wbs
WARWICK BUSINESS SCHOOL
THE UNIVERSITY OF WARWICK

For the Change Makers

Dr. Iman Ahmadi
Assistant Prof. of Marketing

Marketing & Strategy Analytics:

Supervised Learning: Decision Trees

Announcements

- **This week's lecture:**

  Guest Speaker: Jerome Hancock from "SKIM"

- **No seminar for next week**

  Exercise on Datacamp!

Overview about Different Terms

- **Dataset**
  - Data table
  - Flat file
- **Observations**
  - Records
  - Instances, examples
- **(Independent) Variables**
  - Columns
  - Features
  - Attributes
  - Feature vector (= set of independent variables)
- **Dependent variable**
  - Target variable
  - Data class

Description of Decision Trees

Decision tree:

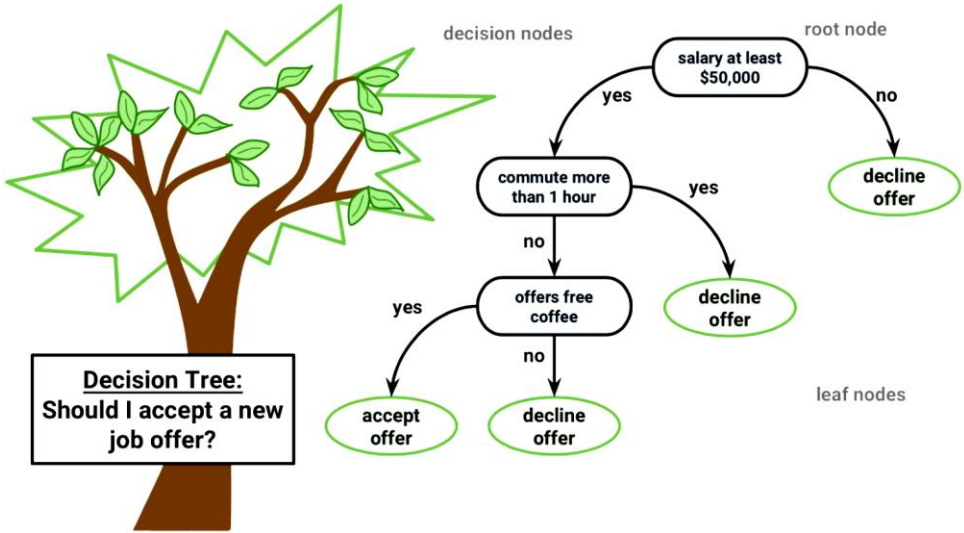
- presents knowledge in form of tree structures
- series of rules to predict an output variable
- examples: credit scoring, satisfaction-churn studies, diagnosis of medical conditions, …

Root node: beginning of tree

Decision nodes: choices to be made

Branches: leads either to outcome or decision node

Leaf/terminal nodes: outcome, e.g. action to be taken



Lantz, B. (2015) *Machine Learning with R (Second edition)*. Birmingham: Packt Publishing. Chapter 5.

Entropy as Splitting Criteria for Information Gain

- **Entropy: measure of impurity**

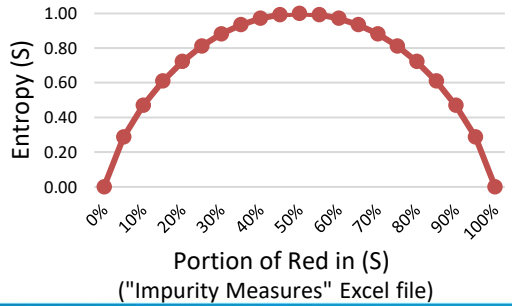
$$Entropy(S) = -\sum_{i=1}^{k} p_i \log_2(p_i),$$

S: subset of data;
k: number of different classes of dependent variable;
p_i: share of values (i.e., probability) falling into class *i*

- **Example: Entropy of subset (S) with two classes, red (60%) and blue (40%)**

$$Entropy(S) = -(0.6 \times \log_2(0.6) + 0.4 \times \log_2(0.4)) = 0.97$$

- **Entropy varies between 0 and 1**: the higher the Entropy, the higher heterogeneity of dependent variable
  - red (100%) and blue (0%) → Entropy = 0
  - red (50%) and blue (50%) → Entropy = 1



Miguel Forte, R. (2015) *Mastering Predictive Analytics with R*. Birmingham: Packt Publishing. Chapter 6.

Using Information Gain to Determine Appropriate Split

- **Information Gain for possible (independent) variable (*F*): difference between entropy in segment**
  - before split
  - after split: weighted sum of Entropy in segments after split

$$Entropy\ after\ split(S \to S_1 \dots S_c) = \sum_{i=1}^{c} \frac{|S_i|}{|S|} E(S_i),$$

S: subset of data;
c: number of subsets after split;
$|S_i|$: size of subset i generated from S;
$|S|$: size of subset S;
$E(S_i)$: Entropy of subset i generated from S

$$Information\ Gain(F) = Entropy(S) - Entropy\ Split(S \to S_1 \dots S_c)$$

- **Pick split with highest Information Gain**

- **Example:**
  - Before split (N = 10):
    - 50% red balls, 50% blue balls: Entropy: 1
  - After split:
    - 100% red balls in class 1, 100% blue balls in class 2; Entropy: 50%×0 + 50%×0 = 0
  - Information Gain: 1 – 0 = 1

Miguel Forte, R. (2015) *Mastering Predictive Analytics with R*. Birmingham: Packt Publishing. Chapter 6.

Entropy as Splitting Criterion in Decision Tree (3/9)

Example (Decision Tree): whether a customer defaulted on a loan or not

Dep. var. (default)	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	No	No
Split on indep. var. A (purpose)	Auto	Auto	Auto	Auto	Tablet	Tablet	Tablet	Tablet	Tablet	House	House	House	House	House	House
Split on indep. var. B (credit history)	Bad	Bad	Bad	Bad	Bad	Bad	Bad	Good	Good	Good	Good	Good	Good	Good	Good

Alternative 1: $purpose_{Auto,Tablet}$ $purpose_{House}$

$$Entropy(default) = -\sum_{i=1}^{k} p_i \log_2(p_i) = -\left(\frac{7}{15} \times \log_2 \frac{7}{15} + \frac{8}{15} \times \log_2 \frac{8}{15}\right) = 0.997$$

$$Entropy\left(default\_purpose_{Auto,Tablet}\right) = -\left(\frac{5}{9} \times \log_2 \frac{5}{9} + \frac{4}{9} \times \log_2 \frac{4}{9}\right) = 0.991$$

$$Entropy\left(default\_purpose_{House}\right) = -\left(\frac{2}{6} \times \log_2 \frac{2}{6} + \frac{4}{6} \times \log_2 \frac{4}{6}\right) = 0.918$$

$$Information\ Gain = 0.997 - \left(\frac{9}{15} \times 0.991 + \frac{6}{15} \times 0.918\right) = 0.997 - 0.962 = 0.035$$

	Both Groups	$purpose_{Auto,Tablet}$	$purpose_{House}$
Entropy	0.997	0.991	0.918
Information Gain		$= 0.035$	

Entropy as Splitting Criterion in Decision Tree (9/9)

How many errors do you make?

Your Prediction for default	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	No	No
Dep. var. (default)	Yes	Yes	No	Yes	Yes	Yes	No	No	No	No	Yes	Yes	No	No	No
Split on indep. var. A (purpose)	Auto	Auto	Auto	Auto	Tablet	Tablet	Tablet	Tablet	Tablet	House	House	House	House	House	House
Split on indep. var. B (credit history)	Bad	Bad	Bad	Bad	Bad	Bad	Bad	Good	Good	Good	Good	Good	Good	Good	Good

Two-way table summarizing actual and predicted defaults

Actually defaulted	Predicted to default: No	Predicted to default: Yes
No	6	2
Yes	2	5

Errors (**error rate**): 4 (i.e., 26.67% = 4/15)

Exercise 4.2 – "Risky Bank Loans"

Download the dataset "Risky Bank Loans_20171009.csv" from My.WBS:

1. Use the variable ‘default’ and other variables as your dependent variable and independent variables, respectively.
2. Build a decision tree in R using C5.0 command to help you to predict whether an applicant will default on the mortgage or not.
3. Check the quality of your model precision (through two by two table)
4. Interpret your results.
5. Investigate the help for C5.0 and change the size of your tree
6. Investigate how changing the (a) number of used independent variables and (b) tree size affect the accuracy rate.

("RBL" R code)

Thank You!

Iman.Ahmadi@wbs.ac.uk
Room No.: 3.207