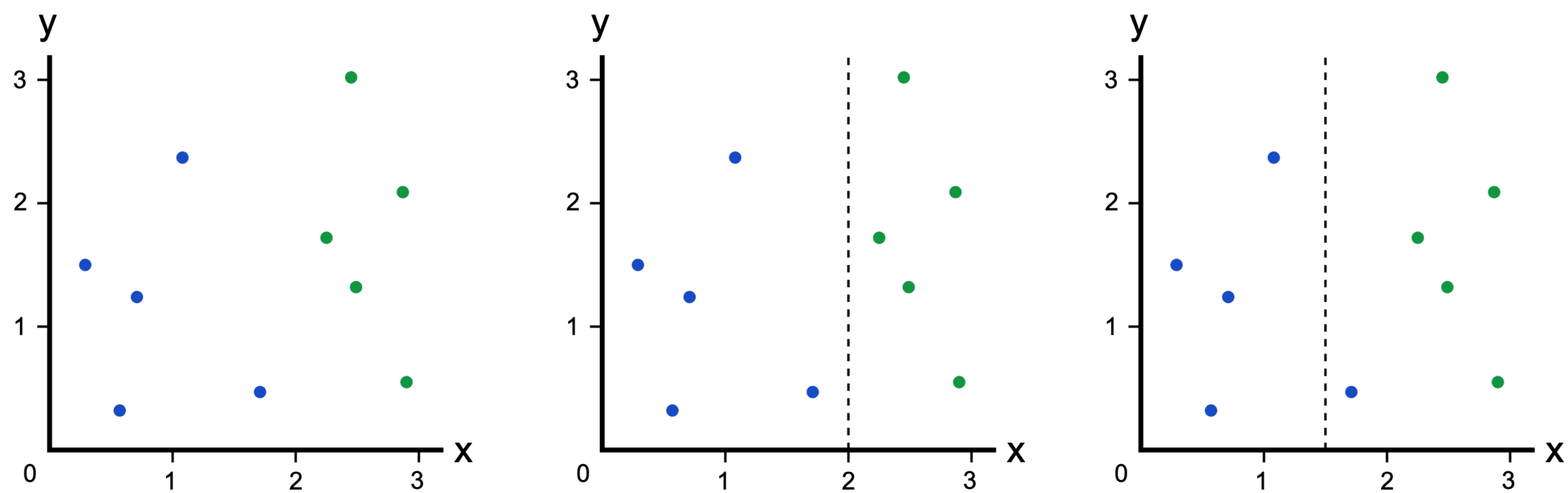
Gini Impurity

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Information Entropy

$$E = -\sum_{i}^{C} p_i \log_2 p_i$$

p(i) – probability of randomly picking an element of class *i*



Dataset Perfect Split Imperfect Split

Images: https://victorzhou.com/blog/gini-impurity/

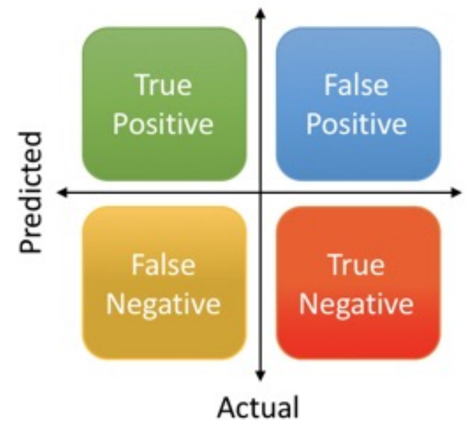
RF Classifier - Metrics

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$

Confusion Matrix

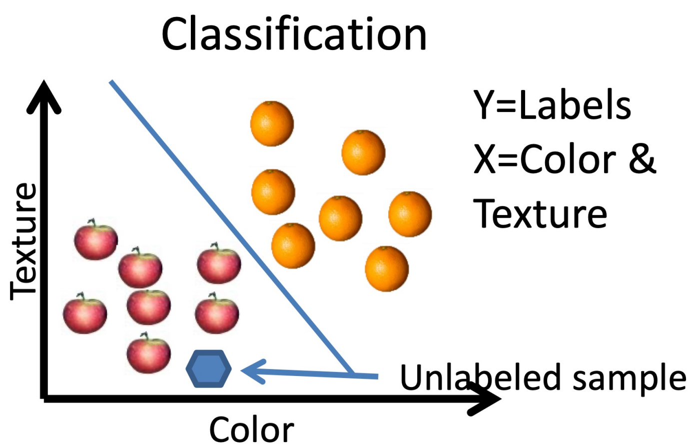


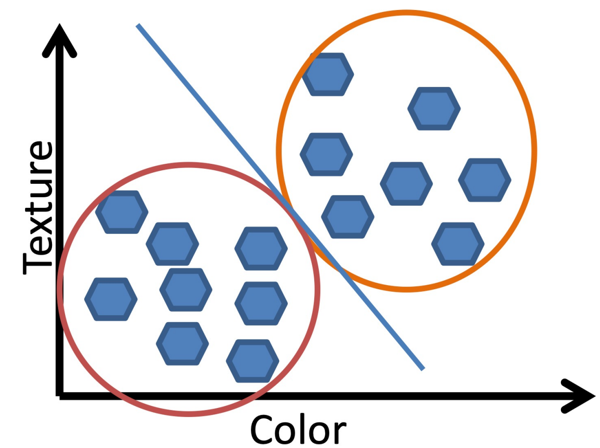
- Precision : reflects how reliable the model is in classifying samples as Positive
- Recall : measures the model's ability to detect Positive samples
- Accuracy : fraction of total correct predictions

Carnegie Mellon University

Unsupervised Learning

- “Unsupervised”: We don’t have output data y (“learning without a teacher”)
- Interested in relationship between the data x
- Learn about x from its distribution
- Cross-validation for algorithm performance isn’t available
    - Performance checked with: Heuristics & Expert analysis





Note: Dimension reduction is the most common application of unsupervised learning

Questions

- Why is it that randomly selecting the validation dataset could be problematic?
- Why OOB score is less than the score for validation dataset?