

# Classification fundamentals

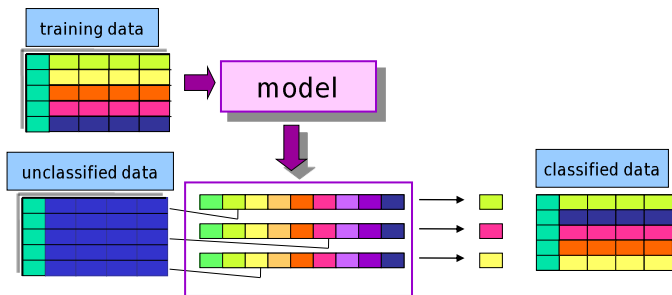


Elena Baralis, Tania Cerquitelli  
*Politecnico di Torino*



## Classification

- Objectives
  - prediction of a class label
  - definition of an interpretable model of a given phenomenon





## Classification: definition

- Given
  - a collection of class labels
  - a collection of data objects labelled with a class label
- Find a descriptive profile of each class, which will allow the assignment of unlabeled objects to the appropriate class



3



## Definitions

- Training set
  - Collection of labeled data objects used to learn the classification model
- Test set
  - Collection of labeled data objects used to validate the classification model



4



## Classification techniques

- Decision trees
- Classification rules
- Association rules
- Neural Networks
- Naïve Bayes and Bayesian Networks
- k-Nearest Neighbours (k-NN)
- Support Vector Machines (SVM)
- ....



5



## Evaluation of classification techniques

- Accuracy
  - quality of the prediction
- Efficiency
  - model building time
  - classification time
- Scalability
  - training set size
  - attribute number
- Robustness
  - noise, missing data
- Interpretability
  - model interpretability
  - model compactness



6

Decision trees



Politecnico di Torino

Example of decision tree

categorical

categorical

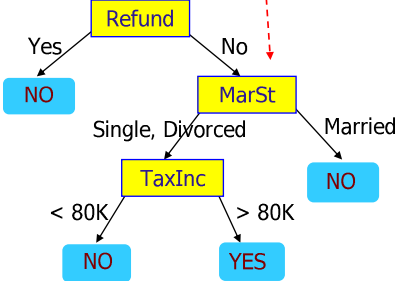
continuous

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


Training Data

Splitting Attributes




```
graph TD; Refund[Refund] -- Yes --> NO1[NO]; Refund -- No --> MarSt[MarSt]; MarSt -- "Single, Divorced" --> TaxInc[TaxInc]; MarSt -- Married --> NO2[NO]; TaxInc -- "< 80K" --> NO3[NO]; TaxInc -- "> 80K" --> YES[YES];
```

Model: Decision Tree



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

8



Another example of decision tree

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

Married

MarSt

Single, Divorced

Refund

Yes

NO

No

TaxInc

< 80K

> 80K

NO


YES

There could be more than one tree that fits the same data!

DBG

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

9



Apply Model to Test Data

Start from the root of tree.

Refund

Yes

NO

No

MarSt

Single, Divorced

Married

TaxInc

< 80K

> 80K

NO

YES


Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

DBG

From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

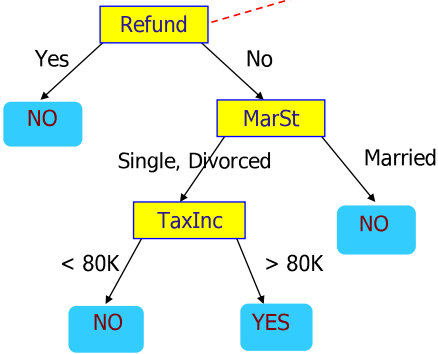
10




Apply Model to Test Data

Test Data


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?





From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

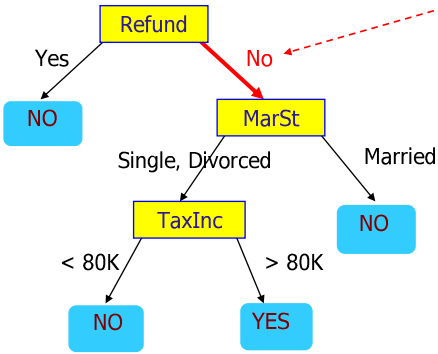
11




Apply Model to Test Data

Test Data


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?





From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

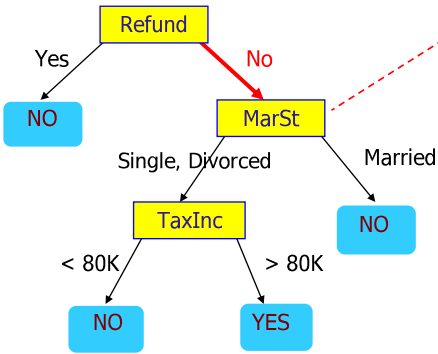
12




Apply Model to Test Data

Test Data


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?





From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

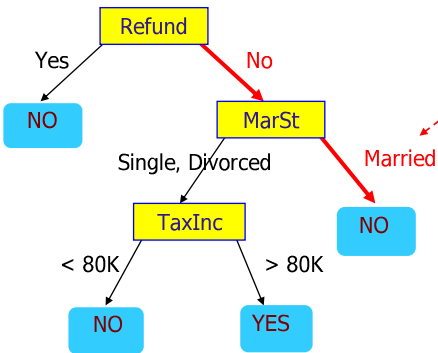
13




Apply Model to Test Data

Test Data


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?





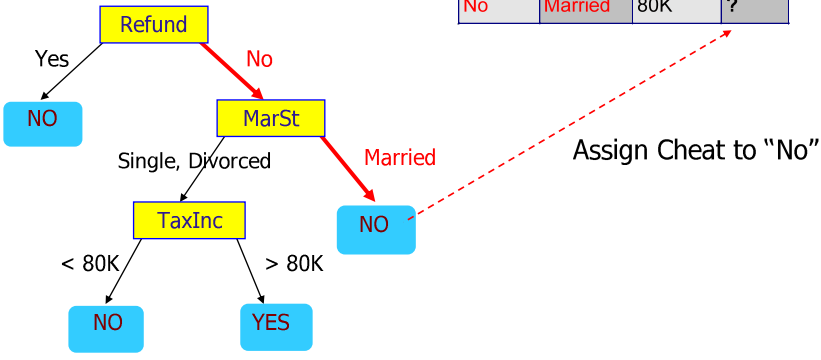
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006


14



# Apply Model to Test Data


Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?






From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

15



# Decision tree induction


- Many algorithms to build a decision tree
  - Hunt's Algorithm (one of the earliest)
  - CART
  - ID3, C4.5, C5.0
  - SLIQ, SPRINT



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

16





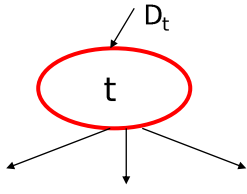
General structure of Hunt's algorithm


Basic steps

- If  $D_t$  contains records that belong to 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
  - select the "best" attribute  $A$  on which to split  $D_t$  and label node  $t$  as  $A$
  - split  $D_t$  into smaller subsets and recursively apply the procedure to each subset
- If  $D_t$  is an empty set
  - then  $t$  is a leaf node labeled as the default (majority) class,  $y_d$

ID	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	80K	Yes
6	No	Married	80K	No
7	Yes	Divorced	220K	No
8	No	Single	80K	Yes
9	No	Married	75K	No
10	No	Single	80K	Yes


$D_t$ , set of training records that reach a node  $t$






From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

17




Decision tree induction

- Adopts a greedy strategy
  - "Best" attribute for the split is selected locally at each step
    - not a global optimum
- Issues
  - Structure of test condition
    - Binary split versus multiway split
  - Selection of the best attribute for the split
  - Stopping condition for the algorithm




18




Structure of test condition

- Depends on attribute type
  - nominal
  - ordinal
  - continuous
- Depends on number of outgoing edges
  - 2-way split
  - multi-way split




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

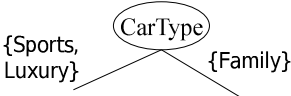
19



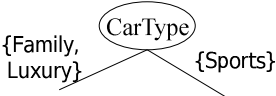
Splitting on nominal attributes


- Multi-way split
  - use as many partitions as distinct values
- Binary split
  - Divides values into two subsets
  - Need to find optimal partitioning






OR





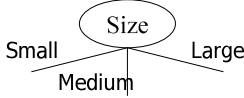
From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

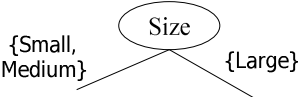
20



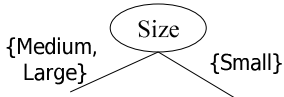
## Splitting on ordinal attributes


- **Multi-way split**
  - use as many partitions as distinct values
- **Binary split**
  - Divides values into two subsets
  - Need to find optimal partitioning






OR






What about this split?



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

21




## Splitting on continuous attributes

- Different techniques
  - **Discretization** to form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – discretize during tree induction


Ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering

- **Binary decision** ( $A < v$ ) or ( $A \geq v$ )
  - consider all possible splits and find the best cut
  - more computationally intensive



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

22



Splitting on continuous attributes

Taxable  
Income  
> 80K?


YesNo

(i) Binary split

Taxable  
Income?


< 10K[10K,25K][25K,50K][50K,80K]> 80K

(ii) Multi-way split



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

23



Selection of the best attribute

Before splitting:10 records of class 0,10 records of class 1

Own  
Car?

YesNo

C0: 6C1: 4

C0: 4C1: 6

Car  
Type?

FamilySportsLuxury

C0: 1C1: 3

C0: 8C1: 0


C0: 1C1: 7

Student  
ID?

$c_1$  $c_{10}$  $c_{11}$  $c_{20}$

C0: 1C1: 0...C0: 1C1: 0C0: 0C1: 1...C0: 0C1: 1

Which attribute (test condition) is the best?




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

24

DataBase and Data Mining Group

12



Selection of the best attribute


- Attributes with *homogeneous* class distribution are preferred
- Need a measure of node impurity

C0: 5  
C1: 5

Non-homogeneous,  
high degree of impurity


C0: 9  
C1: 1

Homogeneous, low  
degree of impurity




From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

25



Measures of node impurity

- Many different measures available
  - Gini index
  - Entropy
  - Misclassification error
- Different algorithms rely on different measures



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

26



## Decision Tree Based Classification

### ■ Advantages

- Inexpensive to construct
- Extremely fast at classifying unknown records
- Easy to interpret for small-sized trees
- Accuracy is comparable to other classification techniques for many simple data sets

### ■ Disadvantages

- accuracy may be affected by missing data



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

27

## Associative classification



*Politecnico di Torino*



## Associative classification

- The classification model is defined by means of association rules

$(Condition) \rightarrow y$

- rule body is an itemset
- Model generation
  - Rule selection & sorting
    - based on support, confidence and correlation thresholds
  - Rule pruning
    - Database coverage: the training set is covered by selecting topmost rules according to previous sort



29



## Associative classification


- Strong points
  - interpretable model
  - higher accuracy than decision trees
    - correlation among attributes is considered
  - efficient classification
  - unaffected by missing data
  - good scalability in the training set size
- Weak points
  - rule generation may be slow
    - it depends on support threshold
  - reduced scalability in the number of attributes
    - rule generation may become unfeasible



30


Neural networks

DBG



Data Base and Data Mining Group of Politecnico di Torino

Politecnico di Torino





Neural networks

■ Inspired to the structure of the human brain

■ Neurons as elaboration units

■ Synapses as connection network






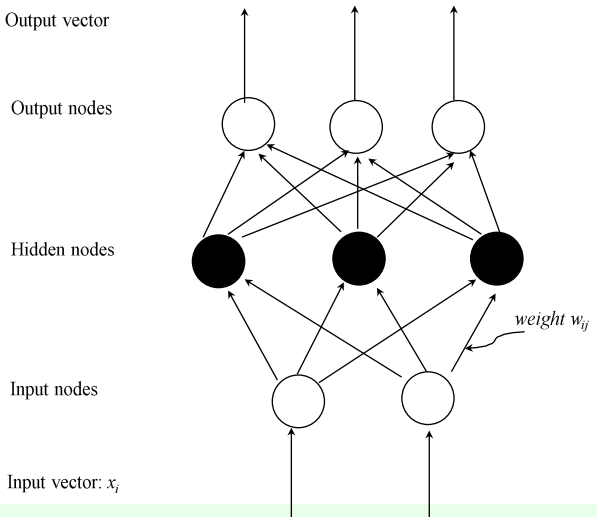
DBG


32






Structure of a neural network



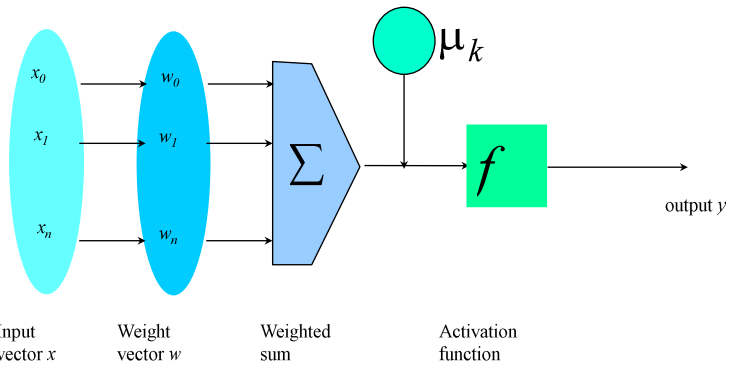



From: Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006

33



Structure of a neuron





From: Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006

34



## Construction of the neural network

- For each node, definition of
  - set of weights
  - offset valueproviding the highest accuracy on the training data
- Iterative approach on training data instances



35



## Neural networks

- Strong points
  - High accuracy
  - Robust to noise and outliers
  - Supports both discrete and continuous output
  - Efficient during classification
- Weak points
  - Long training time
    - weakly scalable in training data size
    - complex configuration
  - Not interpretable model
    - application domain knowledge cannot be exploited in the model



36

# Bayesian Classification



*Politecnico di Torino*



## Bayes theorem

- Let  $C$  and  $X$  be random variables

$$P(C, X) = P(C|X) P(X)$$

$$P(C, X) = P(X|C) P(C)$$

- Hence

$$P(C|X) P(X) = P(X|C) P(C)$$

- and also

$$P(C|X) = P(X|C) P(C) / P(X)$$



38



## Bayesian classification

- Let the class attribute and all data attributes be random variables
  - $C$  = any class label
  - $X = \langle x_1, \dots, x_k \rangle$  record to be classified
- Bayesian classification
  - compute  $P(C|X)$  for all classes
    - probability that record  $X$  belongs to  $C$
  - assign  $X$  to the class with *maximal*  $P(C|X)$
- Applying Bayes theorem
 
$$P(C|X) = P(X|C) \cdot P(C) / P(X)$$
  - $P(X)$  constant for all  $C$ , disregarded for maximum computation
  - $P(C)$  a priori probability of  $C$ 

$$P(C) = N_C / N$$



39




## Bayesian classification

- How to estimate  $P(X|C)$ , i.e.  $P(x_1, \dots, x_k|C)$ ?
- Naïve hypothesis
 
$$P(x_1, \dots, x_k|C) = P(x_1|C) P(x_2|C) \dots P(x_k|C)$$
  - *statistical independence* of attributes  $x_1, \dots, x_k$
  - not always true
    - model quality may be affected
- Computing  $P(x_k|C)$ 
  - for discrete attributes
 
$$P(x_k|C) = |x_{kC}| / N_C$$
    - where  $|x_{kC}|$  is number of instances having value  $x_k$  for attribute  $k$  and belonging to class  $C$
  - for continuous attributes, use probability distribution
- Bayesian networks
  - allow specifying a subset of dependencies among attributes




40




## Bayesian classification: Example

Outlook	Temperature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N



From: Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006

41




## Bayesian classification: Example

outlook	
$P(\text{sunny} \text{p}) = 2/9$	$P(\text{sunny} \text{n}) = 3/5$
$P(\text{overcast} \text{p}) = 4/9$	$P(\text{overcast} \text{n}) = 0$
$P(\text{rain} \text{p}) = 3/9$	$P(\text{rain} \text{n}) = 2/5$
temperature	
$P(\text{hot} \text{p}) = 2/9$	$P(\text{hot} \text{n}) = 2/5$
$P(\text{mild} \text{p}) = 4/9$	$P(\text{mild} \text{n}) = 2/5$
$P(\text{cool} \text{p}) = 3/9$	$P(\text{cool} \text{n}) = 1/5$
humidity	
$P(\text{high} \text{p}) = 3/9$	$P(\text{high} \text{n}) = 4/5$
$P(\text{normal} \text{p}) = 6/9$	$P(\text{normal} \text{n}) = 2/5$
windy	
$P(\text{true} \text{p}) = 3/9$	$P(\text{true} \text{n}) = 3/5$
$P(\text{false} \text{p}) = 6/9$	$P(\text{false} \text{n}) = 2/5$


$P(\text{p}) = 9/14$

$P(\text{n}) = 5/14$




From: Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006

42



## Bayesian classification: Example



- Data to be labeled  
 $X = \langle \text{rain, hot, high, false} \rangle$
- For class p  
$$\begin{aligned} P(X|p) \cdot P(p) &= \\ &= P(\text{rain}|p) \cdot P(\text{hot}|p) \cdot P(\text{high}|p) \cdot P(\text{false}|p) \cdot P(p) \\ &= 3/9 \cdot 2/9 \cdot 3/9 \cdot 6/9 \cdot 9/14 = 0.010582 \end{aligned}$$
- For class n  
$$\begin{aligned} P(X|n) \cdot P(n) &= \\ &= P(\text{rain}|n) \cdot P(\text{hot}|n) \cdot P(\text{high}|n) \cdot P(\text{false}|n) \cdot P(n) \\ &= 2/5 \cdot 2/5 \cdot 4/5 \cdot 2/5 \cdot 5/14 = \textcolor{red}{0.018286} \end{aligned}$$



From: Han, Kamber, "Data mining: Concepts and Techniques", Morgan Kaufmann 2006

43

# Model evaluation



Data Base and Data Mining Group of Politecnico di Torino

*Politecnico di Torino*



## Model evaluation

- Methods for performance evaluation
  - Partitioning techniques for training and test sets
- Metrics for performance evaluation
  - Accuracy, other measures
- Techniques for model comparison
  - ROC curve



45



## Methods of estimation

- Partitioning labeled data in
  - training set for model building
  - test set for model evaluation
- Several partitioning techniques
  - holdout
  - cross validation
- Stratified sampling to generate partitions
  - without replacement
- Bootstrap
  - Sampling with replacement



46



## Holdout

- Fixed partitioning
  - reserve 2/3 for training and 1/3 for testing
- Appropriate for large datasets
  - may be repeated several times
    - repeated holdout



47




## Cross validation

- Cross validation
  - partition data into  $k$  disjoint subsets (i.e., folds)
  - $k$ -fold: train on  $k-1$  partitions, test on the remaining one
    - repeat for all folds
  - reliable accuracy estimation, not appropriate for very large datasets
- Leave-one-out
  - cross validation for  $k=n$
  - only appropriate for very small datasets



48





## Metrics for model evaluation

- Evaluate the predictive accuracy of a model
- Confusion matrix
  - binary classifier


ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
Class=Yes	a	b
Class=No	c	d

a: TP (true positive)

b: FN (false negative)


c: FP (false positive)

d: TN (true negative)



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

49




## Accuracy


- Most widely-used metric for model evaluation

$$\text{Accuracy} = \frac{\text{Number of correctly classified objects}}{\text{Number of classified objects}}$$

- Not always a reliable metric



50




## Accuracy

- For a binary classifier


ACTUAL CLASS	PREDICTED CLASS	
	Class=Yes	Class=No
	<div>Class=Yes</div> <div>a (TP)</div>	<div>b (FN)</div>
<div>Class=No</div> <div>c (FP)</div>	<div>d (TN)</div>	

$$\text{Accuracy} = \frac{a + d}{a + b + c + d} = \frac{TP + TN}{TP + TN + FP + FN}$$



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

51




## Limitations of accuracy

- Consider a binary problem
  - Cardinality of Class 0 = 9900
  - Cardinality of Class 1 = 100
- Model

$() \rightarrow \text{class } 0$

  - Model predicts everything to be class 0
    - accuracy is  $9900/10000 = 99.0\%$
- Accuracy is misleading because the model does not detect any class 1 object



52



## Limitations of accuracy

- Classes may have different importance
  - Misclassification of objects of a given class is more important
  - e.g., ill patients erroneously assigned to the healthy patients class
- Accuracy is not appropriate for
  - unbalanced class label distribution
  - different class relevance



53



## Class specific measures

- Evaluate separately for each class

$$\text{Recall (r)} = \frac{\text{Number of objects correctly assigned to C}}{\text{Number of objects belonging to C}}$$

$$\text{Precision (p)} = \frac{\text{Number of objects correctly assigned to C}}{\text{Number of objects assigned to C}}$$

- Maximize

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



54



## Class specific measures

- For a binary classification problem
  - on the confusion matrix, for the positive class

$$\text{Precision (p)} = \frac{a}{a + c}$$

$$\text{Recall (r)} = \frac{a}{a + b}$$

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



From: Tan, Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

55