Big Data (MHI222956/MHI225101)

7.2 Decision Tree & Linear Regression

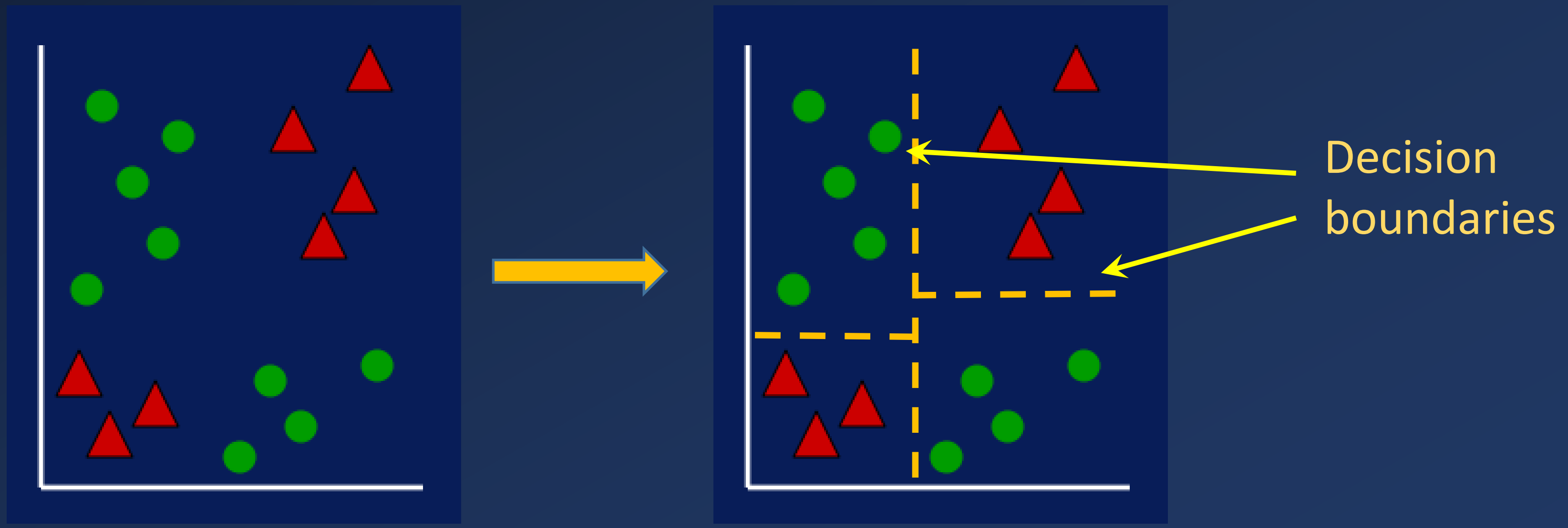
Decision Tree

Decision Tree

- Idea – Split data into “pure” region

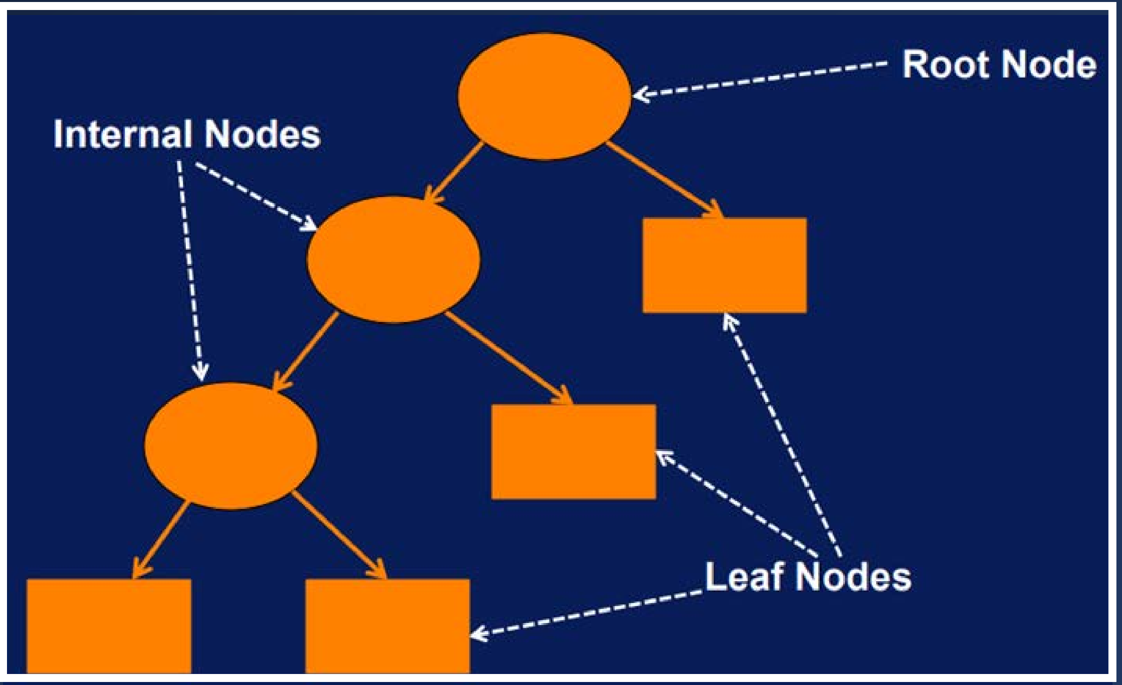
  i.e., each subset belongs to only one class

- With real data completely pure subsets may not be possible. So the goal is to divide the data into subsets that are as pure as possible.

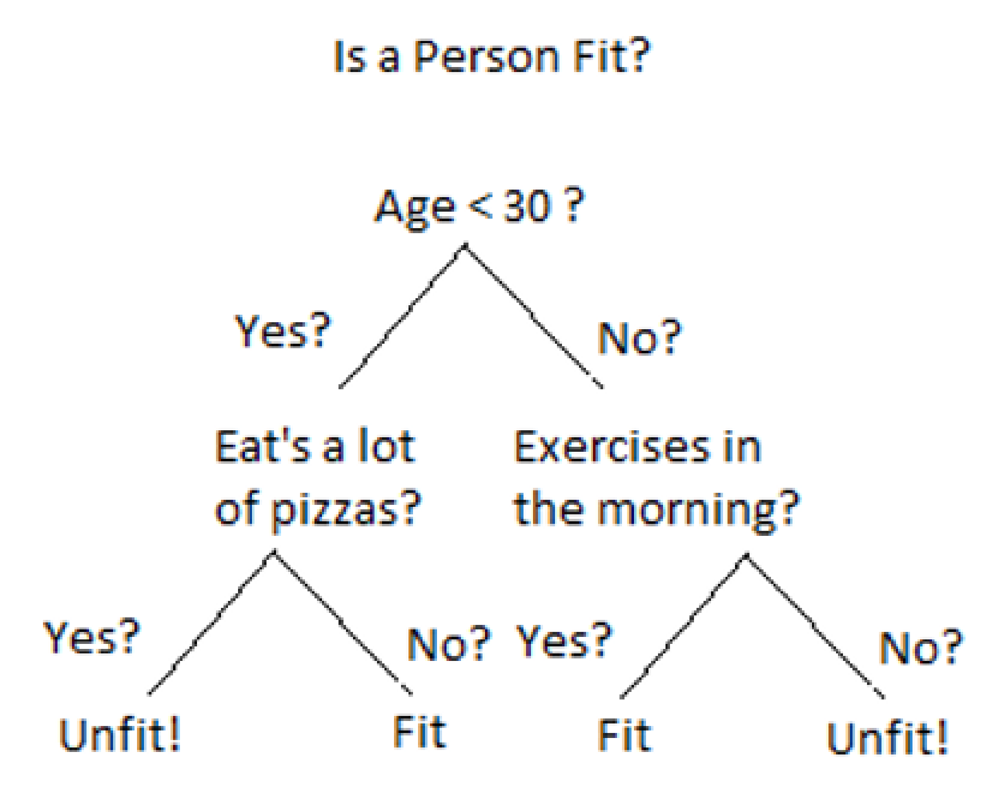


- A decision tree is a hierarchical structure with nodes and directed edges.
    - Root node - the node at the top
    - Leaf nodes - the nodes at the bottom
    - Internal nodes

- The root and internal nodes have test conditions
- Each leaf node has a class label associated with it



- Using a decision tree, a classification decision is made by:
  - Traversing the decision tree starting with the root node.
  - At each node the answer to the test condition determines which branch to traverse to.
  - When a leaf node is reached the category at the leaf node determines the classification decision.

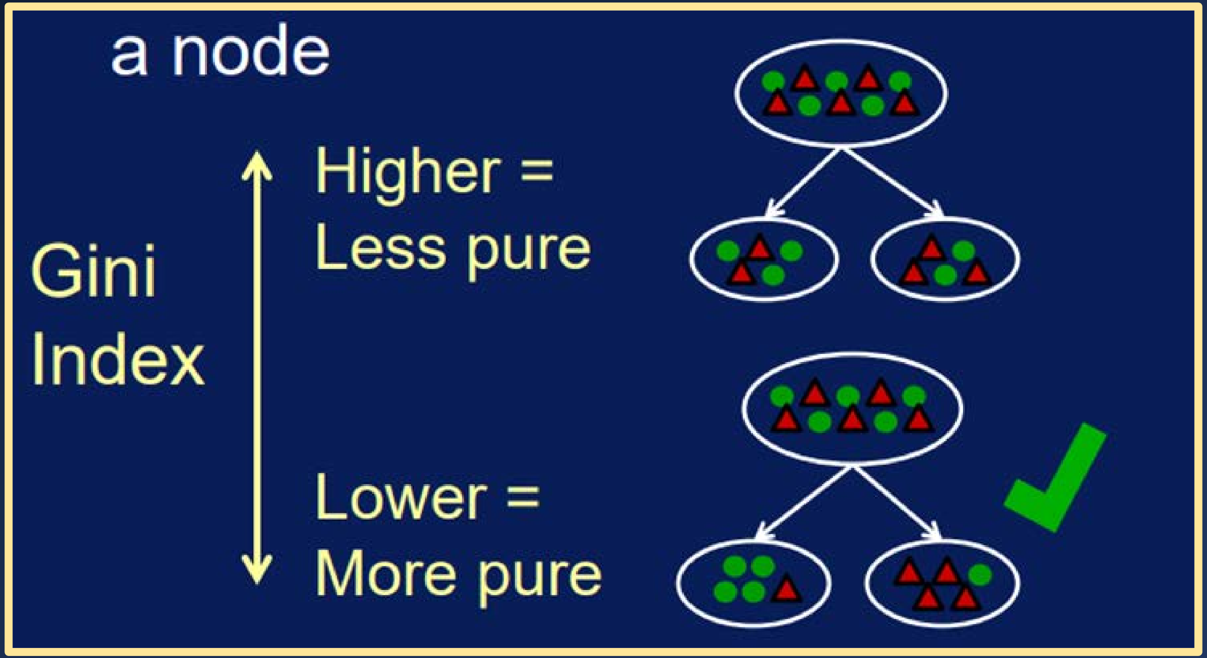


Constructing a Decision Tree

- Start with all samples and a node
- Partition samples into subsets based in the input variables to create subsets of records that are purest – ‘greedy approach’
- Repeatedly partition data into successively purer subsets until some stopping criterion is satisfied

Constructing a Decision Tree

- Impurity measure – Gini index



Constructing a Decision Tree

- What variable to split on?
  - Splits on all variables are tested
- When to stop splitting a node?
  - All samples in the node have the same class label.
  - Number of samples in the node falls below a certain minimum value
  - Change in impurity measure is smaller than threshold
  - Max tree depth is reached
  - etc.

- Resulting tree is often simple and easy to interpret
- Induction is computationally inexpensive
- Greedy approach does not guarantee best solution
- Rectilinear decision boundaries

Linear Regression

Simple linear regression

- A regression algorithm - the <u>output variable is a numeric value</u>

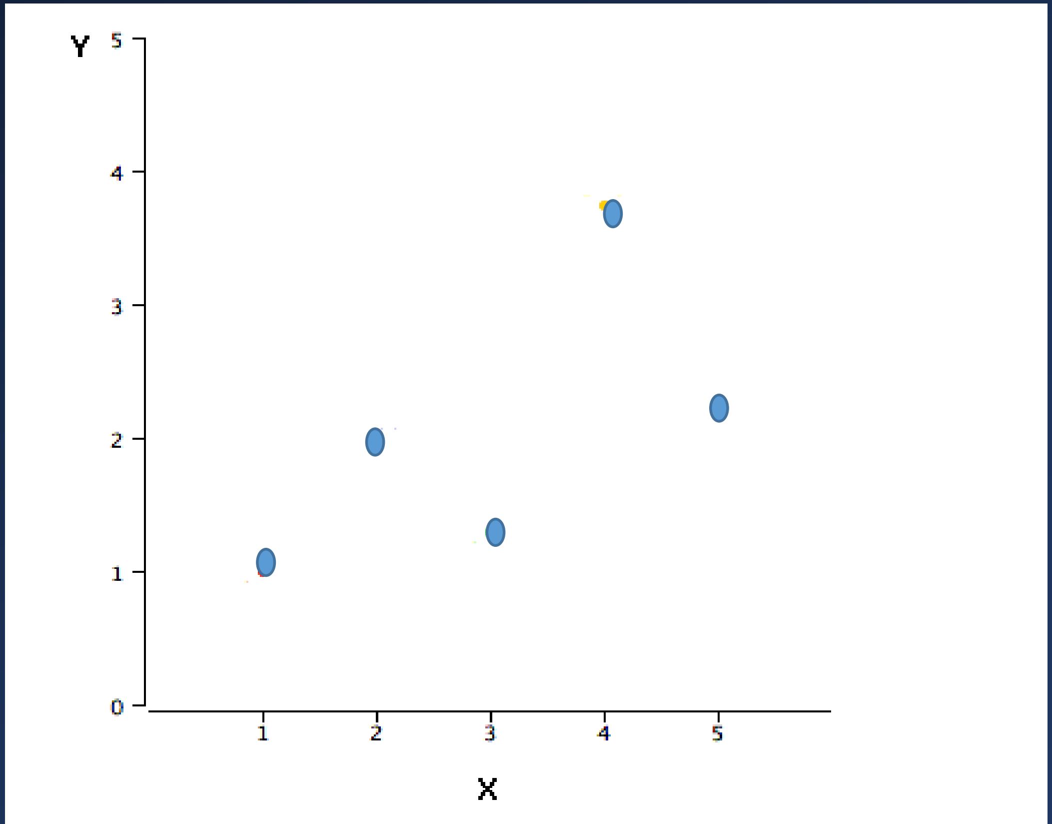
- A statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:
  - One variable, *x*, is regarded as the **predictor, explanatory**, or **independent** variable.
  - The other variable, *y*, is regarded as the **response, outcome**, or **dependent** variable.

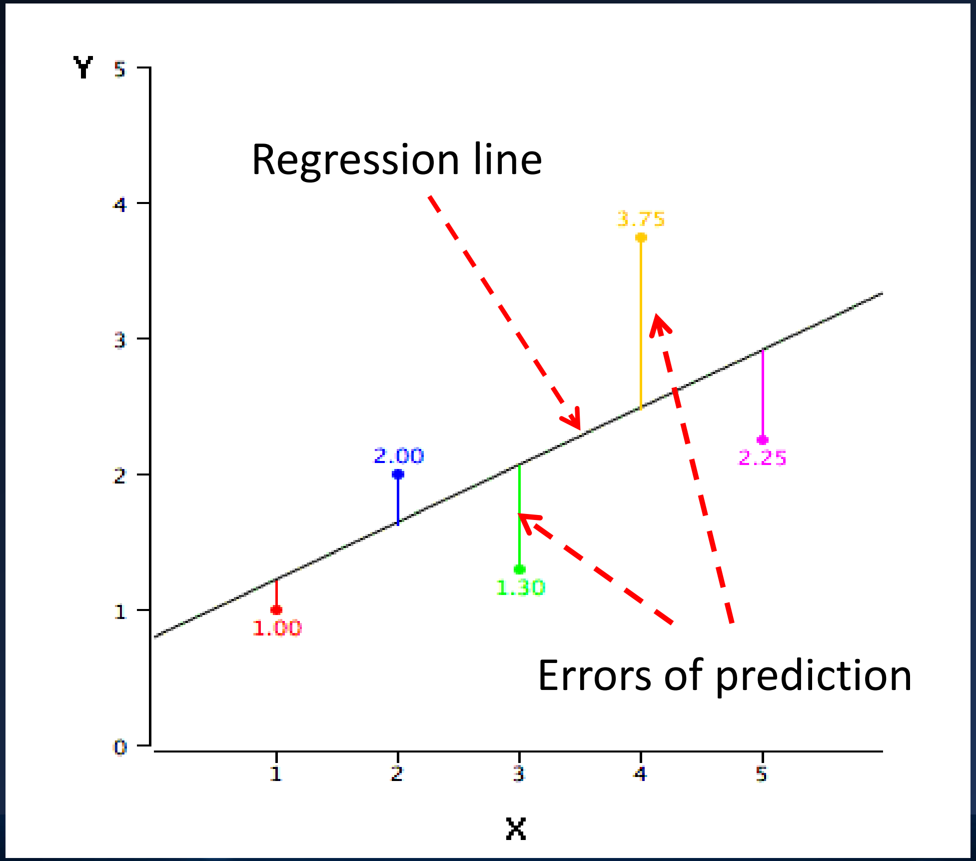
- Simple linear regression concerns the study of only one predictor variable.

- Multiple linear regression concerns the study of two or more predictor variables.

Simple linear regression

X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25





- Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

- Criterion for the best-fitting line:

the line that minimizes the sum of the squared errors of prediction – least squares criterion.

Simple linear regression

- Computing the Regression Line
  - y_i denotes the observed response for experimental unit *i*
  - x_i denotes the predictor value for experimental unit *i*
  - $\tilde{y}_i$ is the predicted response (or fitted value) for experimental unit *i*
  - Then, the equation for the best fitting line is:

$$\tilde{y}_i = b_0 + b_1 x_i$$

- The prediction error is:

$$e_i = y_i - \tilde{y}_i$$

Simple linear regression

- Least squares criterion – find the vales of b_0 and b_1 that minimize:

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Take the derivative with respect to b_0 and b_1, set to 0, and solve for b_0 and b_1, the "**least squares estimates**" for b_0 and b_1 are:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Multiple Linear Regression

- The model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i.$$

- We assume that the ϵ_i have a normal distribution with mean 0 and constant variance σ^2. These are the same assumptions that we used in simple regression with one *x*-variable.

Multiple Linear Regression

- The word "linear" in "multiple linear regression" refers to the fact that the model is *linear in the parameters*, $\beta_0, \beta_1, \ldots, \beta_{p-1}$.

- The estimates of the β coefficients are the values that minimize the sum of squared errors for the sample.

Summary

- Decision Tree
- Linear Regression