CSX4202_ITX4202: Data Mining Lecture 4

Asst. Prof. Dr. Rachsuda Setthawong
Computer Science Department
Assumption University

Outlines

- Classification: Basic Concepts
- Decision Trees and Issues
- Model Evaluation

Classification: Definition

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.

- 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.
  - A *test set* is used to determine the accuracy of the model.

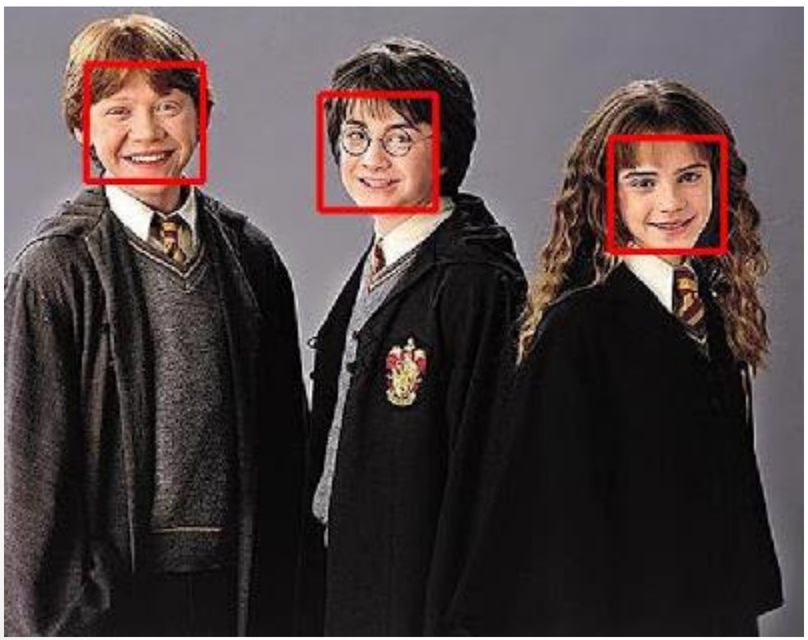
Classification Techniques

- Decision Tree
- Rule-based Methods
- Memory based reasoning
- Naïve Bayes
- K-Nearest Neighbors (kNN)
- Linear Regression
- Neural Network
- Support Vector Machine
- Ensemble Classifiers (Vote)
- Etc.

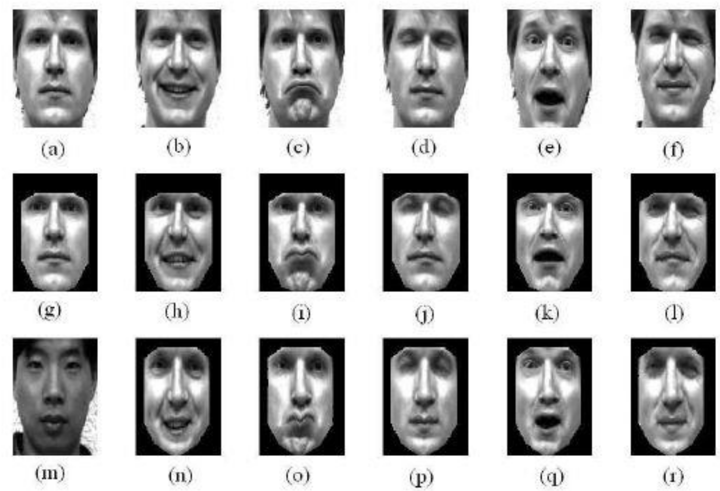
Classification Application – 1

Face recognition and suggested tag

Ron **Harry** **Hermione**



Facial expression classification



Ref: http://www.cs.sfu.ca/~haoz/papers.html

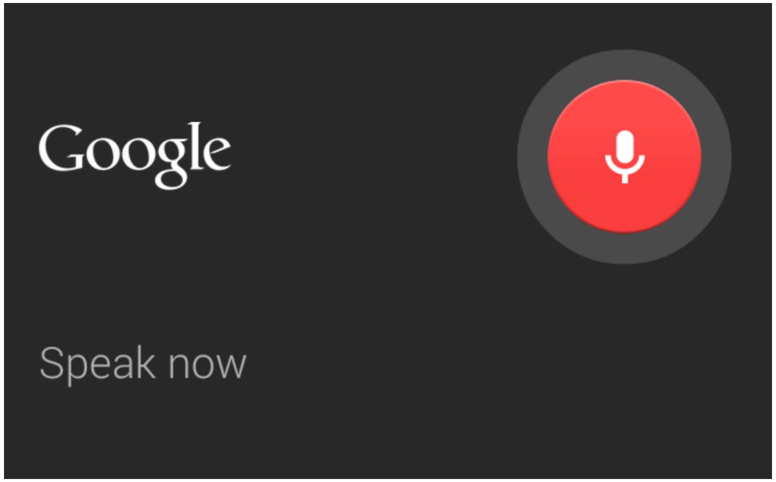
Classification Application – 2

Finger print identification and classification

Ref: http://ntl.bts.gov/lib/jpodocs/repts_te/13769.html

Classification Application – 3

Speech recognition



Classification Application – 4

- Spam e-mail

Try This…

- Determine the spam folder in your email account.
- What kind of pattern(s) do you observe?

Classification Example:
Spam E-mail - 1

ID	Header	Content	Type
1	Medicine$ Buy Here	Be$t Onlline Sohp Hree http://bestherbsreward.ru/?asc1wubobwd	?
2	TODAY DISCOUNT 35%	Click Here [ONLINE PHAMARCY]	?
3	[5850512]传递价值●《远程IO解决方案》资料(70页,5.5MB,PDF)免费下载[b586ow	如果无法正常显示,请点击此	?
4	โปรโมชั่น ! ตั๋ว ท	Show Images	?
5	TODAY DISCOUNT 37%	Click Here [ONLINE PHAMARCY]	?
6	Swiss Rolex Replica Watches 2014 Models Sale	Buy a Rolex, Omega, Breitling at only a fraction of the price! SAVE AN ADDITIONAL 15% ON PURCHASES OF $250 OR MORE	?
7	Apress Android, Swift, and iOS New Book Alert!	Show Images	?
8	My dtac e-service - Quick Payment Confirmation: Mobile No. 081XXXXXXX	Thank you for paying your bill via Quick Payment.	?

Asst. Prof. Dr. Rachsuda Setthawong

Classification Example:
Spam E-mail - 2

ID	Header	Content	Type
1	Medicine$ Buy Here	Be$t Onlline Sohp Hree http://bestherbsreward.ru/?asc1wubobwd	Spam
2	TODAY DISCOUNT 35%	Click Here [ONLINE PHAMARCY]	Spam
3	[5850512]传递价值● 《远程IO解决方案》资料(70页,5.5MB,PDF)免费下载[b586ow	如果无法正常显示,请点击此	Spam
4	โปรโมชั่น ! ตั๋ว ท	Show Images	Normal
5	TODAY DISCOUNT 37%	Click Here [ONLINE PHAMARCY]	Spam
6	Swiss Rolex Replica Watches 2014 Models Sale	Buy a Rolex, Omega, Breitling at only a fraction of the price! SAVE AN ADDITIONAL 15% ON PURCHASES OF $250 OR MORE	Spam
7	Apress Android, Swift, and iOS New Book Alert!	Show Images	Normal
8	My dtac e-service - Quick Payment Confirmation: Mobile No. 081XXXXXXX	Thank you for paying your bill via Quick Payment.	Normal

Asst. Prof. Dr. Rachsuda Setthawong

Classification Example:
Spam E-mail – Transform Data

ID	Header	Content	Type
1	Medicine$ Buy Here	Be$t Online Sohp Hree http://bestherbsreward.ru/?asc1wubobwd	Spam
2	TODAY DISCOUNT 35%	Click Here [ONLINE PHAMARCY]	Spam
3	[5850512]传递价值●《远程IO解决方案》资料(70页,5.5MB,PDF)免费下载[b586ow	如果无法正常显示,请点击此	Spam
4	โปรโมชั่น ! ตั๋ว ท	Show Images	Normal
5	TODAY DISCOUNT 37%	Click Here [ONLINE PHAMARCY]	Spam
6	Swiss Rolex Replica Watches 2014 Models Sale	Buy a Rolex, Omega, Breitling at only a fraction of the price! SAVE AN ADDITIONAL 50% ON PURCHASES OF $250 OR MORE	Spam
7	Apress Android, Swift, and iOS New Book Alert!	Show Images	Normal
8	My dtac e-service - Quick Payment Confirmation: Mobile No. 081XXXXXXX	Thank you for paying your bill via Quick Payment.	Normal

ID	Link?	Discount	Non-Thai/Non-English?	Type
1	Y	N	N	Spam
2	Y	Y	N	Spam
3	N	N	Y	Spam
4	N	N	N	Normal
5	Y	Y	N	Spam
6	N	N	N	Spam
7	N	N	N	Normal
8	N	N	N	Normal

Training Dataset

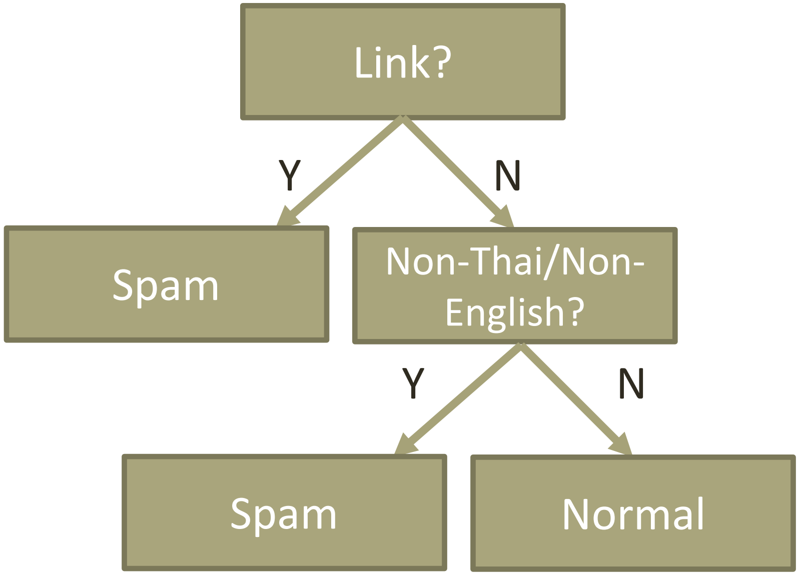
Asst. Prof. Dr. Rachsuda Setthawong

Classification Example:

Spam E-mail – Learn a Model (Induction)

ID	Link?	Discount	Non-Thai/Non-English?	Type
1	Y	N	N	Spam
2	Y	Y	N	Spam
3	N	N	Y	Spam
4	N	N	N	Normal
5	Y	Y	N	Spam
6	N	N	N	Spam
7	N	N	N	Normal
8	N	N	N	Normal

Training Dataset

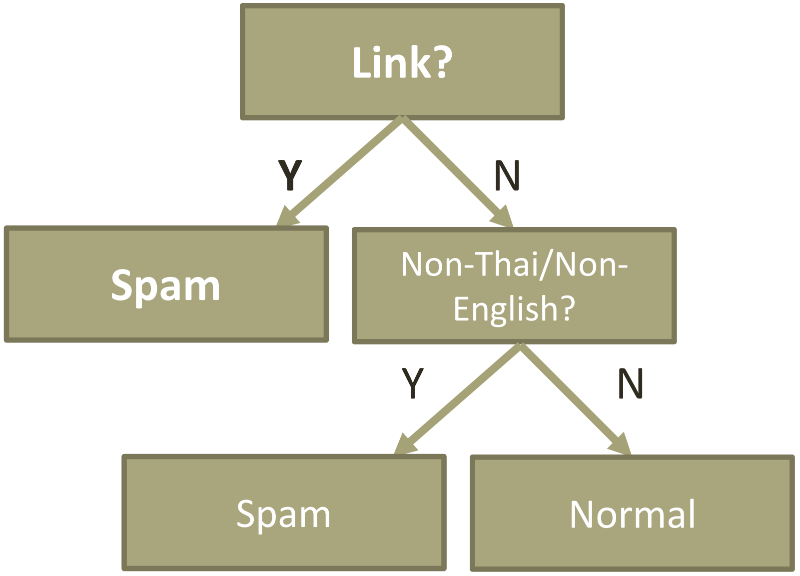


Asst. Prof. Dr. Rachsuda Setthawong

Classification Example:
Spam E-mail – Apply the Model (Deduction)

ID	Link?	Discount	Non-Thai/Non-English?	Type
1	Y	Y	N	?
2	N	N	N	?

Test Dataset



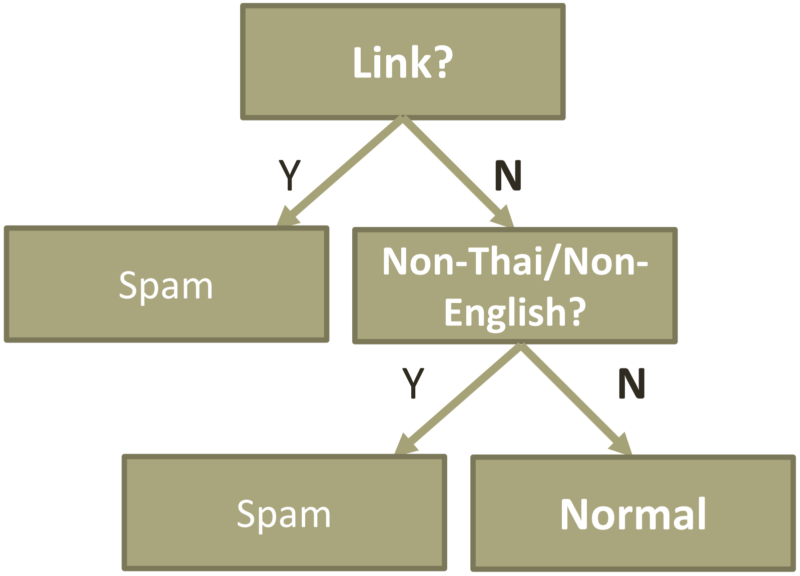
Asst. Prof. Dr. Rachsuda Setthawong

Classification Example:

Spam E-mail - Apply the Model (Deduction)

ID	Link?	Discount	Non-Thai/Non-English?	Type
1	Y	Y	N	?
2	N	N	N	?

Test Dataset



Asst. Prof. Dr. Rachsuda Setthawong

Summary:
Classification Steps

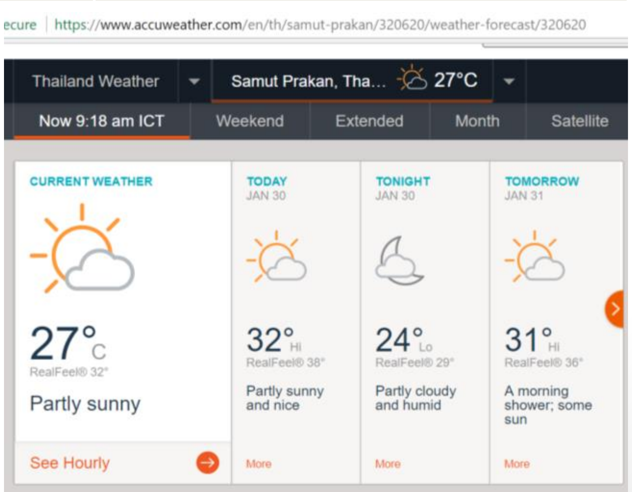
1. Prepare a training dataset
2. Build a classification model (classifier) using a learning algorithm.
3. Apply the model on the test dataset

Try This

- Let’s go through the process of creating another useful classification model in real life.
- Create a simple classification model that will predict whether a student will be interested to take this course or not based on their historical background.
  - What are the relating non-class attributes that should be included in a dataset?
  - What is the class attributes of this dataset?
  - How to apply a classification model for prediction?

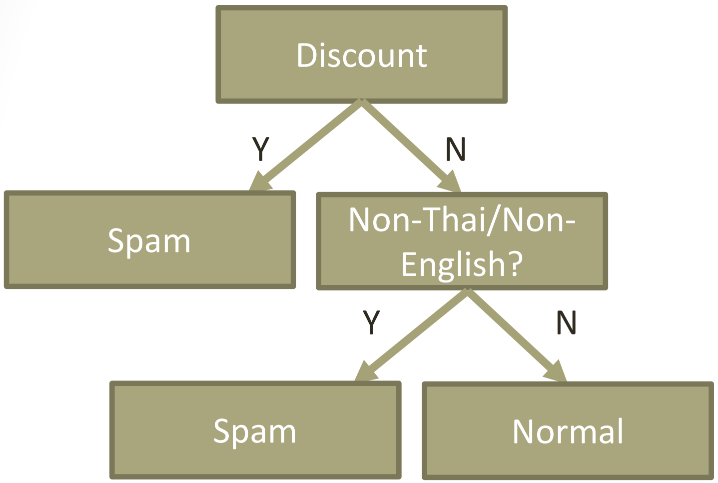
Classification vs Regression

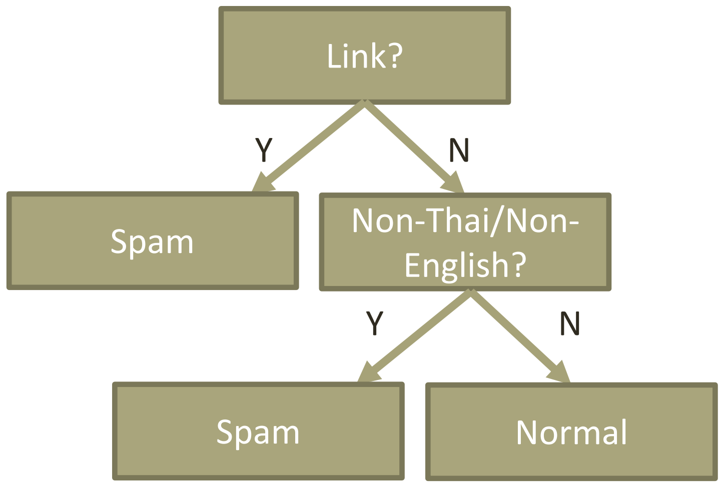
	Classification	Regression
Build a model from a training dataset with class label?	Yes	Yes
Type of class label	Nominal	Numeric
Examples	• Spam e-mail • Rainfall or not	• Predicted sale amount for the next quarter • Weather forecast

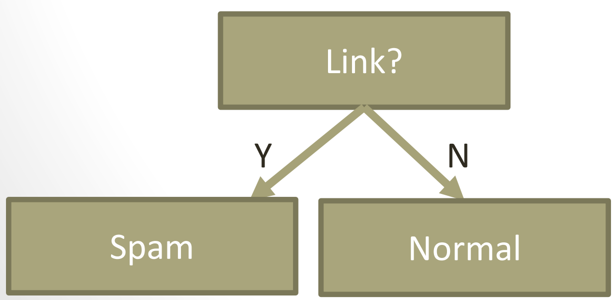


Asst. Prof. Dr. Rachsuda Setthawong

How to build a Decision Tree?







"Which model is the best?"

Outlines

- Classification: Basic Concepts
- Decision Trees and Issues
- Model Evaluation

Asst. Prof. Dr. Rachsuda Setthawong

Decision Tree Induction

- Many Algorithms:
  - Hunt's Algorithm (one of the earliest)
  - CART
  - ID3, C4.5, C5.0
  - SLIQ, SPRINT

"Divide and Conquer Approach"
(Data splitting)

"Look for purity"

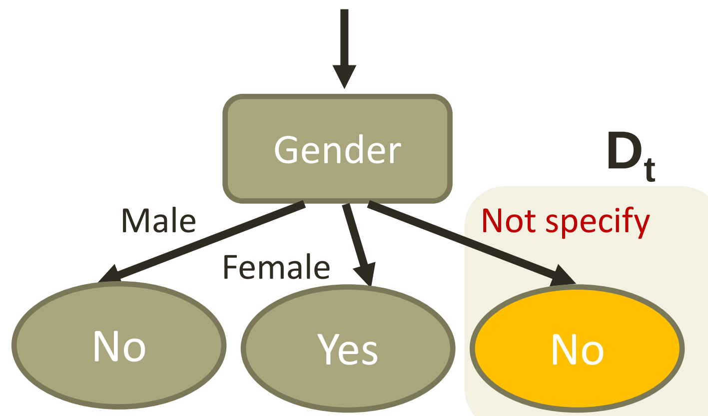
General Structure of Hunt's Algorithm

- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d

    Recursively apply the procedure to each subset.
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets.

    Recursively apply the procedure to each subset.

Age	Gender	PlaySport	PassCourse?
16	Male	No	**No**
17	Female	No	Yes
20	Male	Yes	**No**
20	Male	No	**No**



Age	Gender	Play-Sport	Pass-Course?
16	Male	No	No
20	Male	Yes	No
20	Male	No	No

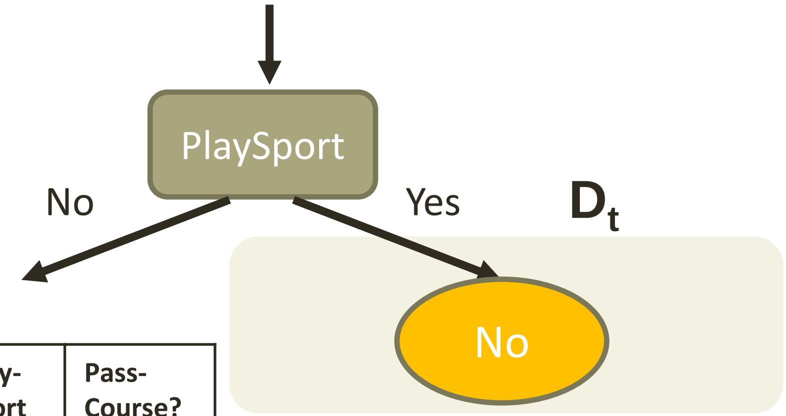
Age	Gender	Play-Sport	Pass-Course?
17	Female	No	Yes

Asst. Prof. Dr. Rachsuda Setthawong

General Structure of Hunt's Algorithm

- Let D_t be the set of training records that reach a node t
- General Procedure:
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t contains records that belong to more than one class, use an attribute test to split the data into smaller subsets.

Age	Gender	PlaySport	PassCourse?
16	Male	**No**	No
17	Female	**No**	Yes
20	Male	**Yes**	No
20	Male	**No**	No



Recursively apply the procedure to each subset.

Age	Gender	Play-Sport	Pass-Course?
16	Male	No	No
17	Female	No	Yes
20	Male	No	No

Age	Gender	Play-Sport	Pass-Course?
20	Male	Yes	**No**

Asst. Prof. Dr. Rachsuda Setthawong

General Structure of Hunt's Algorithm

- Let D_t be the set of training records that reach a node t

- General Procedure:
  - If D_t is an empty set, then t is a leaf node labeled by the default class, y_d
  - If D_t contains records that belong the same class y_t, then t is a leaf node labeled as y_t
  - If D_t contains records that belong to more than one class, use an attribute test (e.g., Gender) to split the data into smaller subsets.

  Recursively apply the procedure to each subset.

Age	Gender	PlaySport	PassCourse?
16	Male	No	**No**
17	Female	No	Yes
20	Male	Yes	**No**
20	Male	No	**No**

PlaySport

No

D_t

Age	Gender	Play-Sport	Pass-Course?
16	Male	No	**No**
17	Female	No	**Yes**
20	Male	No	**No**

??

Yes

Age	Gender	Play-Sport	Pass-Course?
20	Male	Yes	No

Asst. Prof. Dr. Rachsuda Setthawong

Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?

  - Determine when to stop splitting

How to Specify Test Condition?

Depends on attribute types

- Nominal
- Ordinal
- Continuous

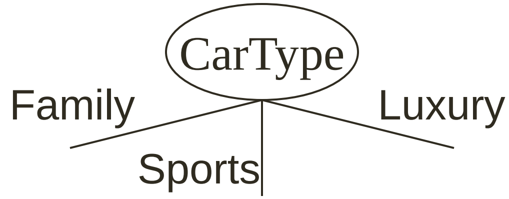
Depends on number of ways to split

- Multi-way split
- 2-way split

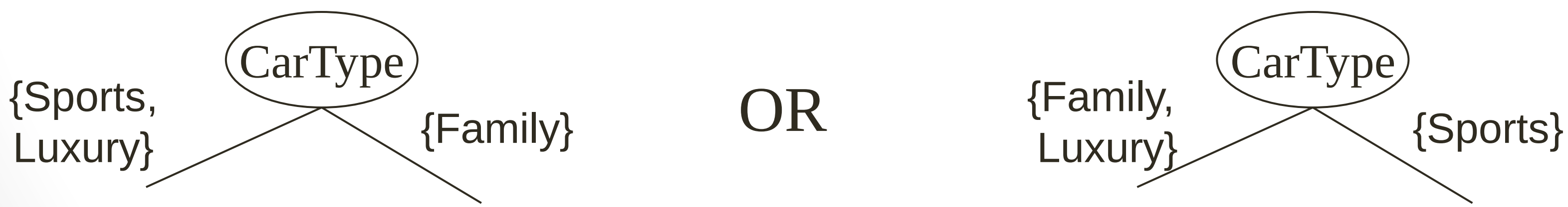
Asst. Prof. Dr. Rachsuda Setthawong

Splitting Based on Nominal Attributes

- Multi-way split: Use as many partitions as distinct values.

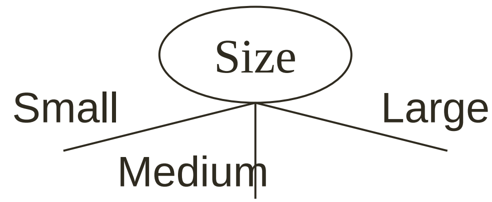


- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.

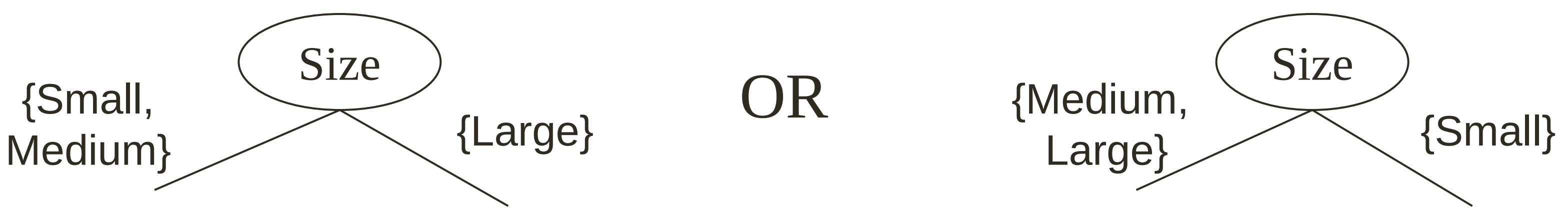


Splitting Based on Ordinal Attributes

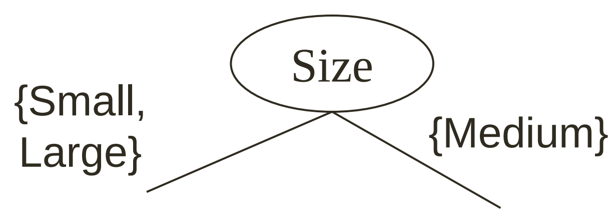
- Multi-way split: Use as many partitions as distinct values.



- Binary split: Divides values into two subsets.
  Need to find optimal partitioning.

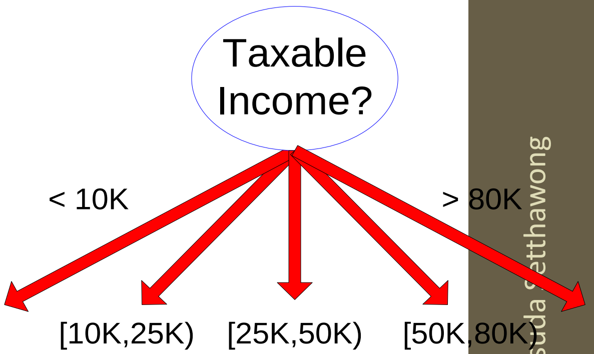


- What about this split?

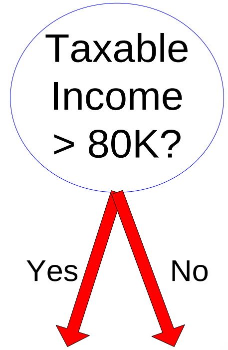


Splitting Based on Continuous Attributes

- Multi-way split:
  - Discretization to form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering.
- Binary split:
  - Binary Decision: $(A < v)$ or $(A \geq v)$
    - consider all possible splits and finds the best cut
    - can be more compute intensive



(ii) Multi-way split



(i) Binary split

Asst. Prof. Dr. Rachsuda Setthawong

Tree Induction

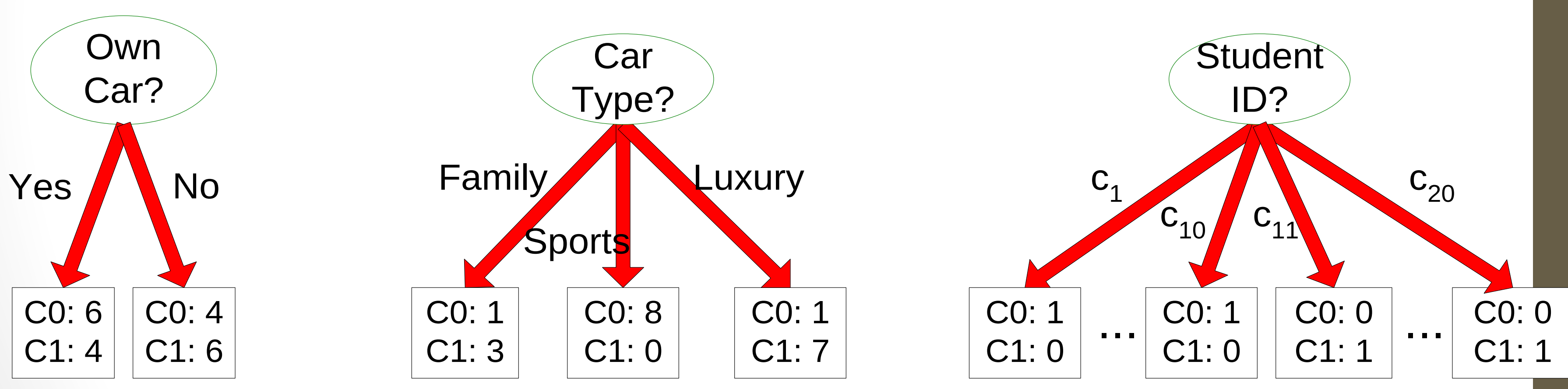
- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?

  - Determine when to stop splitting

How to Determine the Best Split

Before Splitting: **10 records of class 0,**
10 records of class 1



Which test condition is the best?

Asst. Prof. Dr. Rachsuda Setthawong

How to Determine the Best Split

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

C0: 5 C1: 5	C0: 9 C1: 1
Non-homogeneous, **High degree of impurity**	**Homogeneous,** **Low degree of impurity**

Finding the Best Split

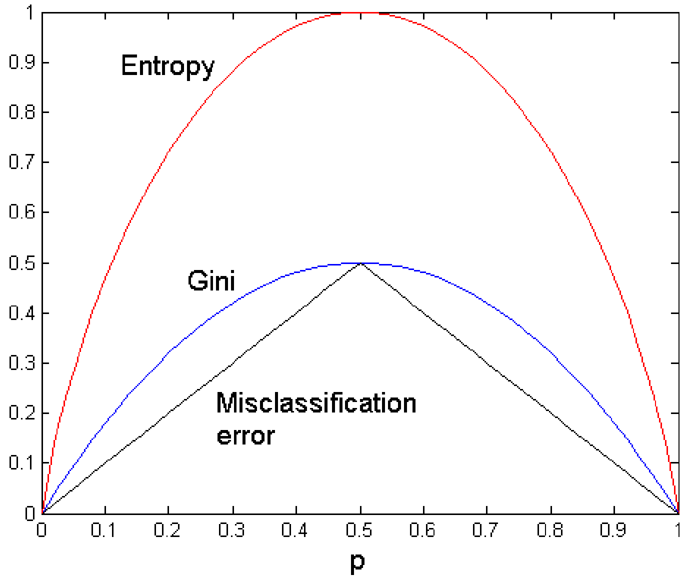
1. Compute impurity measure (P) before splitting
2. Compute impurity measure (M) after splitting
   - Compute impurity measure of each child node
   - M is the weighted impurity of children
3. Choose the attribute test condition that produces *the highest gain*

Gain = P – M

or equivalently, lowest impurity measure after splitting (M)

Measures of Node Impurity

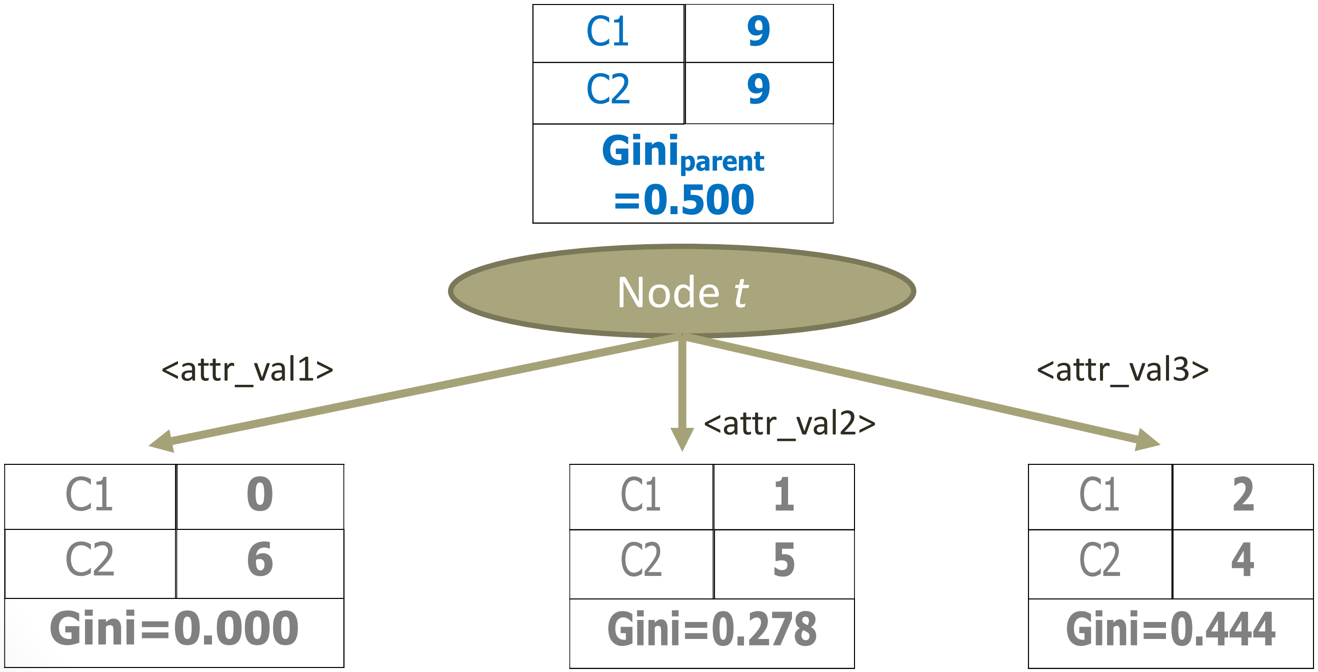
- Gini Index
- Entropy
- Misclassification error



For a 2-class problem

Goal of Splitting

$$\text{Gini}_{\text{Parent}} > \Sigma\, \text{Gini}_{\text{Split}}$$



GINI

- Gini Index for a given node t :

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

(NOTE: $p(j \mid t)$ is the relative frequency of class *j* at node *t*).

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

$Gini_{Split}$ (M)

- Used in CART, SLIQ, SPRINT.
- When a node p is split into k partitions (children), the quality of split is computed as,

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

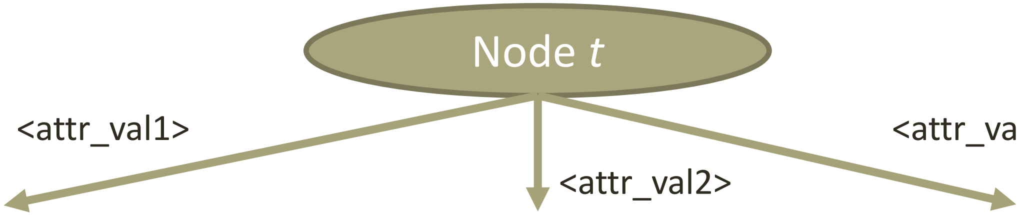
where, n_i = number of records at child i,

n = number of records at node p.

Examples:
Computing GINI

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

C1	9
C2	9
$Gini_{parent}$ =0.500	



C1	0
C2	6
Gini=0.000	

C1	1
C2	5
Gini=0.278	

C1	2
C2	4
Gini=0.444	

P(C1) = 0/6 = 0 P(C2) = 6/6 = 1

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

P(C1) = 1/6 P(C2) = 5/6

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

P(C1) = 2/6 P(C2) = 4/6

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

Asst. Prof. Dr. Rachsuda Setthawong

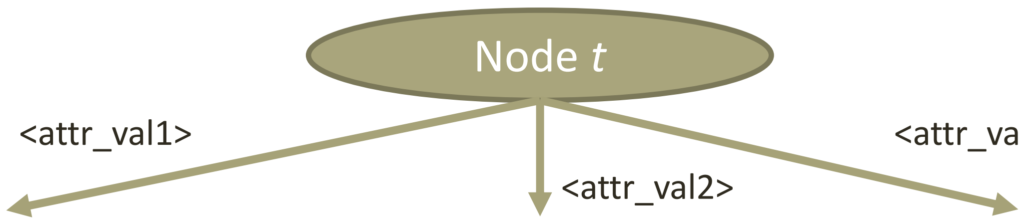
Examples:

Computing GINI_{Split}

(M: weighted aggregation)

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

C1	9
C2	9
Gini_{parent} =0.500	



C1	**0**
C2	**6**
Gini=0.000	

C1	**1**
C2	**5**
Gini=0.278	

C1	**2**
C2	**4**
Gini=0.444	

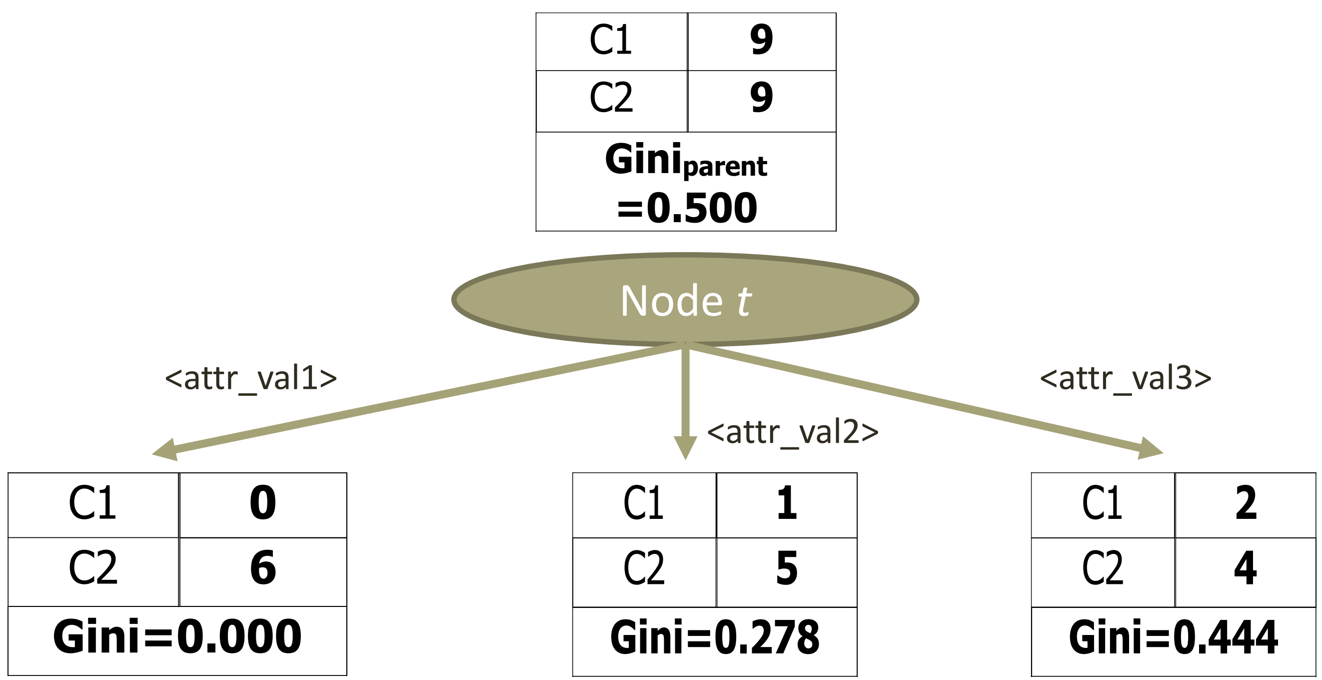
$$\text{Gini}_{Split} = \underbrace{(6/18) * 0}_{\text{<attr\_val1>}} + \underbrace{(6/18) * 0.278}_{\text{<attr\_val2>}} + \underbrace{(6/18) * 0.444}_{\text{<attr\_val3>}} = 0+0.092+0.074 = 0.166$$

Asst. Prof. Dr. Rachsuda Setthawong

Revisit Goal of Splitting:

Gain = P - M

Gain if $Gini_{Parent}$ > $Gini_{Split}$



$Gini_{Split}$ = 0.166

Gain = P – M = 0.5 - 0.166 = 0.334

Asst. Prof. Dr. Rachsuda Setthawong

Alternative Splitting Criteria Based on INFO

- **Entropy** at a given node t:

$$\boxed{Entropy(t) = -\sum_{j} p(j \mid t) \log_2 p(j \mid t)}$$

(NOTE: $p(j \mid t)$ is the relative frequency of class j at node t).

- Measures homogeneity of a node.
  - Minimum (0.0) when all records belong to one class, implying most information
  - Maximum ($\log n_c$) when records are equally distributed among all classes implying least information

Asst. Prof. Dr. Rachsuda Setthawong

Entropy$_{\text{Split}}$

$$Entropy_{split} = \sum_{k=1}^{n} \frac{n_i}{n} Entropy\ (i)$$

where, n_i = number of records at child i,
n = number of records at node p.

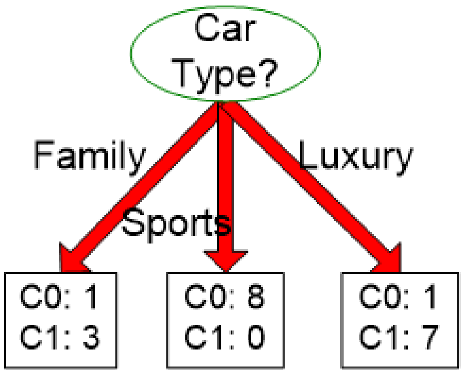
Splitting Based on INFO:
Information Gain

$$Gain_{split} = Entropy\,(parent) - \,Entropy_{split}\,(child)$$

- Choose the split that achieves most reduction (maximizes GAIN)
- Used in ID3 and C4.5
- Disadvantage: Tends to prefer splits that result in large number of partitions, each being small but pure.

Example

$$Entropy(t) = -\sum_j p(j|t)\log_2 p(j|t)$$



$Entropy_{Family}$ = – (1/4) $\log_2$ (1/4) – (3/4) $\log_2$ (3/4) = 0.81

$Entropy_{Sport}$ = – (8/8) $\log_2$ (8/8) – (0/8) $\log_2$ (0/8) = 0 – 0 = 0

$Entropy_{Luxury}$ = – (1/8) $\log_2$ (1/8) – (7/8) $\log_2$ (7/8) = 0.29

$$Entropy_{split} = \sum_{k=1}^{n} \frac{n_i}{n} Entropy\ (i)$$

$Entropy_{CarType}$= (4/20)*0.81 + (8/20)*0 + (8/20)*0.29 = 0.28

$$Gain_{split} = Entropy\ (parent) - Entropy_{split}\ (child)$$

$Entropy_{Parent}$ = – (10/20) $\log_2$ (10/20) – (10/20) $\log_2$ (10/20)

= – (-0.5) – (-0.5) = 1

$Gain_{CarType}$ = 1 – 0.28 = 0.72

Asst. Prof. Dr. Rachsuda Setthawong

Splitting Based on INFO:

Gain Ratio

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

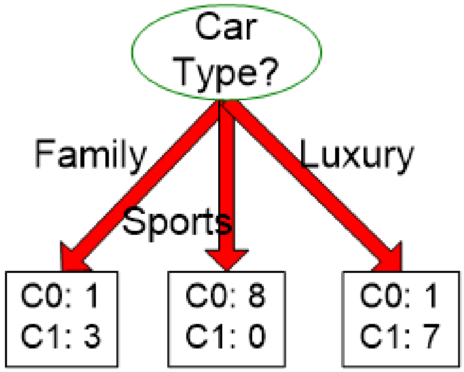
$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

Parent Node, p is split into k partitions

n_i is the number of records in partition I

- Higher entropy partitioning is penalized!
- Used in C4.5
- Designed to overcome the disadvantage of Information Gain

Example



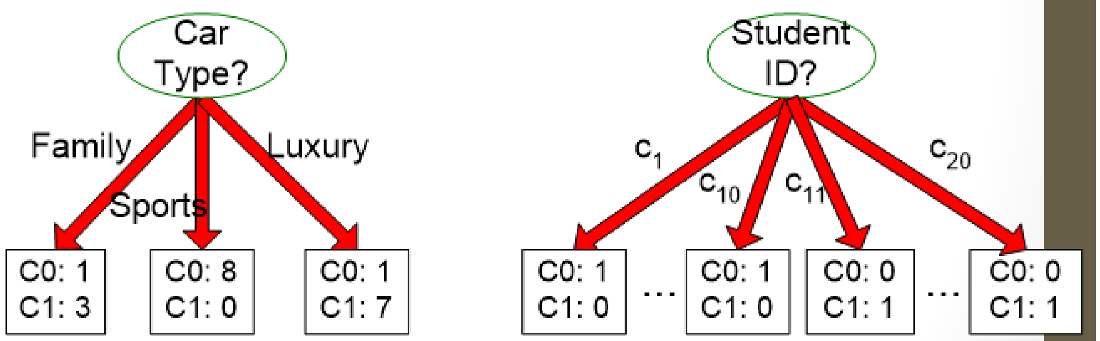
$$SplitINFO = -\sum_{i=1}^{k} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

SplitInfo = – (4/20) $\log_2$ (4/20) – (8/20) $\log_2$ (8/20) – (8/20) $\log_2$ (8/20) = 1.52

$$GainRATIO_{split} = \frac{GAIN_{Split}}{SplitINFO}$$

$Gain_{CarType} = 0.72$

$GainRATIO_{Split} = 0.72 / 1.52 = 0.47$



Which test condition is the best?

$\text{Entropy}_{\textbf{StudentID}} = (1/20) * [- (1/1) \log_2 (1/1) – 0/1 \log_2 (0/1)] * 20 = 0$

$\text{GainRATIO}_{\text{StudentID}} = (1 – 0) \ / \ [[-(1/20) \log_2 (1/20)] * 20] = 0.23$

$\text{GainRATIO}_{\text{CarType}} > \text{GainRATIO}_{\text{StudentID}}$

= 0.47 = 0.23

Splitting Criteria based on **Classification Error**

- Classification error at a node t :

$$\boxed{Error(t) = 1 - \max_{i} P(i \mid t)}$$

- Measures misclassification error made by a node.
  - Minimum (0.0) when all records belong to one class, implying most interesting information
  - Maximum ($1 - 1/n_c$) when records are equally distributed among all classes, implying least interesting information

Asst. Prof. Dr. Rachsuda Setthawong

Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?
      - Binary, Categorical, Continuous Attributes

  - Determine when to stop splitting

How to Determine the Best Split?

Categorical Attributes

Multi-way split

	CarType		
	Family	Sports	Luxury
C1	1	2	1
C2	4	1	1
Gini	0.393		

Two-way split

(find best partition of values)

	CarType	
	{Sports, Luxury}	{Family}
C1	3	1
C2	2	4
Gini	0.400	

	CarType	
	{Sports}	{Family, Luxury}
C1	2	2
C2	1	5
Gini	0.419	

Note: ID3 and C4.5 allows Categorical Attributes

How to Determine the Best Split?
Continuous Attributes – 1

For each attribute,

1. Sort the attribute on values
2. Linearly scan these values, each time updating the count matrix and computing gini index
3. Choose the split position that has the least Gini index

Note: C4.5 allows Continuous Attributes

How to Determine the Best Split?

Continuous Attributes - 2

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

Sorted Values →

Split Positions →

Cheat	**No**	**No**	**No**	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**No**
	Taxable Income									
	60	**70**	**75**	**85**	**90**	**95**	**100**	**120**	**125**	**220**

	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**	

Tree Induction

- Greedy strategy.
  - Split the records based on an attribute test that optimizes certain criterion.

- Issues
  - Determine how to split the records
    - How to specify the attribute test condition?
    - How to determine the best split?

  - Determine when to stop splitting

Stopping Criteria for Tree Induction

- Stop expanding a node when all the records belong to the same class

- Stop expanding a node when all the records have similar attribute values

- Early termination (to be discussed later)

Advantages
of Decision Tree Based Classification

- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees
- Robust to noise (especially when methods to avoid overfitting are employed)
- Can easily handle redundant or irrelevant attributes
- Accuracy is comparable to other classification techniques for many simple data sets

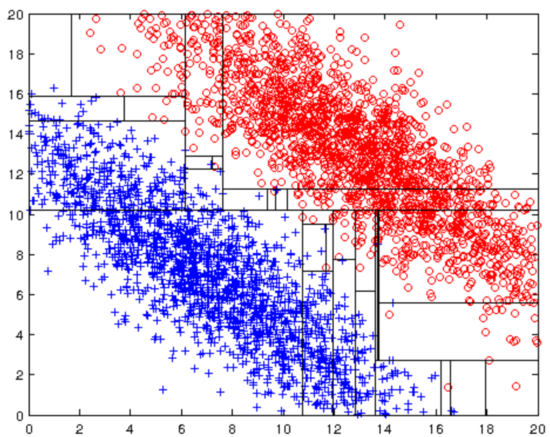
Asst. Prof. Dr. Rachsuda Setthawong

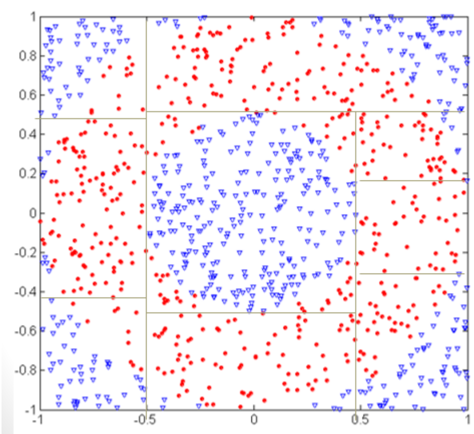
Disadvantages
of Decision Tree Based Classification

- Space of possible decision trees is exponentially large. Greedy approaches are often unable to find the best tree.
- Does not take into account interactions between attributes
- Each decision boundary involves only a single attribute

Examples of Infeasible Datasets for Decision Tree

Both positive (+) and negative (o) classes generated from skewed Gaussians with centers at (8,8) and (12,12) respectively.





Circular points:

$0.5 \leq \text{sqrt}(x_1^2+x_2^2) \leq 1$

Triangular points:

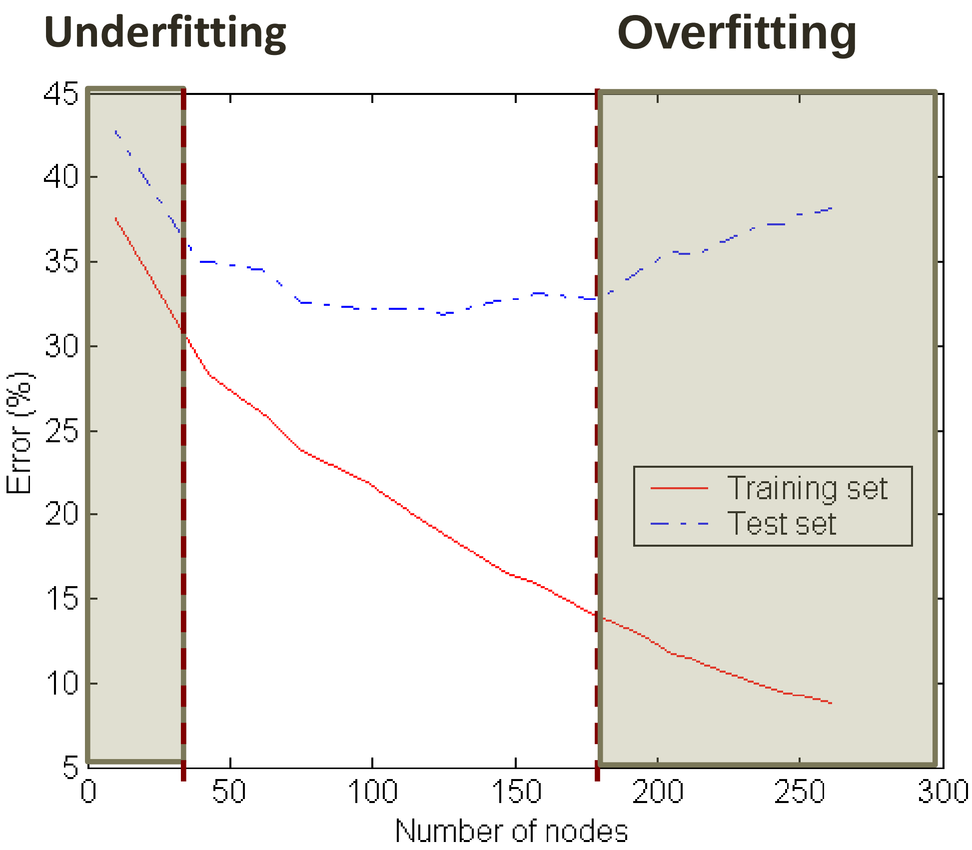
$\text{sqrt}(x_1^2+x_2^2) > 0.5$ or

$\text{sqrt}(x_1^2+x_2^2) < 1$

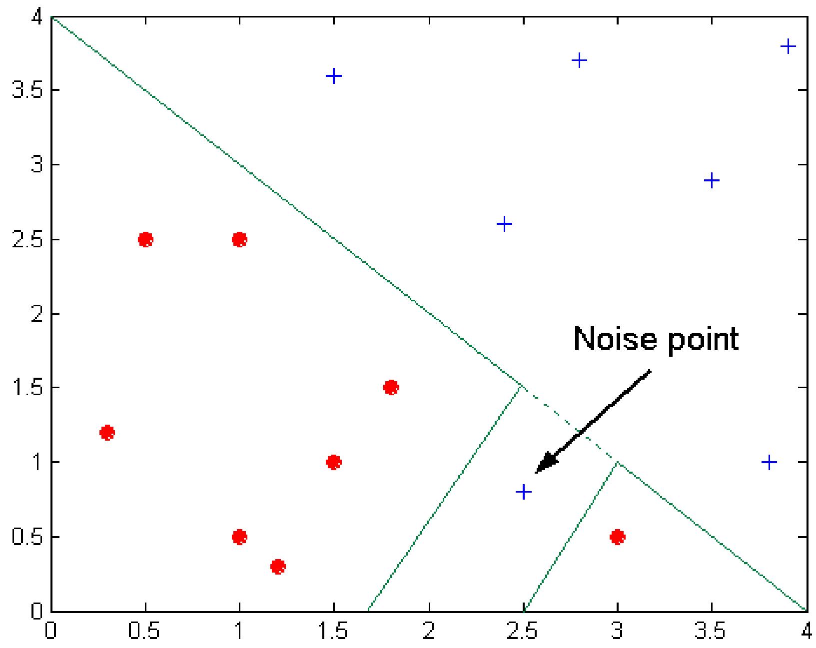
Practical Issues of Classification

- Underfitting and Overfitting
- Costs of Classification
- Missing Values

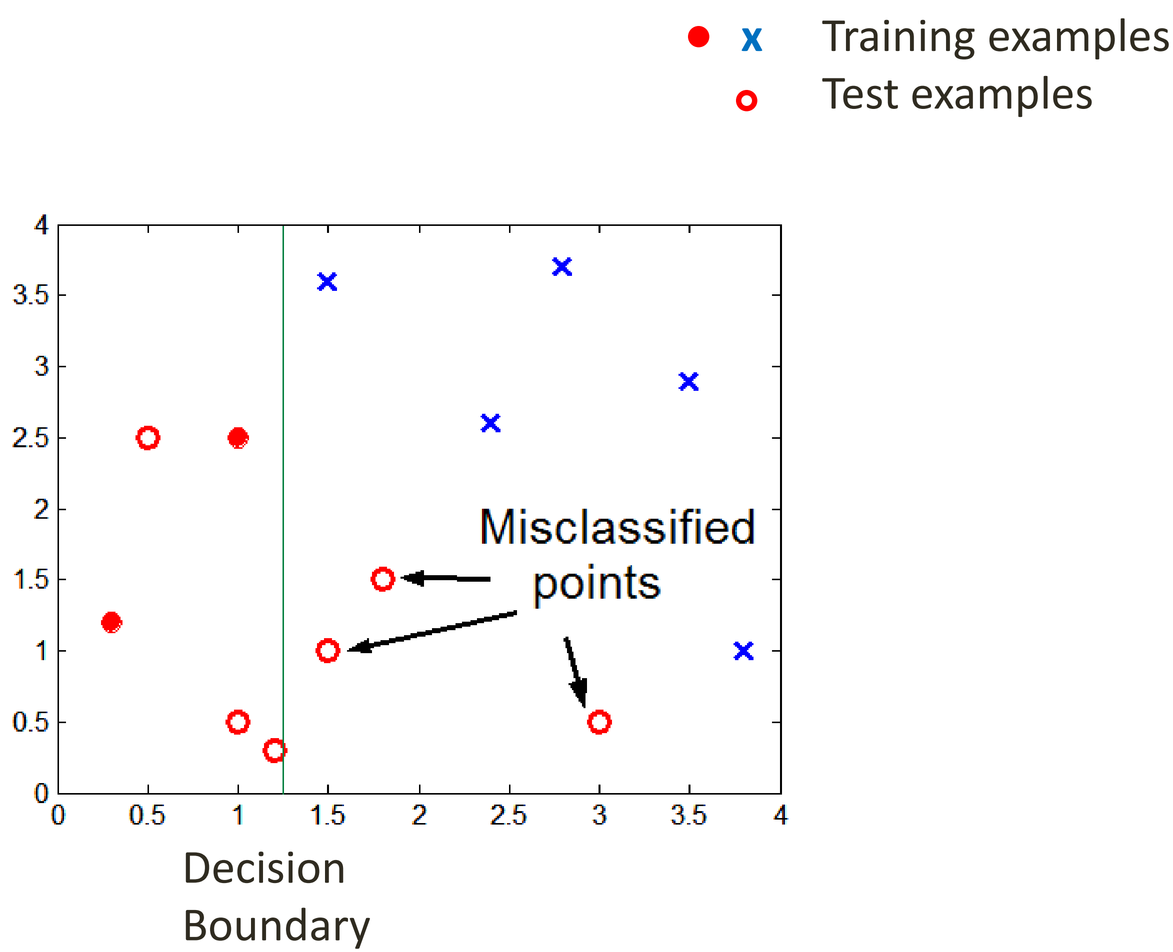
Underfitting and Overfitting



Overfitting due to Noise



Overfitting due to Insufficient Examples



Notes on Overfitting

- Overfitting results in decision trees that are more complex than necessary.
- Training error no longer provides a good estimate of how well the tree will perform on previously unseen records.
- Need new ways for estimating errors.

Estimating Generalization Errors

- Re-substitution errors: error on training (Σ e(t))
- Generalization errors: error on testing (Σ e'(t))
- Methods for estimating generalization errors:
  - Optimistic approach: e'(t) = e(t)

    For a tree with 30 leaf nodes and 10 errors on training (out of 1000 instances):

    Training error = 10/1000 = 0.01 = 1%
  - Pessimistic approach:
    - For each leaf node: e'(t) = e(t) + 0.5
    - Total errors: e'(T) = e(T) + N $\times$ 0.5 (N: number of leaf nodes)
    - For a tree with 30 leaf nodes and 10 errors on training (out of 1000 instances):

      Training error = 10/1000 = 1%

      $\text{Generalization\_error}_{\text{pessimistic}} = (10 + 30\times0.5)/1000 = 0.025 = 2.5\%$
  - Reduced error pruning (REP):
    - Uses validation data set to estimate generalization error

Occam's Razor

- Given two models of similar generalization errors, one should prefer the simpler model over the more complex model

- For complex models, there is a greater chance that it was fitted accidentally by errors in data

How to Address Overfitting

Pre-Pruning (Early Stopping Rule)

- Stop the algorithm before it becomes a fully-grown tree
- Typical stopping conditions for a node:
  - Stop if all instances belong to the same class
  - Stop if all the attribute values are the same
- More restrictive conditions:
  - Evaluate number of instances wrt. a user-specified threshold
  - Evaluate independency of the available features of instances' class distribution (e.g., using χ^2 test)
  - Evaluate Impurity measures (e.g., Gini or information gain).

How to Address Overfitting

Post-pruning

- Grow decision tree to its entirety.
- Trim the nodes of the decision tree in a bottom-up fashion.
    - If generalization error improves after trimming, replace sub-tree by a leaf node.
    - Class label of leaf node is determined from majority class of instances in the sub-tree

Asst. Prof. Dr. Rachsuda Setthawong

Example of Post-Pruning

Class = Yes	20
Class = No	10
Error = 10/30	

Optimistic approach:

Training Error (Before splitting) = 10/30

Training Error (After splitting) = 9/30

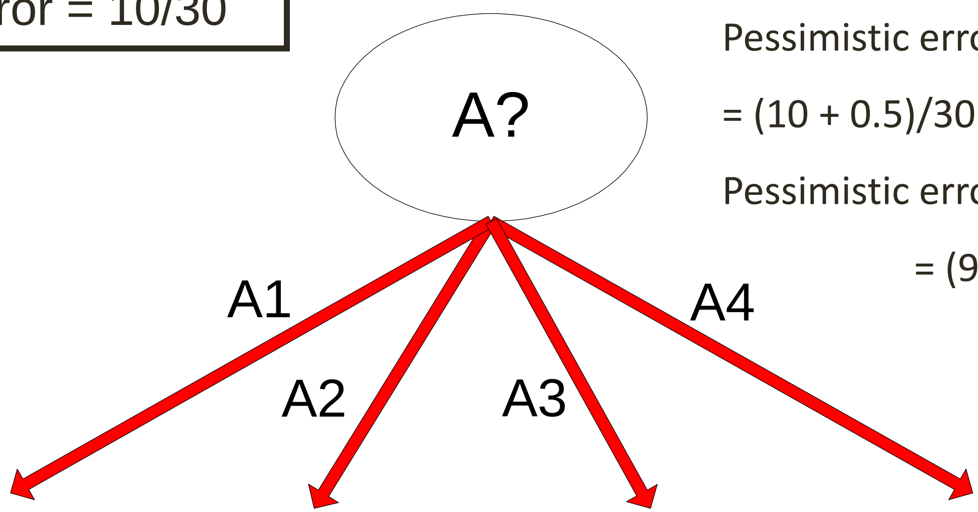
Pessimistic approach:

Pessimistic error (Before splitting)

= (10 + 0.5)/30 = 10.5/30

Pessimistic error (After splitting)

= (9 + 4 × 0.5)/30 = 11/30



Class = Yes	8
Class = No	4

Class = Yes	3
Class = No	4

Class = Yes	4
Class = No	1

Class = Yes	5
Class = No	1

Total errors: e’(T) = e(T) + N × 0.5 (N: number of leaf nodes)

Effects of Missing Values on Decision Tree

- Training phase:
  - Affects how impurity measures are computed
  - Affects how to distribute instance with missing value to child nodes
- Testing phase:
  - Affects how a test instance with missing value is classified

Asst. Prof. Dr. Rachsuda Setthawong

Computing Impurity Measure with Missing Value Exists
Way 1: Ignore the missing value

Tid	Refund	Marital Status	Taxable Income	Class
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	?	Single	90K	Yes

Missing value

(Ignore Tid 10 for the calculation)

Before Splitting:

Entropy(Parent)
= -0.3 log(0.3)-(0.7)log(0.7) = 0.8813

	Class = Yes	Class = No
Refund=Yes	0	3
Refund=No	2	4
Refund=?	1	0

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

Split on Refund:

Entropy(Refund=Yes) = 0

Entropy(Refund=No)
= -(2/6)log(2/6) – (4/6)log(4/6) = 0.9183

Entropy(Children)
= **0.3** (0) + **0.6** (0.9183) = 0.551

Gain = 0.8813 – 0.551 = 0.3303

Asst. Dr. Rachsuda Setthawong

Computing Impurity Measure with Missing Value Exists

Way 2: Include the instance with missing by distributing the instance

Tid	Refund	Marital Status	Taxable Income	Class
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

Refund

Yes

Class=Yes	**0**
Class=No	**3**

No

Cheat=Yes	**2**
Cheat=No	**4**

Missing value

(Distribute it proportionally to both classes)

Tid	Refund	Marital Status	Taxable Income	Class
10	?	Single	90K	Yes

Refund

Yes

Class=Yes	**0 + 3/9**
Class=No	**3**

No

Class=Yes	**2 + 6/9**
Class=No	**4**

Probability that Refund=Yes is 3/9

Probability that Refund=No is 6/9

Assign record to the left child with weight = 3/9 and to the right child with weight = 6/9

Asst. Prof. Dr. Rachsuda Setthawong

Outlines

- Classification: Basic Concepts
- Decision Trees and Issues
- Model Evaluation

Model Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?

- Methods for Performance Evaluation
  - How to obtain reliable estimates?

Model Evaluation

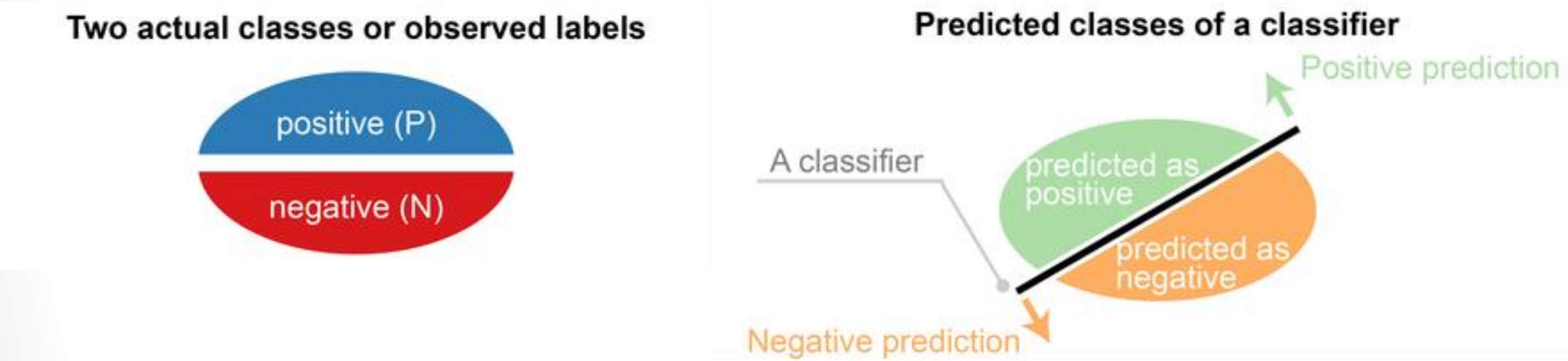
- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?

Straightforward

Metrics	Description / Example
Speed (Computational Time)	**How fast** does it take to classify or build models?
Scalability	**How large** is a dataset can it be applied?
Predictive capability of a model	e.g., Accuracy, Cost, Precision, Recall, F-measure, Weighted-accuracy

Asst. Prof. Dr. Rachsuda Setthawong

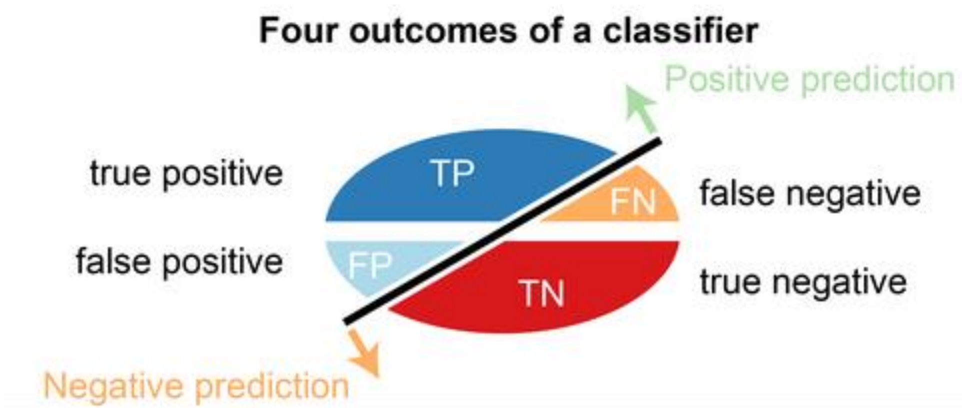
Predictions on Test Datasets



Ref: https://classeval.wordpress.com/introduction/basic-evaluation-measures/

Asst. Prof. Dr. Rachsuda Setthawong

Metrics for Performance Evaluation: Confusion Matrix



	PREDICTED CLASS	
ACTUAL CLASS	Class=Yes	Class=No
Class=Yes	TP (true positive)	FN (false negative)
Class=No	FP (false positive)	TN (true negative)

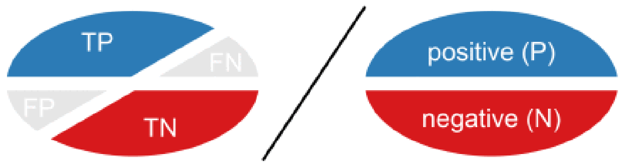
- True positive (TP): correct positive prediction
- False positive (FP): incorrect positive prediction
- True negative (TN): correct negative prediction
- False negative (FN): incorrect negative prediction

Ref: https://classeval.wordpress.com/introduction/basic-evaluation-measures/

Metrics for Performance Evaluation

Accuracy

Accuracy: (TP + TN) / (P + N)



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{n}$$

n: number of records in a dataset

	PREDICTED CLASS	
	Class=Yes	Class=No
ACTUAL CLASS Class=Yes	TP (true positive)	FN (false negative)
ACTUAL CLASS Class=No	FP (false positive)	TN (true negative)

Asst. Prof. Dr. Rachsuda Setthawong

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any class 1 example

Cost Matrix

Cost Matrix	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	**Class=Yes**	**Class=No**
	Class=Yes	C(Yes|Yes)	C(Yes|No)
	Class=No	C(No|Yes)	C(No|No)

Ct(M) = c(Yes|Yes) x TP +
c(No|No) x TN +
c(No|Yes) x FP +
c((Yes|No) x FN

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Ct(M) = c(Yes|Yes) x TP +
c(No|No) x TN +
c(No|Yes) x FP +
c((Yes|No) x FN

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

$\text{Accuracy}_{M1} = 80\%$

$$Cost_{M1} = (-1 * 150) + (0 * 250) + (1 * 60) + (100 * 40) = 3910$$

$\text{Accuracy}_{M2} = 90\%$

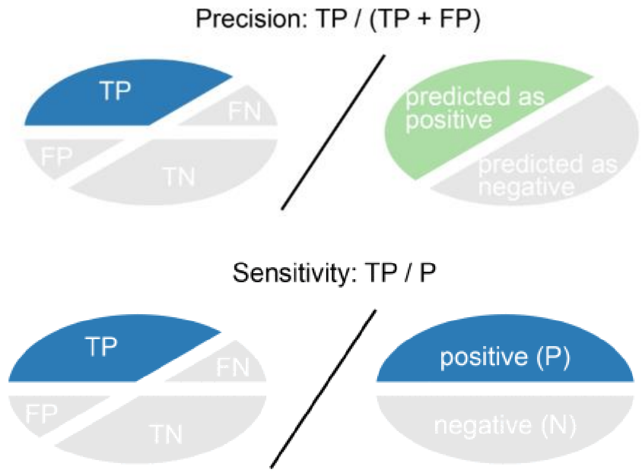
$$Cost_{M2} = (-1 * 250) + (0 * 200) + (1 * 5) + (100 * 45) = 4255$$

Cost-Sensitive Measures

$$Precision\ (p) = \frac{TP}{TP + FP}$$

$$Recall\ (r) = \frac{TP}{TP + FN}$$

$$F - measure\ (F) = \frac{2rp}{r + p}$$



$$Weighted\ Accuracy = \frac{w_1 TP + w_4 TN}{w_1 TP + w_2 FN + w_3 FP + w_4 TN}$$

Precision determines the **fraction of records** that **actually turns out to be positive** *in the group the classifier has declared as a* ***positive class***.

Recall measures the fraction of **positive examples correctly predicted** by the classifier.

Ref: https://classeval.wordpress.com/introduction/basic-evaluation-measures/

Asst. Prof. Dr. Rachsuda Setthawong

Model Evaluation

- Metrics for Performance Evaluation
  - How to evaluate the performance of a model?

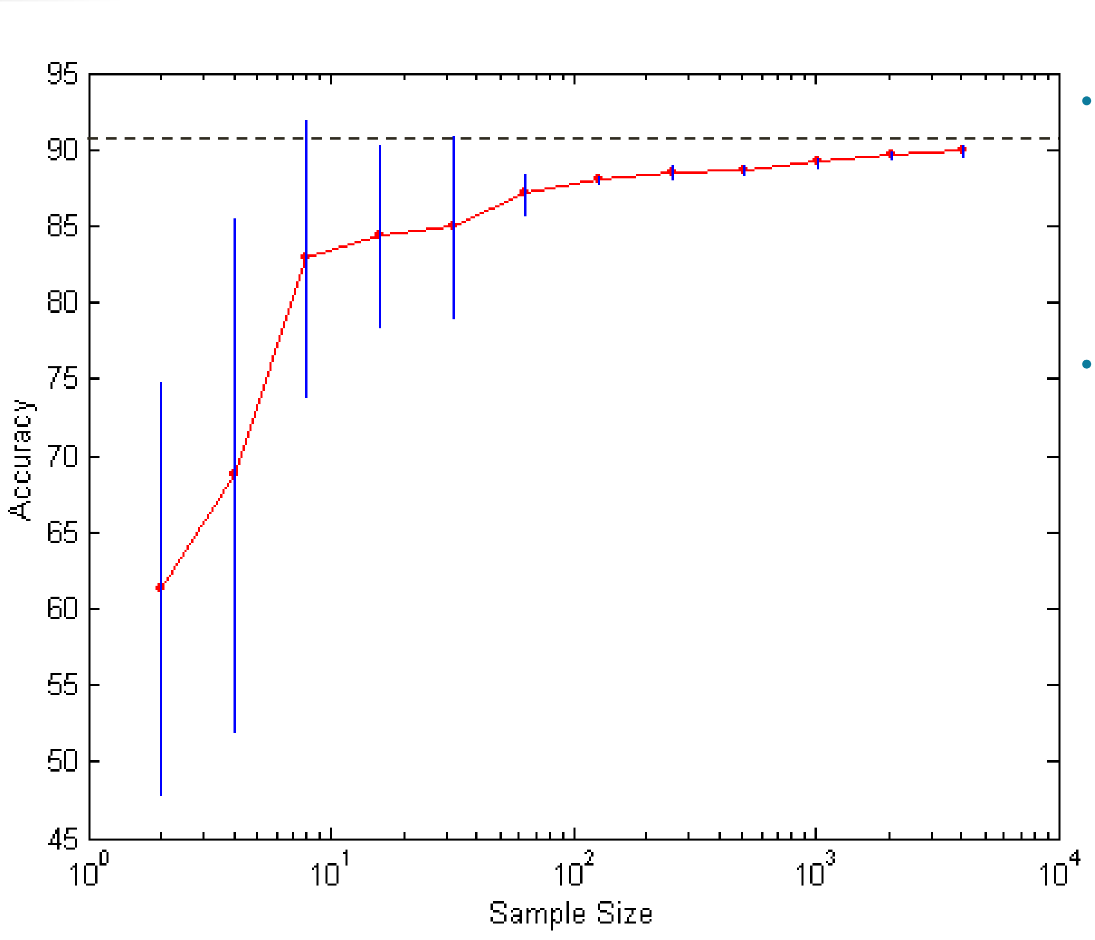
- Methods for Performance Evaluation
  - How to obtain reliable estimates?

Methods for Performance Evaluation

How to obtain a reliable estimate of performance?

- Performance of a model ***depends on several factors***
  - ***E.g.,***
  - Learning algorithm
  - Class distribution (balance/imbalance dataset)
  - Cost of misclassification
  - Size of training and test sets

Learning Curve



- Learning curve shows how accuracy changes with varying sample size
- Effect of small sample size:
  - Bias in the estimate
  - Variance of estimate

Maximizing the Utilization of Instances in the (Small) Dataset

- Holdout
  - Reserve 2/3 for training and 1/3 for testing

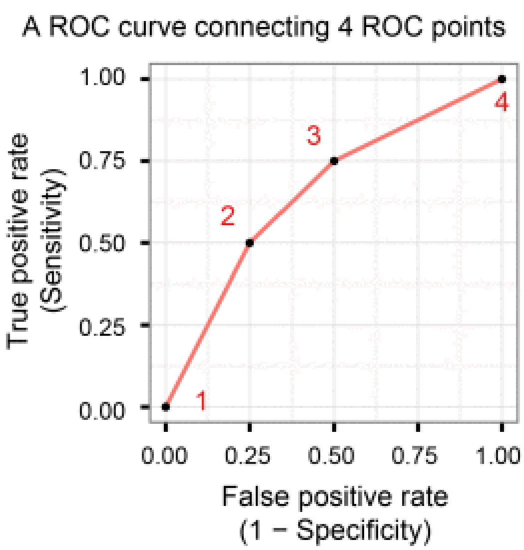
- Cross validation
  - Partition data into k disjoint subsets
  - k-fold (repetitive process): train on k-1 partitions, test on the remaining one
  - Leave-one-out: k=n

- Stratified sampling
  - Split the data into several partitions;
  - then draw random samples from each partition

- Bootstrap
  - Sampling with replacement

ROC (Receiver Operating Characteristic)

- Developed in 1950s for signal detection theory to analyze noisy signals
  - Characterize the trade-off between positive hits and false alarms
- ROC curve plots TPR (on the y-axis) against FPR (on the x-axis)
- Performance of each classifier represented as a point on the ROC curve
  - changing the threshold of algorithm, sample distribution or cost matrix changes the location of the point



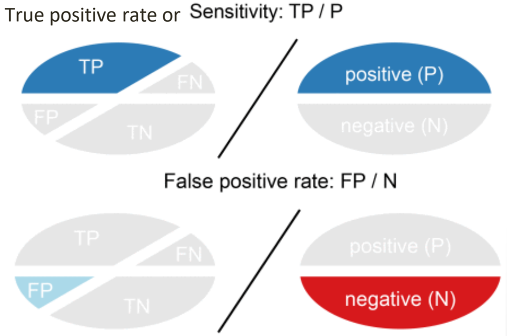
Asst. Prof. Dr. Rachsuda Setthawong

Ref: https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/

How to Construct an ROC curve

InstanceID	P(+|A)	True Class
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+

1. Use classifier that produces predicted result (posterior probability for each test instance P(+|A))
2. Sort the instances according to P(+|A) in decreasing order
3. Apply threshold at each unique value of P(+|A)
4. Count the number of TP, FP, TN, FN at each threshold and plot it
   - TP rate, TPR = TP/(TP+FN)
   - FP rate, FPR = FP/(FP+TN)



Ref: https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/

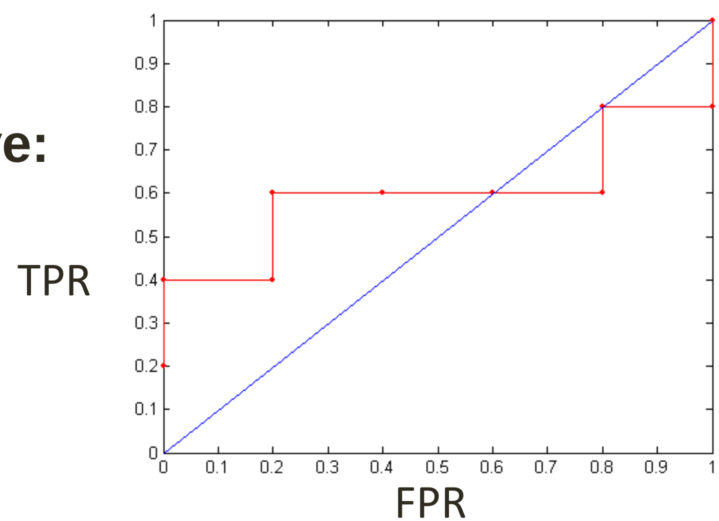
Asst. Prof. Dr. Rachsuda Setthawong

How to construct an ROC curve

The instances sorted according to P(+|A) in decreasing order

InstanceID	10	9	8	7	6	5	4	3	2	1	
True Class	+	-	+	-	-	-	+	-	+	+	
Classification Threshold >= (P(+|A))	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

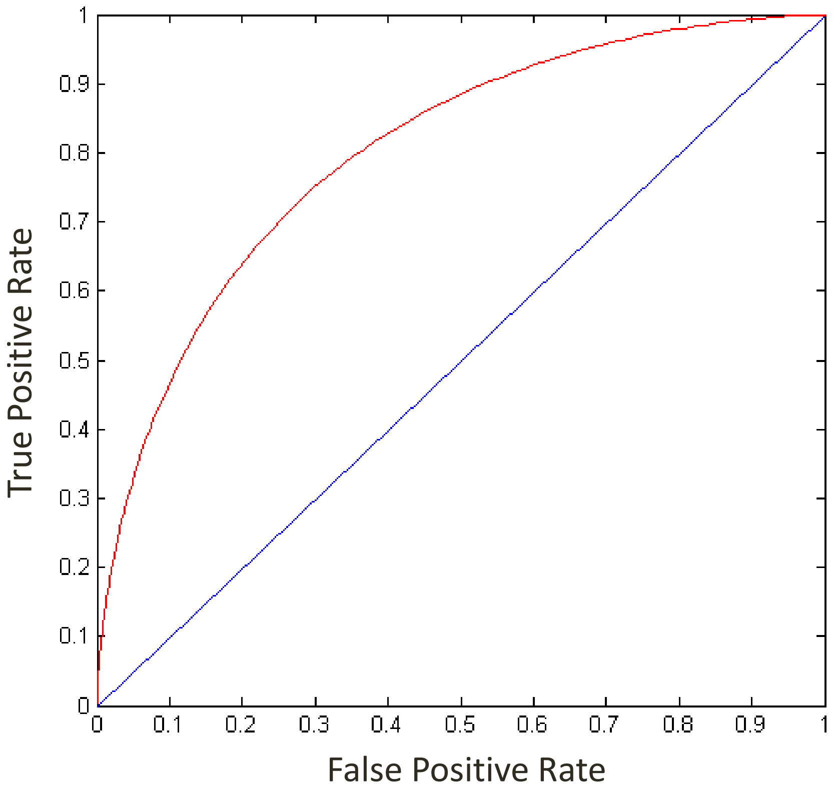
ROC Curve:



- TPR = TP/(TP+FN)
- FPR = FP/(FP+TN)

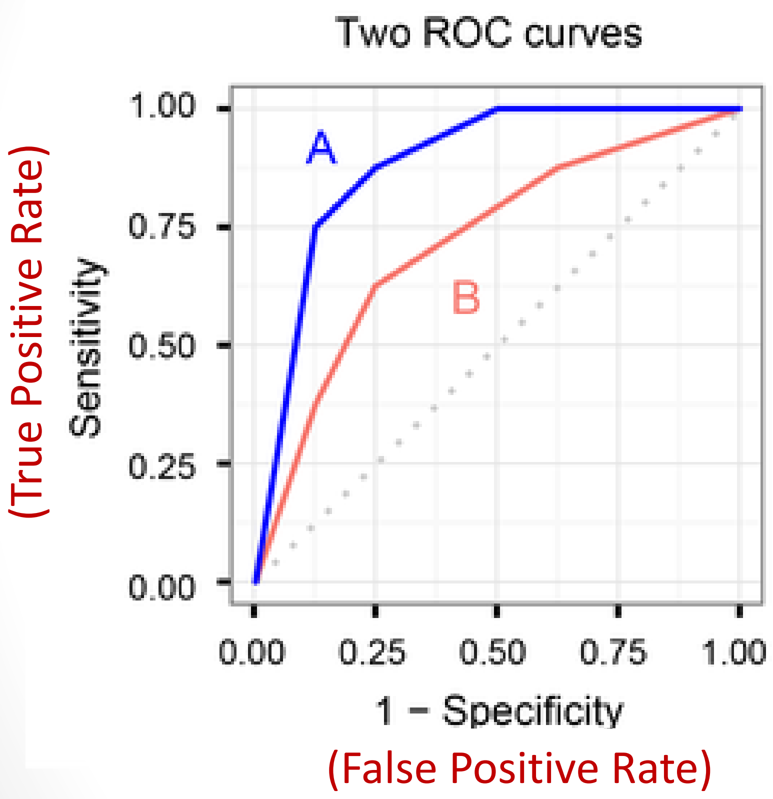
The Performance of a Classifier Indicated in ROC Curve

- Above diagonal line:
  - The larger area under ROC curve (AUC) the better performance of an algorithm is.
- Diagonal line:
  - Random guessing
    - 50/50
- Below diagonal line:
    - Worst than random guessing
      - prediction is opposite of the true class



Asst. Prof. Dr. Rachsuda Setthawong

Using ROC for Model Comparison



- Typical usage:
  - Compare performance of classification models
- Results of ROC in the example:
  - Classifier A clearly outperforms classifier B.

Asst. Prof. Dr. Rachsuda Setthawong

Ref: https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot/

References

- Slides: Lecture Notes for Chapter 4, Introduction to Data Mining by Tan, Steinbach, Kumar
- Practical Data Mining with RapidMiner Studio 6 by Eakasit P., Data Cube (Thailand)