Data Mining :: Unit-3

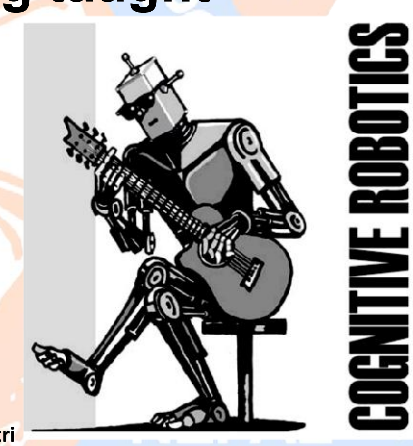
(Classification – Basics and Background)

Er. Dinesh Baniya Kshatri
(Lecturer)

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

Machine Learning

- **To learn = To acquire knowledge via self-study, experience or by being taught**
- **Basic categories of learning**
  - Supervised
  - Unsupervised
  - Reinforcement



Supervised Learning

- Train machines using data which is well labeled i.e. data is already tagged with the correct answer
- Construction of a proper training, validation and test set is crucial
- New data (Test data) is evaluated based on training set
- Examples:
  - **Classification**: Output variable is a category, such as "Red" or "Blue" or "Disease" and "No Disease"
  - **Regression**: Output variable is a real value, such as "dollars" or "weight"

Unsupervised Learning

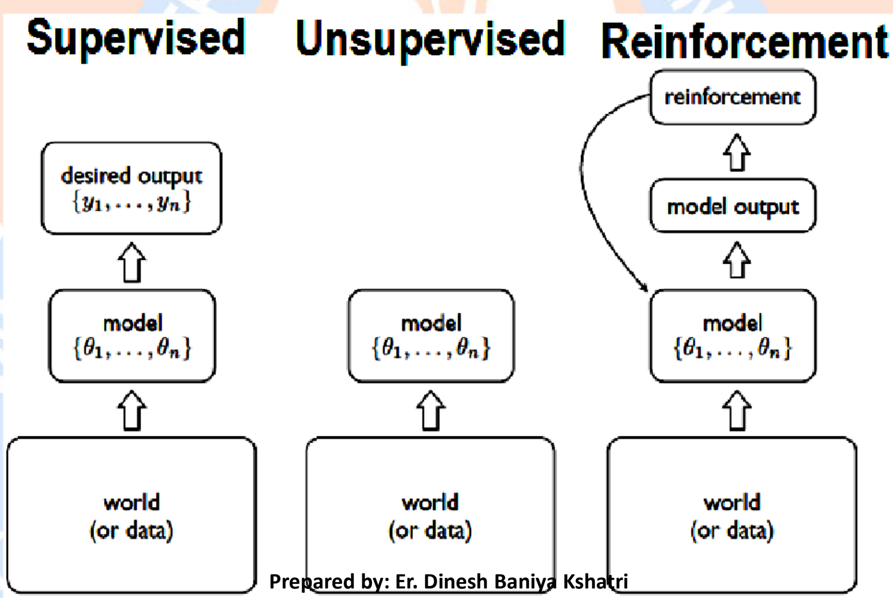
- Train machines using information that is neither classified nor labeled
- It allows the algorithm to act on the information without guidance
- Groups unsorted information according to similarities, patterns and differences without any prior training
- Examples:
  - **Clustering**: Discover inherent groupings in the data, such as grouping customers by purchasing behavior
  - **Association**: Discover rules that describe large portions of data, such as people that buy X also tend to buy Y

Reinforcement Learning

- **Machine learns by obtaining either rewards or penalties for the actions it performs**
  - The goal of the machine is to maximize the total reward

- **The designer sets the reward policy, however the model receives no hints or suggestions to solve a particular task**
  - It is up to the model to figure out how to perform a task to maximize the reward
  - The machine starts from totally random trials and finishes with sophisticated tactics and superhuman skills

Types of Learning



Introduction to Classification – 1

- **Questions:**
  - What is shown in the figure?
  - Why do you know?
  - How have you gained that knowledge?

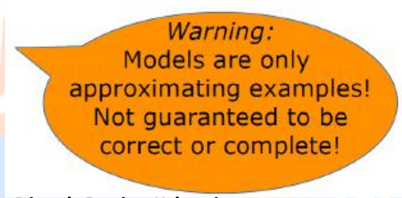
Introduction to Classification – 2

- **Train a model for recognizing a concept such as trees**
  - Requires training data

Introduction to Classification – 3

- **Learning algorithm observes both positive and negative examples from training data**
  - A classifier model is derived
    - Example: “Trees are big green plants that have a trunk”
  - What happens during classification of unseen instances?

Tree?



Agenda and Approach of Classification

Goal: Previously unseen records should be assigned a class from a given set of classes as accurately as possible.

- **Approach:**
  - Given a collection of records (*training set)*
  - Each record contains a set of *attributes*
  - One of the attributes is the *class (label) that should be predicted*
  - Learn a *model for the class attribute as a function of the* values of other attributes

Example – Classification Data

Example: Data Table with class attribute C

Rec	a1	a2	a3	a4	C
o1	1	1	m	g	c1
o2	0	1	v	g	c2
o3	1	0	m	b	c1

This data consists of tuples (examples, instances):

o1= (1, 1, m, g) with the **class label** c1

o2= (0, 1, v, g) with the **class label** c2

o3 =(1, 0, m, b) with the **class label** c1

More Classification Examples

- Credit Risk Assessment
  - Attributes: your age, income, debts, ...
  - Class: Are you getting credit by your bank?
- Marketing
  - Attributes: previously bought products, browsing behaviour
  - Class: Are you a target customer for a new product?
- Tax Fraud
  - Attributes: the values in your tax declaration
  - Class: Are you trying to cheat?
- SPAM Detection
  - Attributes: words and header fields of an e-mail
  - Class: Is it a spam e-mail?

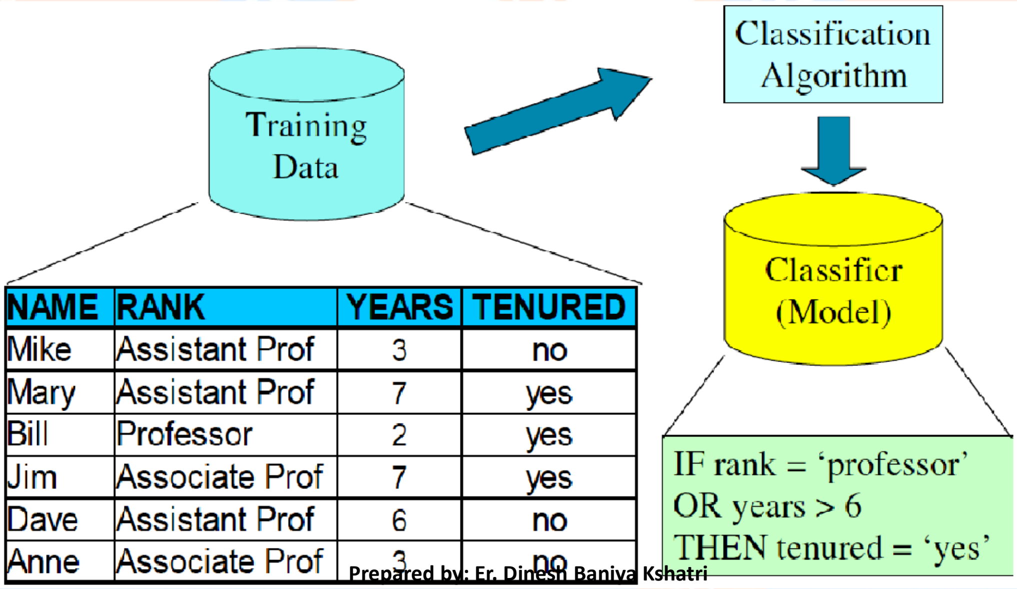
Datasets for Classification Problems

- **Training Dataset**
  – Includes data used for learning where the target value is known
- **Validation Dataset**
  – Portion of data from training dataset that is withheld
  – Used to tune the architecture / parameters of a classifier and get a rough estimate of error
- **Test Dataset**
  – Used only to asses the performance of a classifier
  – Is not used during the training and validation process
  – Is used to give an unbiased estimate of the error in the final tuned model

Classification
(A Two Step Process)

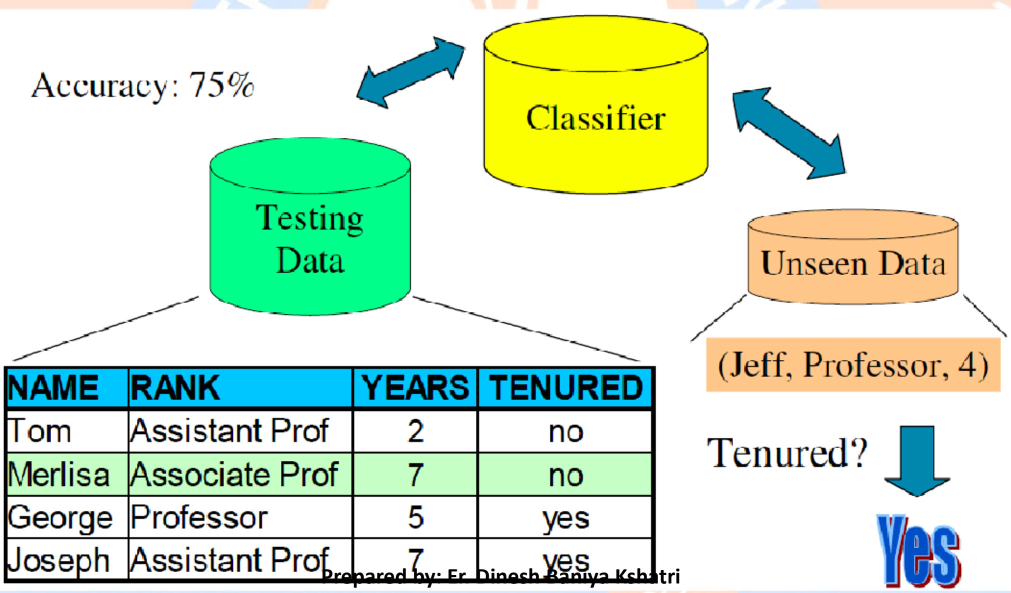
- **Model Construction**
  – Describes a set of predetermined classes
  – Each tuple, sample, record is assumed to belong to a predefined class, as determined by the class label attribute
- **Model Usage**
  – Used for classifying future or unknown objects
  – Accuracy rate is the percentage of test set samples that are correctly classified by the model
  – If the accuracy is acceptable, use the model to classify data objects whose class labels are not known

Classification (Model Construction)



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Classification (Model Usage)



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

Classification vs. Prediction

- Classification
  - Predicts categorical class labels (discrete or nominal)
  - Classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction
  - Models continuous-valued functions, (predicts unknown or missing values)

Evaluating Classification Methods

- **High Accuracy**
  - Classifier accuracy: Predicting class label
  - Predictor accuracy: Guessing value of predicted attributes
- **Speed and Complexity**
  - Time to construct the model (training time)
  - Time to use the model (classification time)
- **Scalability –** Ability to adapt to changing data size
- **Robustness –** Handling noise and outliers
- **Interpretability –** Providing useful insights

What is classification?

	categorical	categorical	continuous	class
Tid	Refund	Marital Status	Taxable Income	Cheat
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

- Classification is the task of learning a target function (f) that maps attribute set (x) to one of the predefined class labels (y)

One of the attributes is the class attribute
In this case: Cheat

Two class labels (or classes): Yes (1), No (0).

Why Classification?

- **Descriptive Modeling**
  - Create an explanatory tool to distinguish between objects of different classes
  - E.g. understand why people cheat on their taxes
- **Predictive Modeling**
  - Predict a class of a previously unseen record

Practical Classification Tasks

- Predicting tumor cells as benign or malignant
- Classifying credit card transactions as legitimate or fraudulent
- Categorizing news stories as finance, weather, entertainment, sports
- Identifying spam email, spam web pages
- Understanding if a web query has commercial intent or not

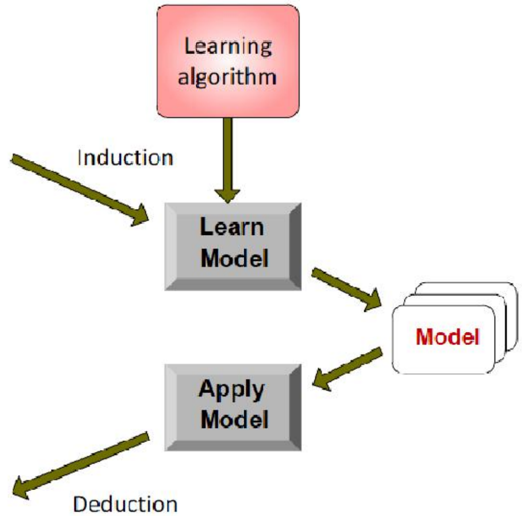
General Classification Approach

- Training set consists of records with known class labels
- Training set is used to build a classification model
- A labeled test set of previously unseen data records is used to evaluate the quality of the model.
- The classification model is applied to new records with unknown class labels

Illustrating Classification Process

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set



Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set

Example: Catching Tax Evasion

Tax-return data for year 2011

Tid	Refund	Marital Status	Taxable Income	Cheat
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

- Learn a method for discriminating between records of different classes (cheater vs. non-cheaters)

A new tax return for 2012
Is this a cheating tax return?

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

Evaluation of Classification Models

- Count the number of test records that are correctly (or incorrectly) predicted by the classification model

Confusion matrix

Actual Class	Predicted Class: Class = 1	Predicted Class: Class = 0
Class = 1	f_{11}	f_{10}
Class = 0	f_{01}	f_{00}

$$\text{Accuracy} = \frac{\#\text{ correct predictions}}{\text{total}\#\text{ of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

$$\text{Error rate} = \frac{\#\text{ wrong predictions}}{\text{total}\#\text{ of predictions}} = \frac{f_{10} + f_{01}}{f_{11} + f_{10} + f_{01} + f_{00}}$$

Classification Algorithms

- Decision Tree Classifier
- Rule Based Classifier
- Nearest Neighbor Classifier
- Naïve Bayes and Bayesian Belief Classifier
- Artificial Neural Network Classifier

Background Information
(Entropy)

- Entropy (Information Theory)
  - A measure of uncertainty associated with a random variable
  - Calculation: For a discrete random variable *Y* taking *m* distinct values $\{y_1, \ldots, y_m\}$,
    - $H(Y) = -\sum_{i=1}^{m} p_i \log(p_i)$, where $p_i = P(Y = y_i)$
  - Interpretation:
    - Higher entropy => higher uncertainty
    - Lower entropy => lower uncertainty

Background Information
(Example – Entropy Calculation)

Entropy = $\sum_i - p_i \log_2 p_i$

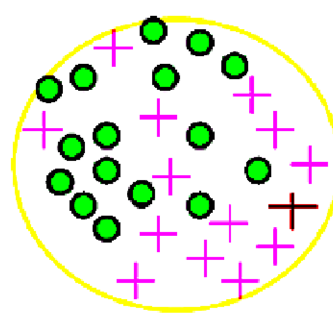
p_i is the probability of class i

Compute it as the proportion of class i in the set.

16/30 are green circles; 14/30 are pink crosses

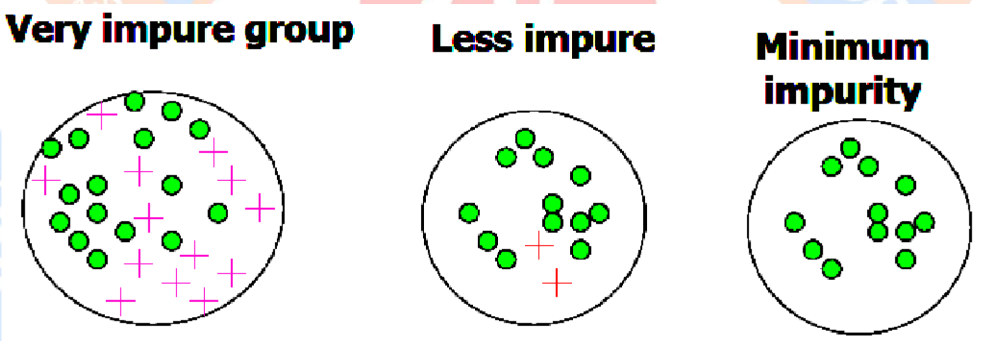
$\log_2(16/30)$ = -.9; $\log_2(14/30)$ = -1.1

Entropy = -(16/30)(-.9) –(14/30)(-1.1) = .99



Entropy & Impurity – 1

- **Entropy measures the level of impurity in a group**
- **Higher the entropy the more the information content**



Prepared by: Er. Dinesh Baniya Kshatri 29

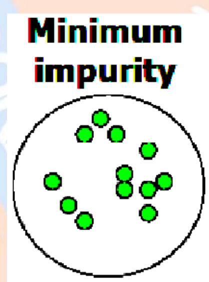
Entropy & Impurity – 2

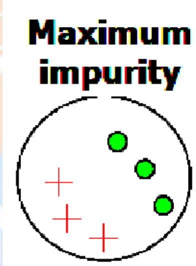
- What is the entropy of a group in which all examples belong to the same class?
  – entropy = $-1 \log_2 1 = 0$

  not a good training set for learning

- What is the entropy of a group with 50% in either class?
  – entropy = $-0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

  good training set for learning





Prepared by: Er. Dinesh Baniya Kshatri 30

Properties of Entropy

Maximized when elements are heterogeneous (impure):

$$\text{If } p_k = \frac{1}{k}, \text{then}$$

$$\text{Entropy} = H = -K \cdot \frac{1}{k} \log_2 \frac{1}{k} = \log_2 K$$

Minimized when elements are homogenous (pure):

$$\text{If } p_i = 1 \text{ or } p_i = 0, \text{then}$$

$$\text{Entropy} = H = 0$$

Information Gain (IG)

- We want to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.

- Information gain tells us how important a given attribute of the feature vectors is.

Information Gain (IG)

- **IG measures how much "information" a feature gives us about a class:**
  - Features that perfectly partition a dataset provide maximal information gain
  - Unrelated features give no information
  - It measures the reduction in entropy

Prepared by: Er. Dinesh Baniya Kshatri 33

Information Gain

With entropy defined as:

$$H = -\sum_{i=1}^{K} p_k \log_2 p_k$$

Then the change in entropy, or *Information Gain*, is defined as:

$$\Delta H = H - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

where m is the total number of instances, with m_k instances belonging to class k, where $K = 1, \ldots, k$.

Prepared by: Er. Dinesh Baniya Kshatri 34

Splitting Criterion

Suppose we want to split on the first variable (x_1):

x_1	1	2	3	4	5	6	7	8
y	0	0	0	1	1	1	1	1

If we split at $x_1 < 3.5$, we get an optimal split.
If we split at $x_1 < 4.5$, we make a mistake (misclassification).

Idea: *A better split should make the samples "pure" (homogeneous).*

How to determine the Best Split?

- Greedy approach:
  - Nodes with purer class distribution are preferred
- Need a measure of node impurity:

C0: 5 C1: 5	C0: 9 C1: 1
High degree of impurity	**Low degree of impurity**

Measures for Selecting the Best Split

Impurity measures include:

$$\text{Entropy} = -\sum_{i=1}^{K} p_k \log_2 p_k$$

$$\text{Gini} = 1 - \sum_{i=1}^{K} {p_k}^2$$

$$\text{Classification error} = 1 - \max_i p_k$$

where p_k denotes the proportion of instances belonging to class k ($K = 1, \ldots, k$), and $0 \log_2 0 = 0$.

Example – Measuring Node Impurity (Question)

C1	0
C2	6

C1	1
C2	5

C1	2
C2	4

- **For each of the cases given on the left calculate:**
  - Entropy
  - Gini
  - Classification Error

Example – Measuring Node Impurity (Answers)

$P(C1) = 0/6 = 0 \quad P(C2) = 6/6 = 1$

Gini $= 1 - P(C1)^2 - P(C2)^2 = 1 - 0 - 1 = 0$

Entropy $= - 0 \log 0 - 1 \log 1 = - 0 - 0 = 0$

Error $= 1 - \max(0, 1) = 1 - 1 = 0$

$P(C1) = 1/6 \quad P(C2) = 5/6$

Gini $= 1 - (1/6)^2 - (5/6)^2 = 0.278$

Entropy $= - (1/6) \log_2 (1/6) - (5/6) \log_2 (1/6) = 0.65$

Error $= 1 - \max(1/6, 5/6) = 1 - 5/6 = 1/6$

$P(C1) = 2/6 \quad P(C2) = 4/6$

Gini $= 1 - (2/6)^2 - (4/6)^2 = 0.444$

Entropy $= - (2/6) \log_2 (2/6) - (4/6) \log_2 (4/6) = 0.92$

Error $= 1 - \max(2/6, 4/6) = 1 - 4/6 = 1/3$

Simple Example – Information Gain

Training Set: 3 features and 2 classes

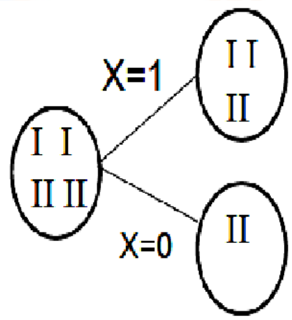
X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II

How would you distinguish class I from class II?

How should the training record be split?

Simple Example – Information Gain
(Split on Attribute X – Based on Entropy)

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II



$E_{child1} = -(1/3)\log_2(1/3) - (2/3)\log_2(2/3)$
$= .5284 + .39$
$= .9184$

$E_{child2} = 0$

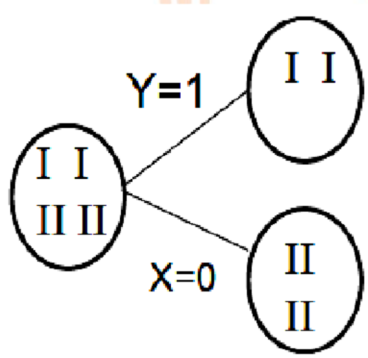
$E_{parent} = 1$

GAIN $= 1 - (3/4)(.9184) - (1/4)(0) = .3112$

Prepared by: Er. Dinesh Baniya Kshatri 41

Simple Example – Information Gain
(Split on Attribute Y – Based on Entropy)

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II



$E_{child1} = 0$

$E_{child2} = 0$

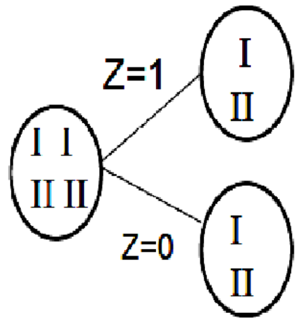
$E_{parent} = 1$

GAIN $= 1 - (1/2)\,0 - (1/2)0 = 1$; BEST ONE

Prepared by: Er. Dinesh Baniya Kshatri 42

Simple Example – Information Gain
(Split on Attribute Z – Based on Entropy)

X	Y	Z	C
1	1	1	I
1	1	0	I
0	0	1	II
1	0	0	II



$E_{child1} = 1$

$E_{child2} = 1$

$E_{parent} = 1$

GAIN = 1 – (1/2)(1) – (1/2)(1) = 0 ie. NO GAIN; WORST

Complex Example – Information Gain [1]

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	*No*
Sunny	Hot	High	True	*No*
Overcast	Hot	High	False	*Yes*
Rainy	Mild	High	False	*Yes*
Rainy	Cool	Normal	False	*Yes*
Rainy	Cool	Normal	True	*No*
Overcast	Cool	Normal	True	*Yes*
Sunny	Mild	High	False	*No*
Sunny	Cool	Normal	False	*Yes*
Rainy	Mild	Normal	False	*Yes*
Sunny	Mild	Normal	True	*Yes*
Overcast	Mild	High	True	*Yes*
Overcast	Hot	Normal	False	*Yes*
Rainy	Mild	High	True	*No*

$$\mathrm{H} = -\sum_{i=1}^{K} p_k \log_2 p_k$$
$$= -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$
$$= 0.94$$

Complex Example – Information Gain [2]

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	**High**	False	No
Sunny	Hot	**High**	True	No
Overcast	Hot	**High**	False	Yes
Rainy	Mild	**High**	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	**High**	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	**High**	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	**High**	True	No

$$InfoGain(\text{Humidity}) = H - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

$$0.94 - \frac{7}{14} H_L - \frac{7}{14} H_R$$

$$H_L = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.592$$

$$H_R = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$$

$$InfoGain(\text{Humidity}) = 0.94 - 0.296 - 0.4925 = 0.1515$$

Prepared by: Er. Dinesh Baniya Kshatri 45

Complex Example – Information Gain [3]

- Information gain for each feature:
  - Outlook = 0.247
  - Temperature = 0.029
  - Humidity = 0.152
  - Windy = 0.048
- Initial split is on outlook, because it is the feature with the highest information gain.

Prepared by: Er. Dinesh Baniya Kshatri 46

Gini Index Properties

Maximized when elements are heterogeneous (impure):

$$\text{If } p_k = \frac{1}{k}, \text{then}$$

$$\text{Gini} = 1 - \sum_{k=1}^{K} \frac{1}{k^2} = 1 - \frac{1}{k}$$

Minimized when elements are homogenous (pure):

$$\text{If } p_i = 1 \text{ or } p_i = 0, \text{then}$$

$$\text{Gini} = 1 - 1 - 0 = 0$$

Gini Index Example

Suppose we want to split on the first variable (x_1):

x_1	1	2	3	4	5	6	7	8
y	0	0	0	1	1	1	1	1

$$Gini = 1 - \left({}^{3}/_{8}\right)^2 - \left({}^{5}/_{8}\right)^2 = {}^{15}/_{32}$$

If we split at $x_1 < 3.5$: $\Delta Gini = \frac{15}{32} - \frac{3}{8} \cdot 0 - \frac{5}{8} \cdot 0 = {}^{15}/_{32}$

If we split at $x_1 < 4.5$: $\Delta Gini = \frac{15}{32} - \frac{4}{8} \cdot \frac{3}{8} - \frac{4}{8} \cdot 0 = {}^{9}/_{32}$

Classification Error Properties

Tends to create impure nodes:

x_1	1	2	3	4	5	6	7	8
y	0	0	0	1	1	0	0	0

a b

Splitting at b has lower classification error than a, but results in both nodes being impure.

Comparison among Impurity Measures

(Valid for a Two Class (Binary Valued) Problem)