

DATA SCIENCE BOOTCAMP

MODULE 2

Acknowledgement

- The Regression example of the slides in this presentation are taken from the online lecture notes posted by Prof. H. D. Vinod of Fordham University
- Han and Kimber (Data Mining Concepts and Techniques)
- Tan, Steinbach and Kumar (Introduction to Data Mining)

Module Outline

- Brief overview of Machine Learning
- Type of Machine Learning
- Supervised Learning
 - Classification (Decision Tree Algorithm)
 - Regression (Linear Regression)
- Unsupervised Learning
 - K-Means Algorithm
- Introduction To Azure ML Studio

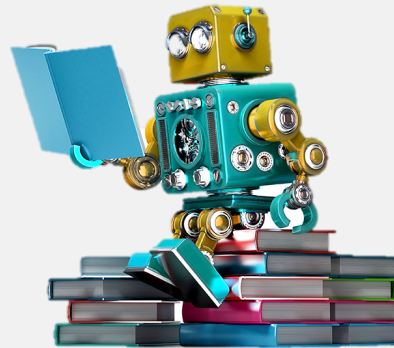
What Is Machine Learning?



Microsoft



- Machine learning is the science of getting computers to act without being explicitly programmed.



Statistics vs. Machine Learning



Microsoft



- Data science has its origins in various disciplines, of which the two most important are statistics and machine learning.
- Statistics has its roots in mathematics, and therefore, there has been an emphasis on mathematical rigor, a desire to establish that something is sensible on theoretical grounds before testing it in practice.
- In contrast, the machine learning/Data Science community has its origin very much in computer practice.

Types Of Machine Learning



Microsoft



- Machine learning comes in many different flavors, depending on the algorithm and its objectives. You can divide machine learning algorithms into three main groups based on their purpose:
- Supervised learning
- Unsupervised learning
- Reinforcement learning

Supervised Learning

- Supervised learning occurs when an algorithm learns from example data and associated target responses that can consist of numeric values or string labels, such as classes or tags, in order to later predict the correct response when posed with new examples.

Unsupervised Learning

- Unsupervised learning occurs when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own.

Reinforcement Learning

- Reinforcement learning occurs when you present the algorithm with examples that lack labels, as in unsupervised learning. However, you can accompany an example with positive or negative feedback according to the solution the algorithm proposes.

Classification (Overview)



Microsoft



- Supervised Learning
- Provided we have a collection of records/observations (Training Set)
- Each historic observation/record has a **Label** which answers our ML problem
- Find a model for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.

Classification (Cont'd)

- Common Algorithms
 - Logistic Regression (with L1 and L2 Regularization)
 - Decision Trees/ Classification Trees
 - Support Vector Machines
 - Random Forests
 - Neural Networks
- You don't need to program these.

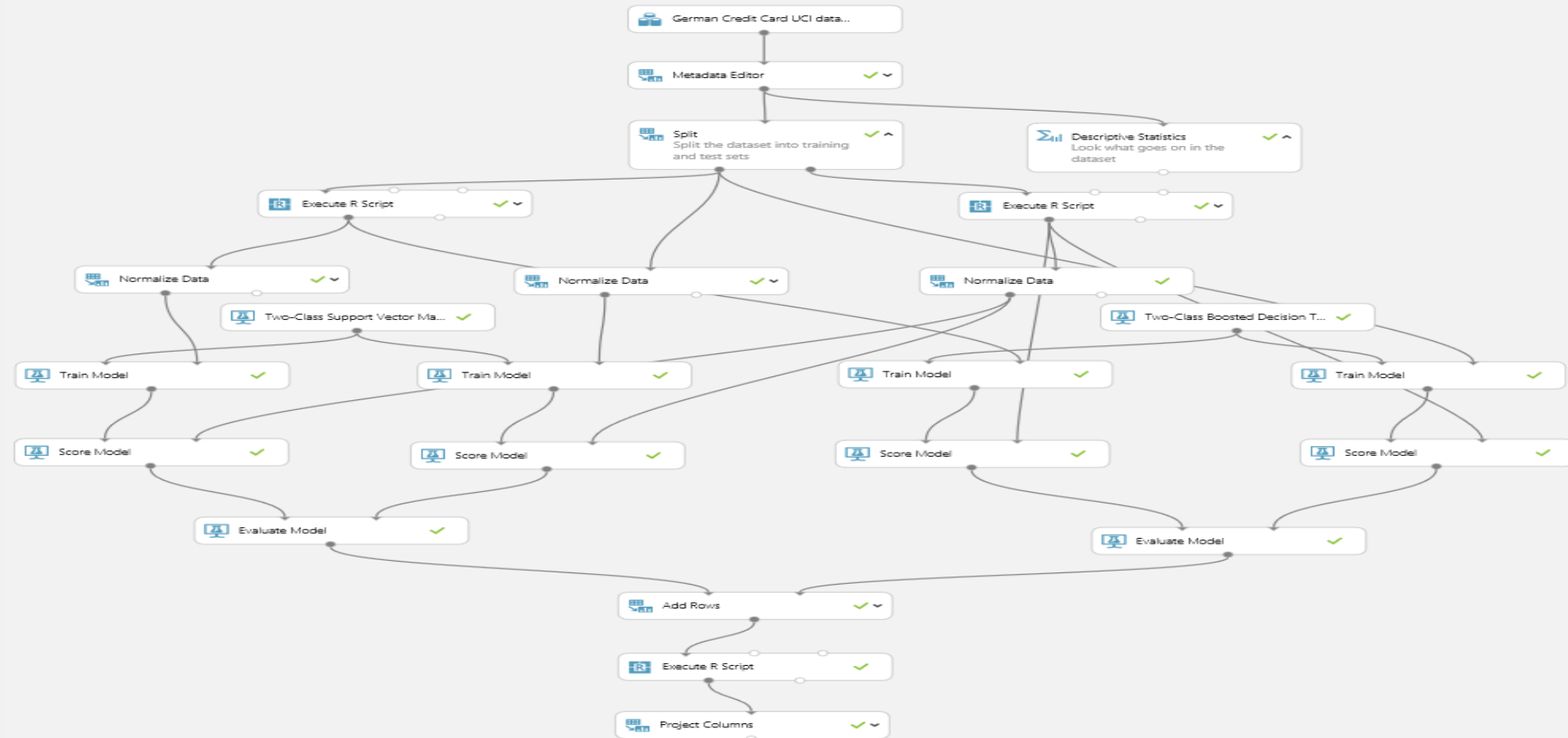
Classification (Cont'd)

- **Use Case**
- Binary Classification: Credit Risk Prediction
- **Cost-sensitive binary classification** to predict credit risk based on the information given on a credit application. The classification problem in this experiment is a cost-sensitive one because the cost of misclassifying the positive samples is five times the cost of misclassifying the negative samples.

Classification (Sneak Peak)



Microsoft



Classification (Azure ML Studio)



Microsoft



- Log on to <https://studio.azureml.net/>
- More coming in later modules

Classification (Decision Tree)

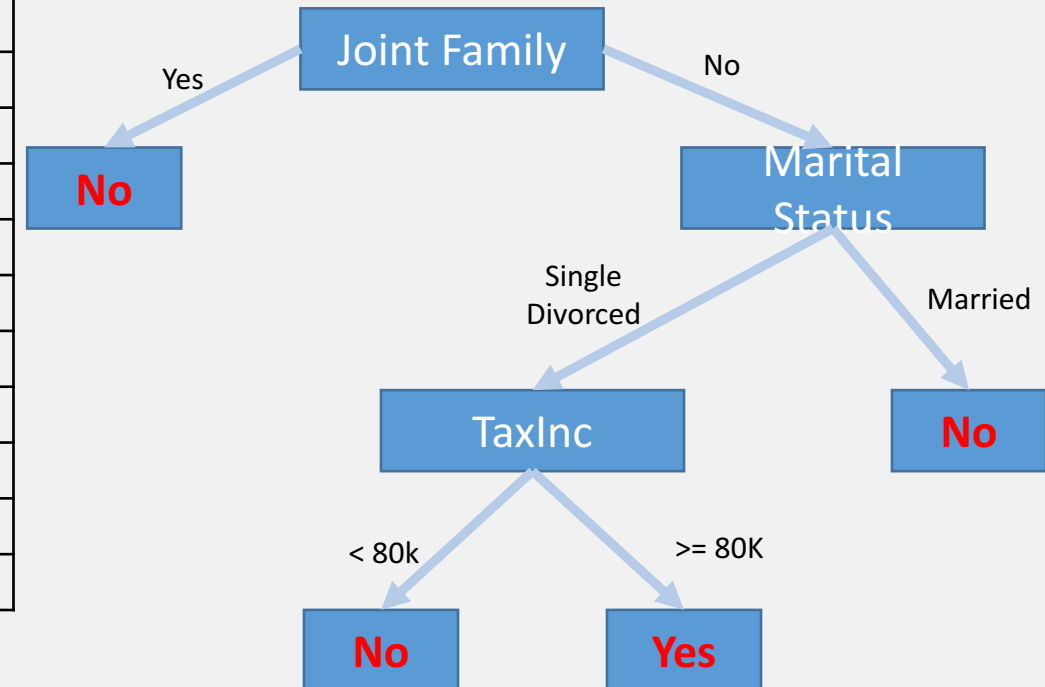


Microsoft



	Categorical	Categorical	Continuous	Class
id	Joint Family	Marital Status	Taxable Income	Buy Car
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

Classification (Decision Tree Cont'd)



Microsoft

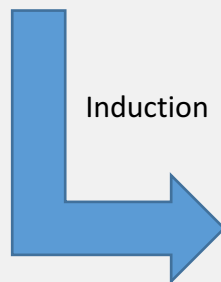


	Categorical	Categorical	Continuous	Class
id	Joint Family	Marital Status	Taxable Income	Buy Car
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data

Id	Joint Family	Marital Status	Taxable Income	Buy Car
1	Yes	Single	99K	?
2	No	Divorced	70K	?
3	Yes	Single	84K	?
4	Yes	Married	125K	?
5	No	Married	20K	?

Testing Data

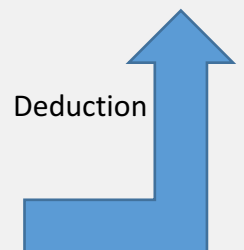


Induction

Learn Model

Model
Decision Tree

Apply Model



Deduction

Classification (Decision Tree)

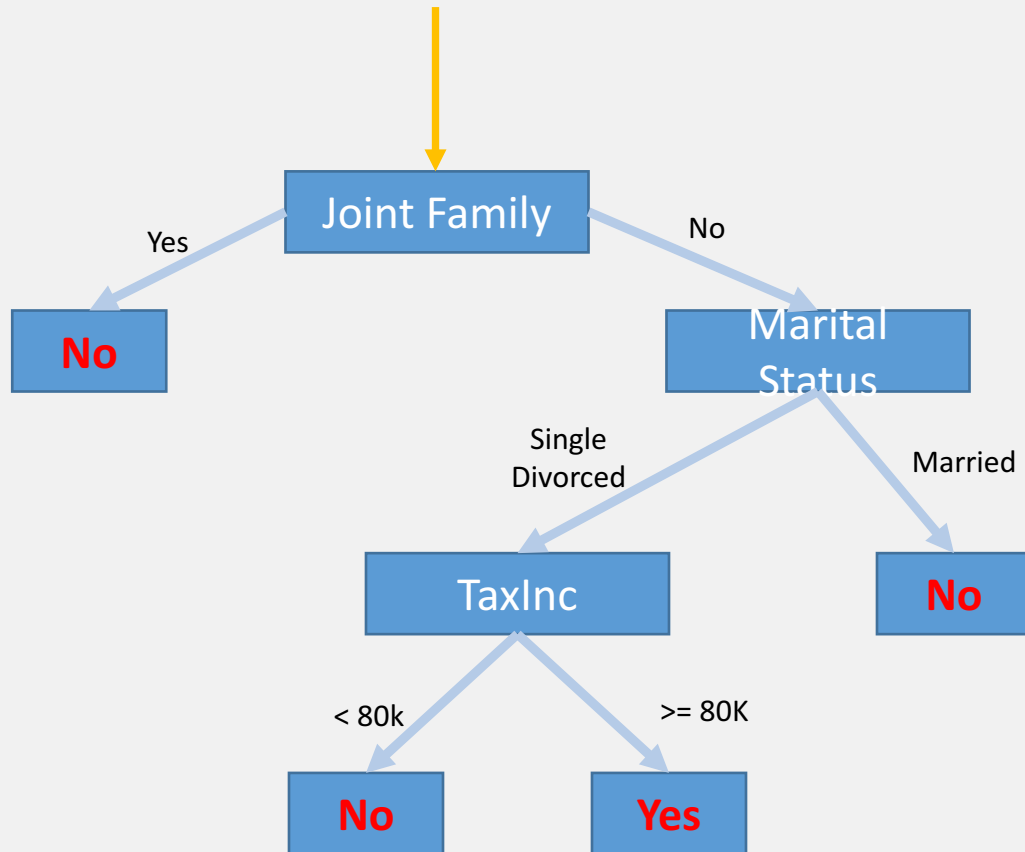
Apply Model To Test Data



Microsoft



Start From the root of the tree



Model: Decision Tree

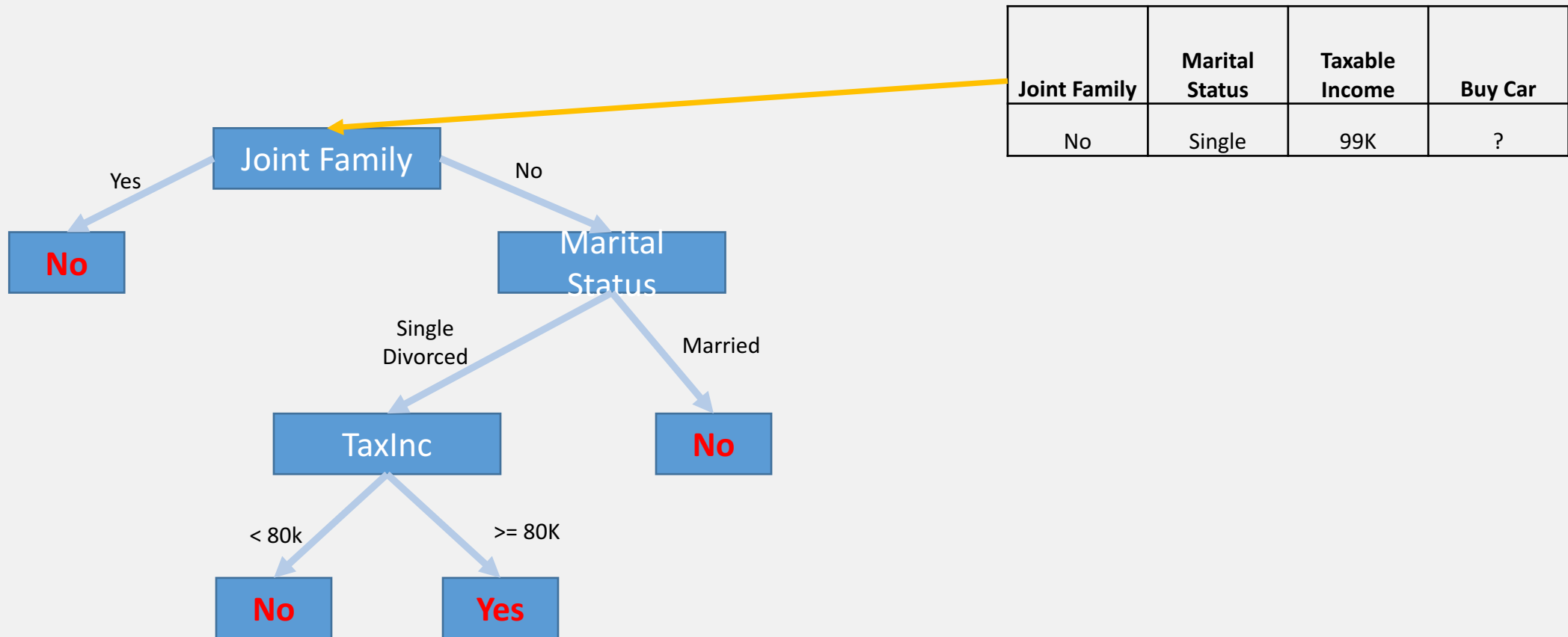
Joint Family	Marital Status	Taxable Income	Buy Car
Yes	Single	99K	?

Classification (Decision Tree)

Apply Model To Test Data



Microsoft



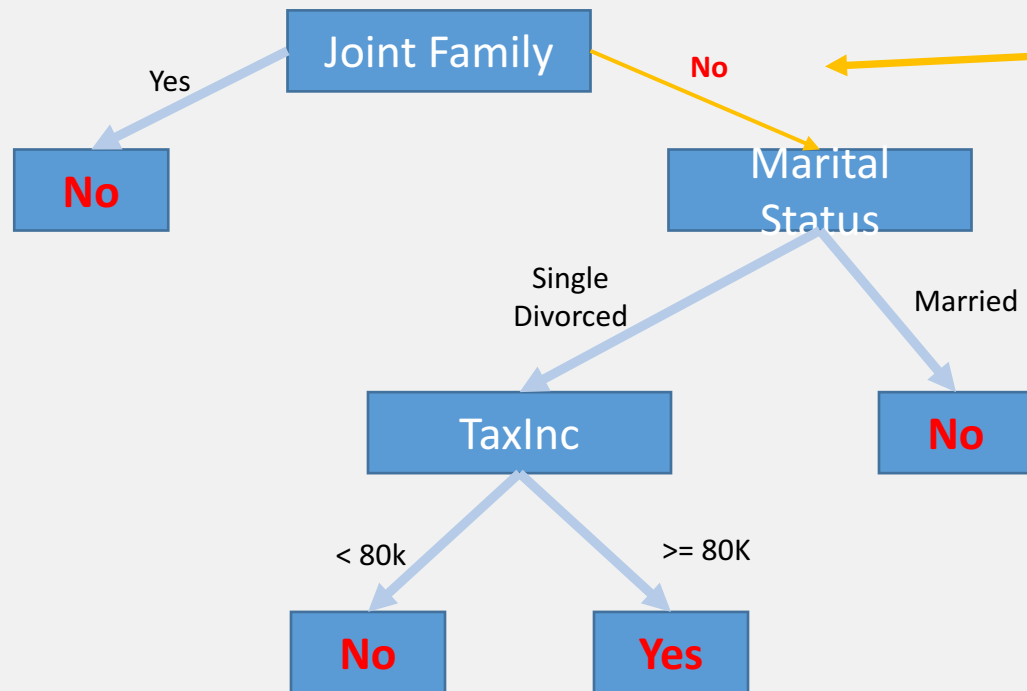
Model: Decision Tree

Classification (Decision Tree)

Apply Model To Test Data



Microsoft



Joint Family	Marital Status	Taxable Income	Buy Car
No	Single	99K	?

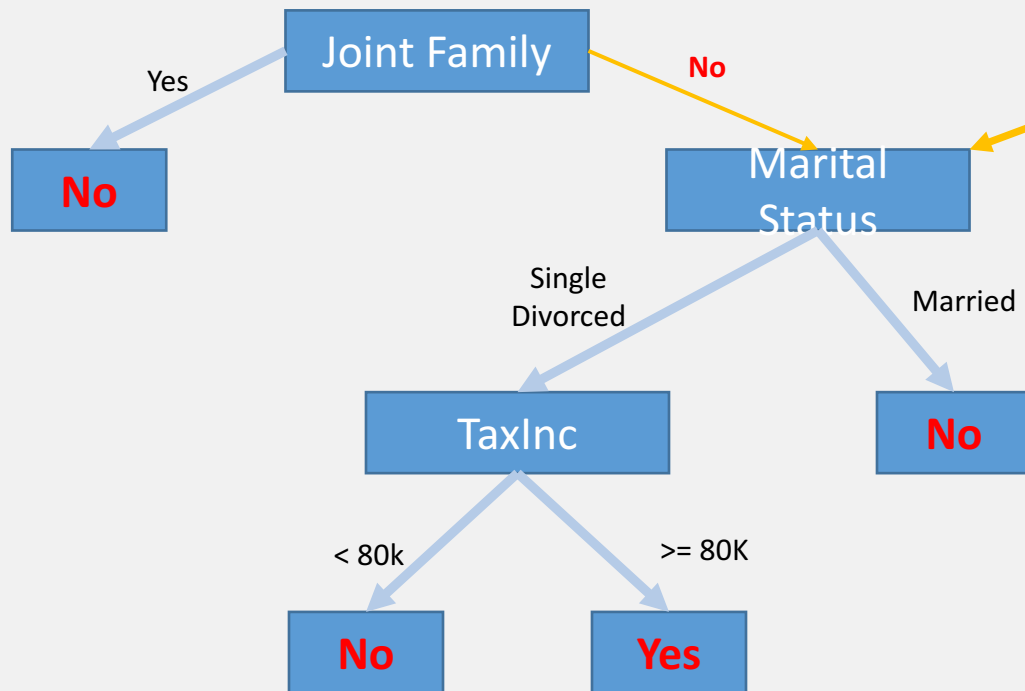
Model: Decision Tree

Classification (Decision Tree)

Apply Model To Test Data



Microsoft



Joint Family	Marital Status	Taxable Income	Buy Car
No	Single	99K	?

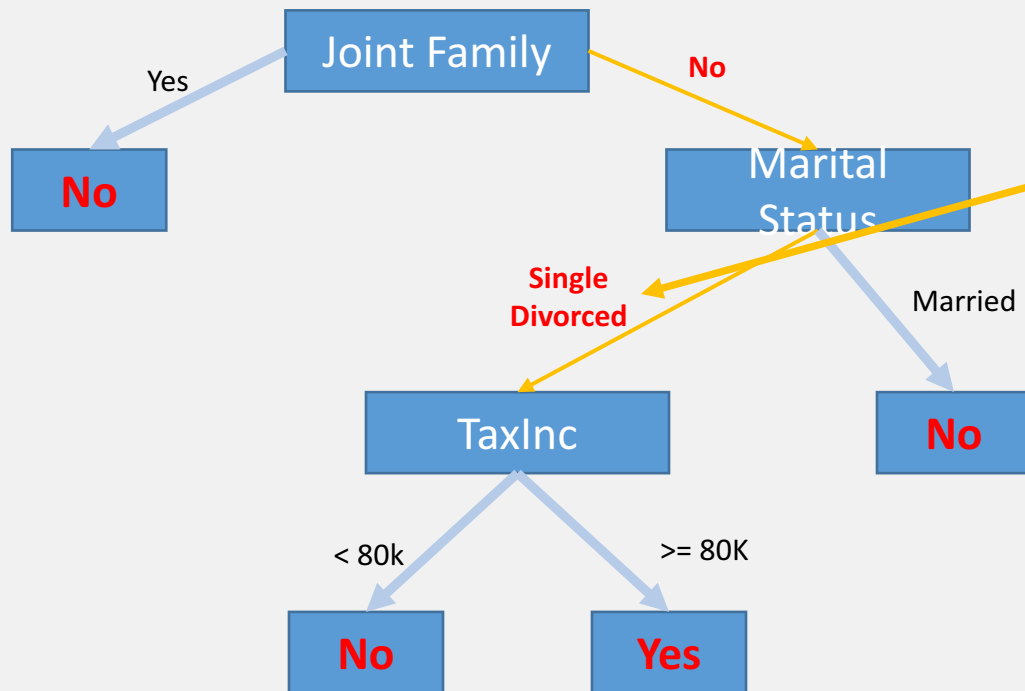
Model: Decision Tree

Classification (Decision Tree)

Apply Model To Test Data



Microsoft



Joint Family	Marital Status	Taxable Income	Buy Car
No	Single	99K	?

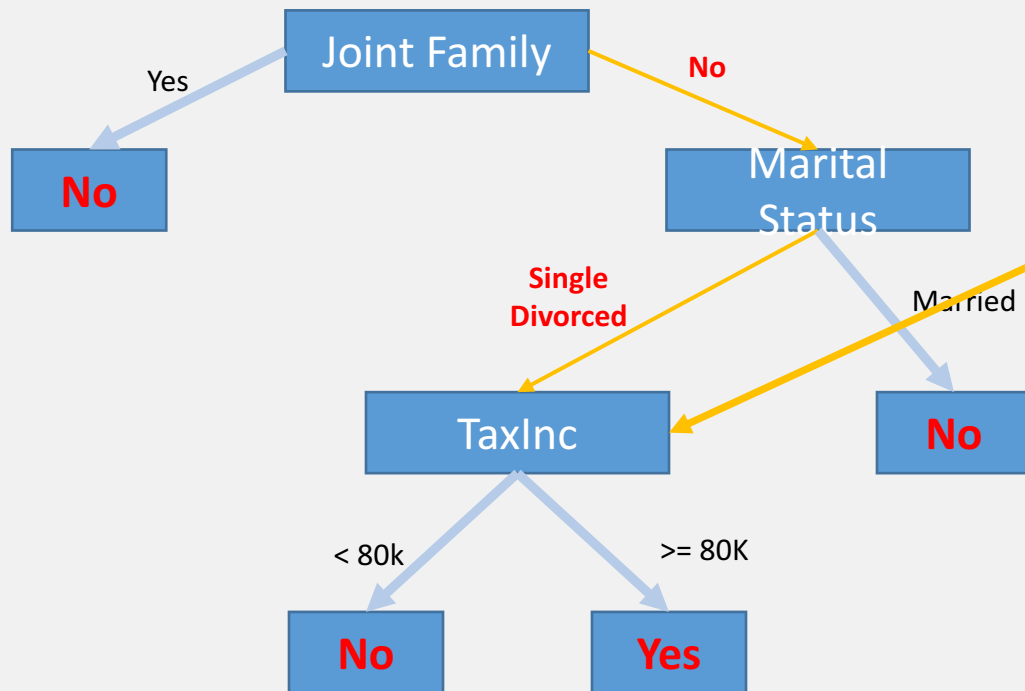
Model: Decision Tree

Classification (Decision Tree)

Apply Model To Test Data



Microsoft



Joint Family	Marital Status	Taxable Income	Buy Car
No	Single	99K	?

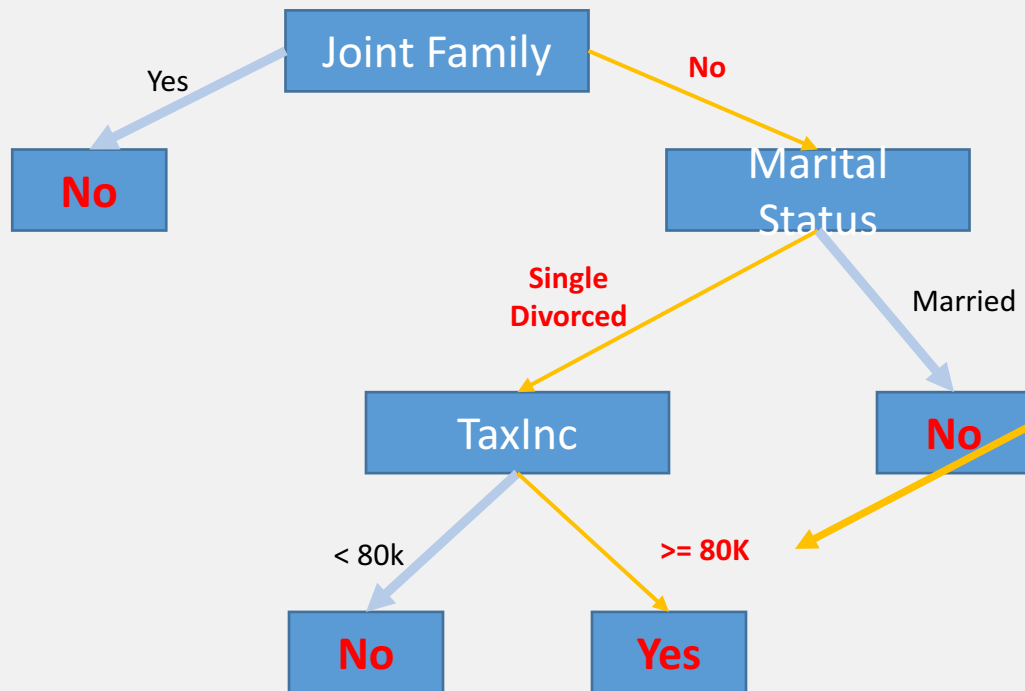
Model: Decision Tree

Classification (Decision Tree)

Apply Model To Test Data



Microsoft



Joint Family	Marital Status	Taxable Income	Buy Car
No	Single	99K	?

Model: Decision Tree

Decision Tree

Determining The Best Split

- The best split is determined by finding the node impurity.
 - Gini Index
 - Entropy

Decision Tree

Measure of Impurity: GINI

- GINI Index for a given node:

$$\text{GINI}(t) = 1 - \sum [p(j|t)]^2$$

- Maximum $(1 - 1/n_c)$ when records are equally distributed among all classes, implying least interesting information
- Minimum (0.0) when all records belong to one class, implying most interesting information

Decision Tree

Examples for computing GINI



Microsoft



C1	0
C2	6

$$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$$
$$\text{Gini} = 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$$

C1	1
C2	5

$$P(C1) = 1/6 \quad P(C2) = 5/6$$
$$\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$$

C1	2
C2	4

$$P(C1) = 2/6 \quad P(C2) = 4/6$$
$$\text{Gini} = 1 - (2/6)^2 - (4/6)^2 = 0.444$$

Decision Tree

Classification: Motivation



Microsoft

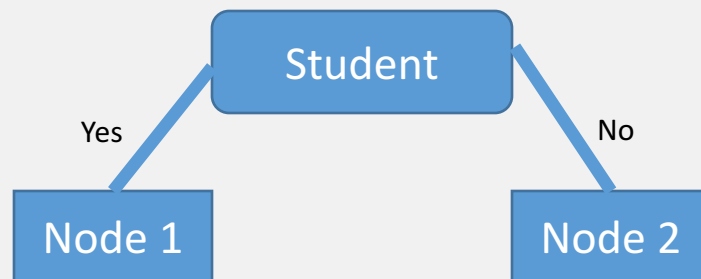


age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
>40	medium	no	excellent	no
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes

Decision Tree

Computing GINI Index

- Computing the overall GINI of a feature (e.g. Student)



Gini(N1)

$$= 1 - (6/7)^2 - (1/7)^2 = \mathbf{0.24}$$

Gini(N2)

$$= 1 - (3/7)^2 - (4/7)^2 = \mathbf{0.49}$$

	Buy Computer		
student	Yes	No	
Yes	6	1	7
No	3	4	7
			14

Gini(Student)

$$= (7/14 * 0.24) + (7/14 * 0.49)$$

= ??

Decision Tree

Numeric Attributes: Computing GINI Index



Microsoft



- For efficient computation: for each attribute, sort the attribute on values.
- Linearly scan these values, each time updating the count matrix and computing GINI index
- Choose the split position that has the least GINI index

		No		No		No		Yes		Yes		Yes		No		No		No		No			
		Taxable Income																					
Sorted Values →		60		70		75		85		90		95		100		120		125		220			
Split Positions →		55		65		72		80		87		92		97		110		122		172		230	
		<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes		0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No		0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini		0.420		0.400		0.375		0.343		0.417		0.400		0.300		0.343		0.375		0.400		0.420	

Decision Tree

Inducing a Decision Tree

- There are many possible trees
- How to find the most compact one, that is consistent with the data?
- The key to building a decision tree, which attribute to choose in order to branch.
- The heuristic is to choose the attribute with the minimum GINI/Entropy.

Decision Tree

Extracting Classification Rules from Trees

- Represent the knowledge in the form of IF-THEN rules
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction.
- The leaf node holds the class prediction
- Rules are easier for humans to understand
- Example
 - IF age = " ≤ 30 " AND student = "no" THEN buys_computer = "no"
 - IF age = " ≤ 30 " AND student = "yes" THEN buys_computer = "yes"
 - IF age = "31...40" THEN buys_computer = "yes"
 - IF age = " > 40 " AND credit_rating = "excellent" THEN buys_computer = "yes"
 - IF age = " ≤ 30 " AND credit_rating = "fair" THEN buys_computer = "no"

Regression

Introduction to Regression Analysis



Microsoft



- Regression analysis is used to:
 - Predict the value of a dependent variable based on the value of at least one independent variable
 - Explain the impact of changes in an independent variable on the dependent variable
- Dependent variable: the variable we wish to explain
- Independent variable: the variable used to explain the dependent variable

Regression

Regression vs. Classification

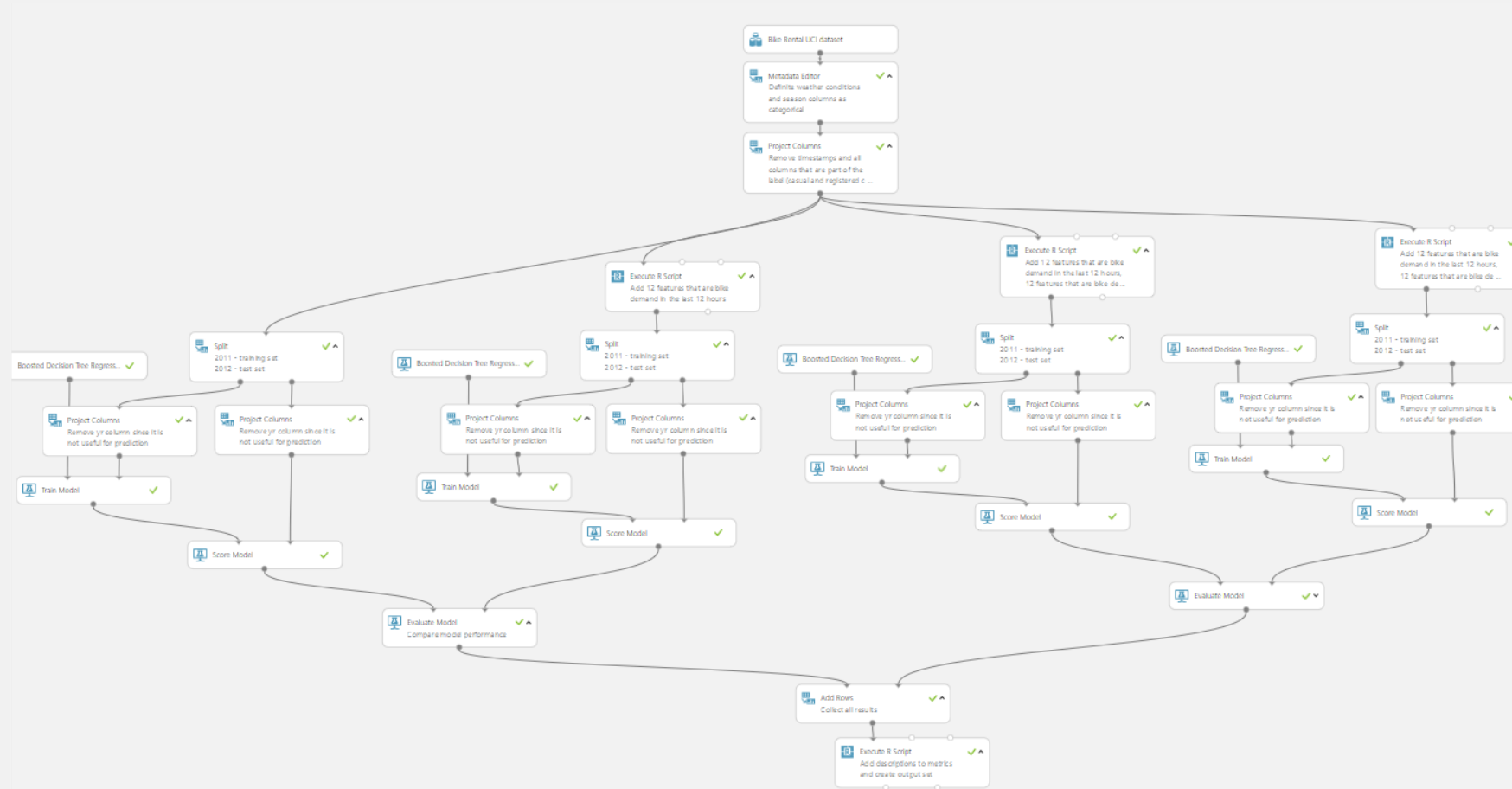
- Regression
 - Output variable takes continuous values
 - Example:—
 - To estimate median housing prices in a neighborhood
 - To estimate a student's CGPA
- Classification
 - Output variable takes class labels
 - Example:—
 - To predict loan default
 - To predict churn
 - To predict if a student will accept the offer letter

- **Use Case:**
- Demand estimation
- The Bike Rental UCI dataset is used as the input raw data for this experiment. This dataset is based on real data from the Capital Bikeshare company, which operates a bike rental network in Washington DC in the United States.
- The experiment demonstrates demand estimation using regression with UCI bike rental data.

Regression (Sneak Peak)



Microsoft



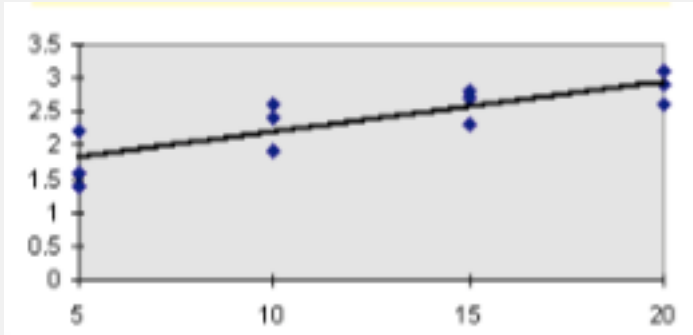


- Log on to <https://studio.azureml.net/>
- More info:
<https://gallery.cortanaintelligence.com/Experiment/Regression-Demand-estimation-4>

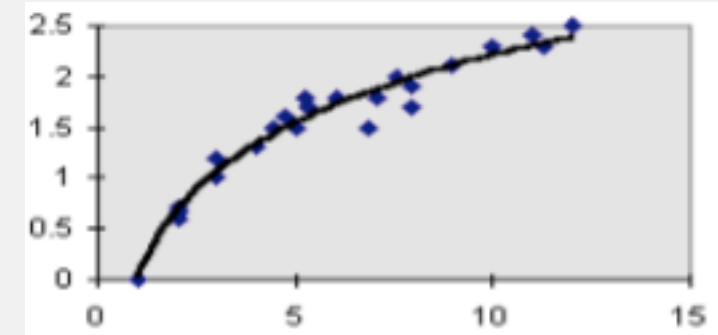
Regression

Types of Regression Models

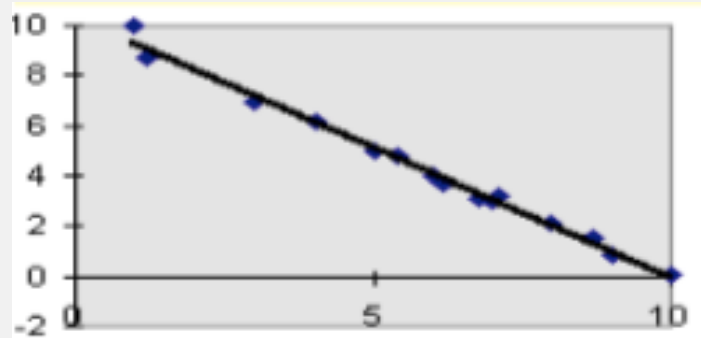
Positive Linear Relationship



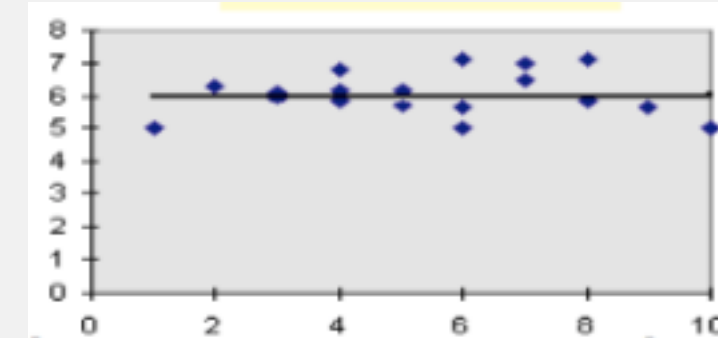
Relationship not Linear



Negative Linear Relationship



No Relationship



Regression

Scatter Plots and Correlation

- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
 - Only concerned with strength of the relationship
 - No causal effect is implied

Regression

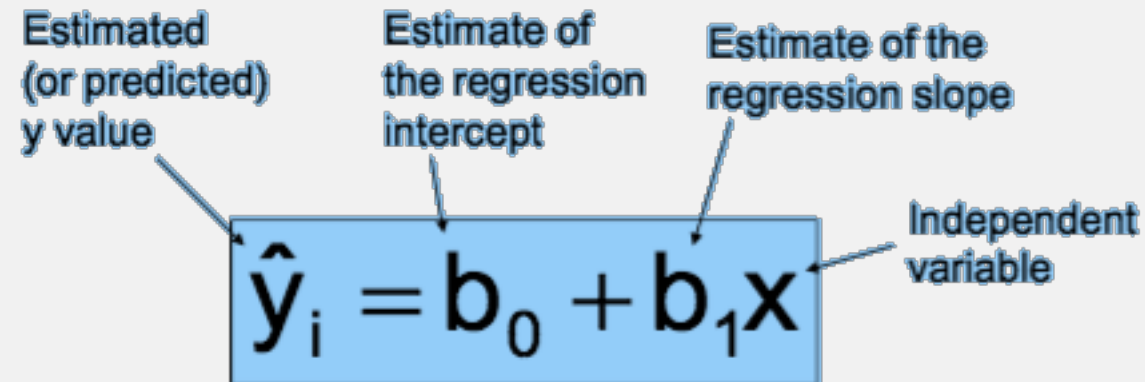
Estimating Regression Model



Microsoft



- The sample regression line provides an estimate of the population regression line



The diagram shows the regression equation $\hat{y}_i = b_0 + b_1x$ inside a light blue box. Four labels with arrows point to different parts of the equation: 'Estimated (or predicted) y value' points to \hat{y}_i , 'Estimate of the regression intercept' points to b_0 , 'Estimate of the regression slope' points to b_1 , and 'Independent variable' points to x .

$$\hat{y}_i = b_0 + b_1x$$

The individual random error terms e_i have a mean of zero

Regression

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

Regression

The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

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

Regression

Evaluation Techniques

- Coefficient of Determination (R^2)
- Residual Analysis
 - Histogram of residuals
 - Scatter plot of residual vs. target variable
- Scatter plot of actual vs. estimated values from hold-out data

Regression

Coefficient of Determination, R^2



Microsoft



- Coefficient of determination

$$R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}$$

Un-Supervised Clustering

(K-Mean)

- **RECAP:**
- Unsupervised learning occurs when an algorithm learns from plain examples without any associated response, leaving to the algorithm to determine the data patterns on its own.

Cluster Analysis

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups
- Cluster: a collection of data objects
 - Similar to one another within the same cluster – Dissimilar to the objects in other clusters
- Cluster analysis :
 - Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters

Examples of Clustering Applications

- **Marketing:** Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
- **Land use:** Identification of areas of similar land use in an earth observation database
- **Insurance:** Identifying groups of motor insurance policy holders with a high average claim cost
- **City-planning:** Identifying groups of houses according to their house type, value, and geographical location
- **Earth-quake studies:** Observed earth quake epicenters should be clustered along continent faults

What Is Good Clustering?

- A good clustering method will produce high quality clusters with
 - high intra-class similarity
 - low inter-class similarity

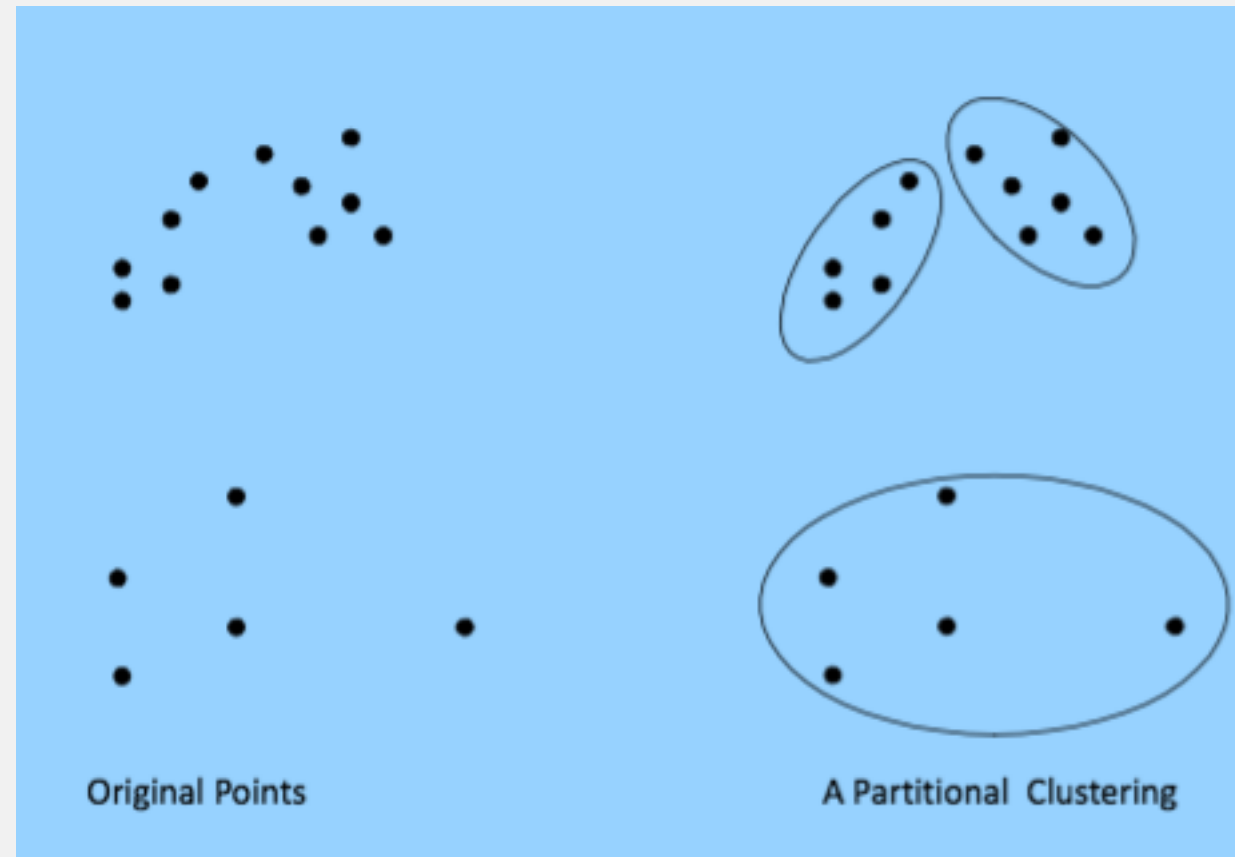
Types of Clustering

- Partition Clustering
 - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- Hierarchical clustering
 - A set of nested clusters organized as a hierarchical tree

Partition Clustering



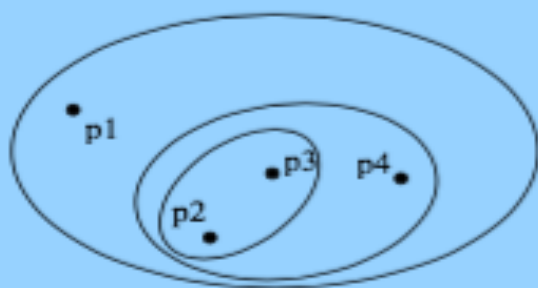
Microsoft



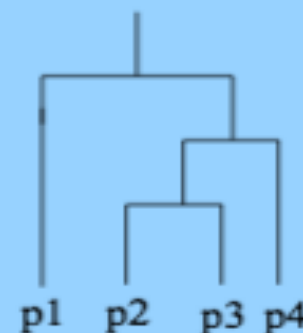
Hierarchical Clustering



Microsoft



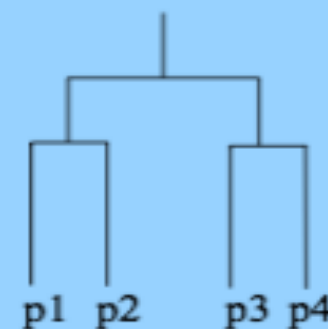
Traditional Hierarchical Clustering



Traditional Dendrogram



Non-traditional Hierarchical Clustering



Non-traditional Dendrogram

The K-Means Clustering Method



Microsoft



1. Choose a value for K , the total number of clusters to be determined.
2. Choose K instances within the dataset at random. These are the initial cluster centers.
3. Use simple Euclidean distance to assign the remaining instances to their closest cluster center.
4. Use the instance in each cluster to calculate a new mean for each cluster.
5. If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centers and repeat steps 3-5.

Working of the K-Mean Algorithm



Microsoft



Instance #:	1	2	3	4	5	6
X:	1	1	2	2	3	3
Y:	1.5	4.5	1.5	3.5	2.5	6.0

- Let's pick Instances #1 and #3 as the initial centroids.

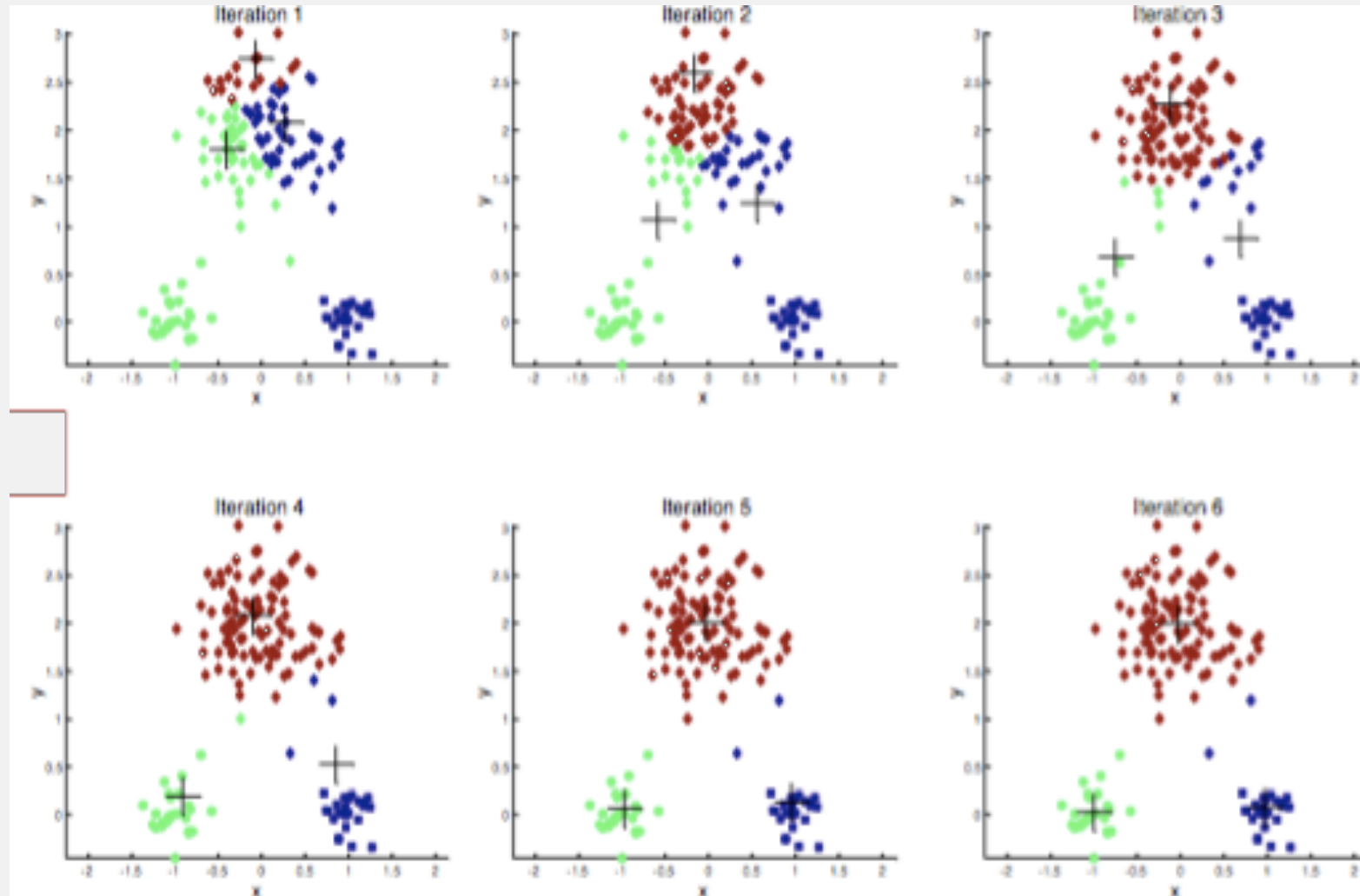
	Distance with Centroid1	Distance with Centroid2
Instance #2	3	3.16
Instance #4	2.24	2
Instance #5	2.24	1.41
Instance #6	6.02	5.41

- New centroids are (1, 3) and (2.5, 3.4)

Graphical View K-Mean



Microsoft



Graphical View K-Mean (cont'd)

- <https://www.youtube.com/watch?v=5l3Ei69l40s>

END OF MODULE 2
