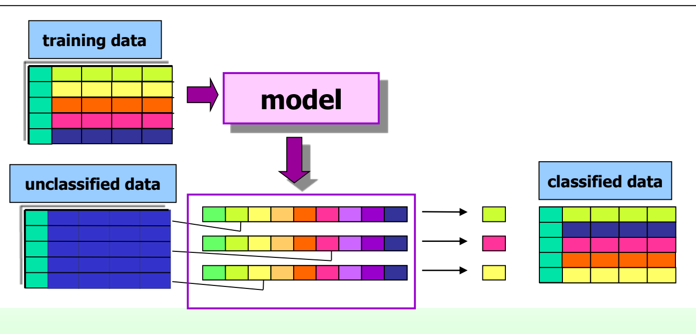
Classification fundamentals

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

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

Classification techniques

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

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

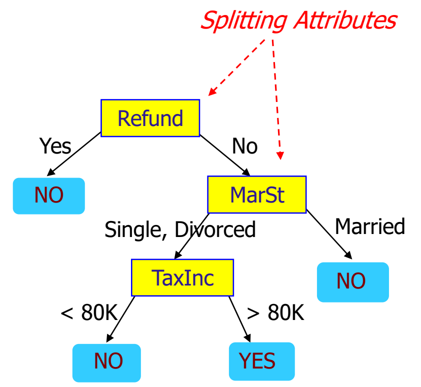
Decision trees

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



Model: Decision Tree

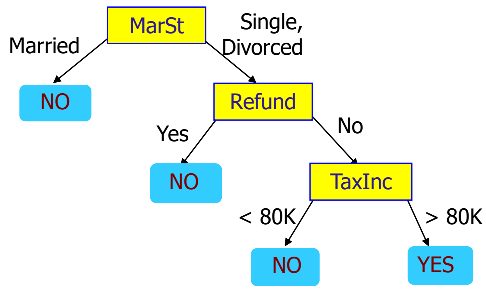
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

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



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

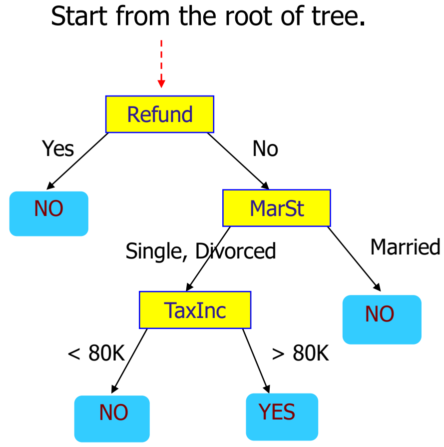
DBMG

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

9

Apply Model to Test Data

Start from the root of tree.



Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?

DBMG

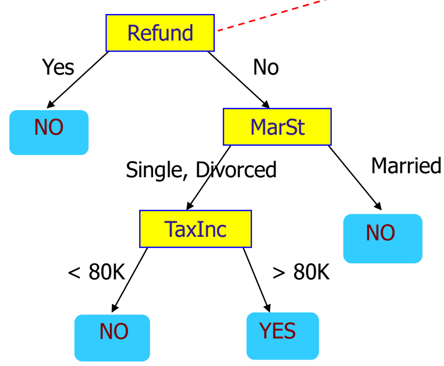
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

10

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



DBMG

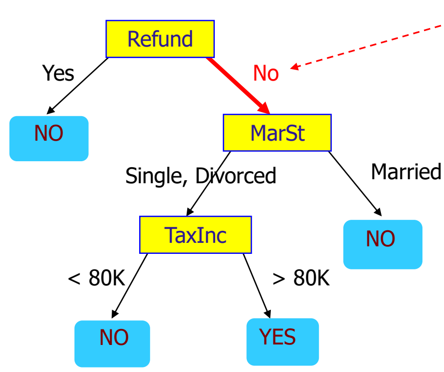
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

11

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



DBMG

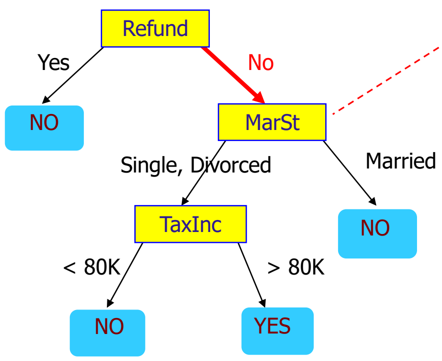
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

12

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



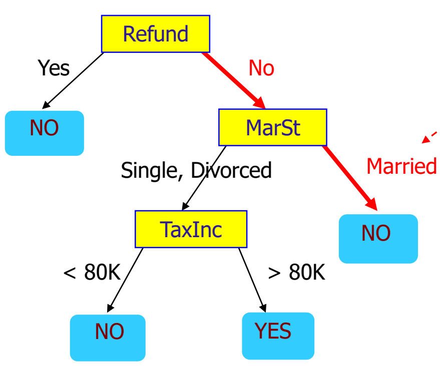
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

13

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



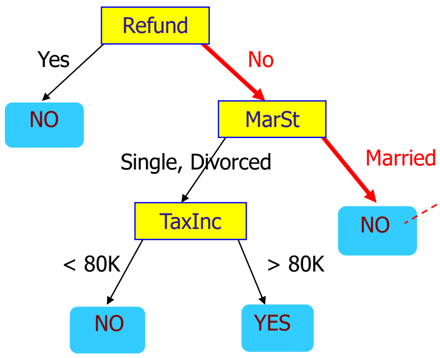
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

14

Apply Model to Test Data

Test Data

Refund	Marital Status	Taxable Income	Cheat
No	Married	80K	?



Assign Cheat to "No"

DBMG

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

DBMG

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

16

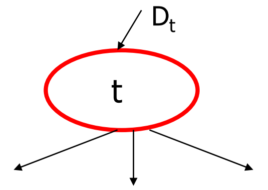
General structure of Hunt's algorithm

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

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

D_t, set of training records that reach a node t



DMG

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

DMG

18

Structure of test condition

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

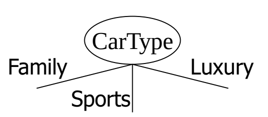
DBMG

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

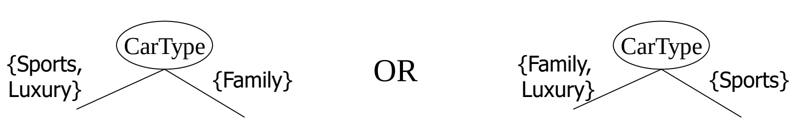
19

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



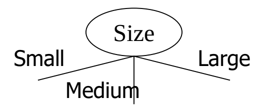
DBMG

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

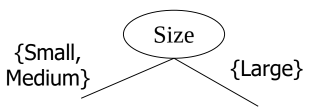
20

Splitting on ordinal attributes

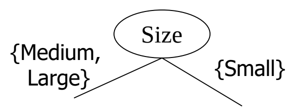
- Multi-way split
  - use as many partitions as distinct values



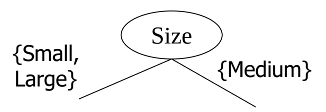
- Binary split
  - Divides values into two subsets
  - Need to find optimal partitioning



OR



What about this split?



DBMG

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

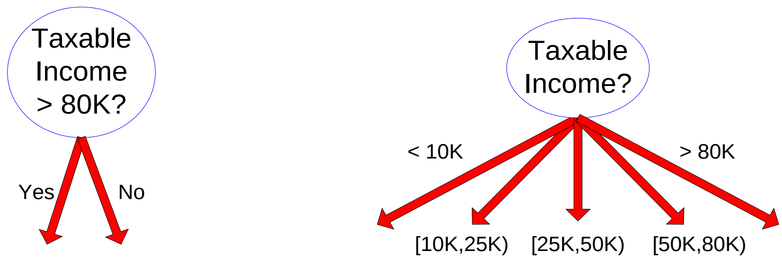
    Ranges can be found by equal interval bucketing, equal frequency bucketing (percentiles), or clustering
  - Binary decision $(A < v)$ or $(A \geq v)$
    - consider all possible splits and find the best cut
    - more computationally intensive

DBMG

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

22

Splitting on continuous attributes



(i) Binary split

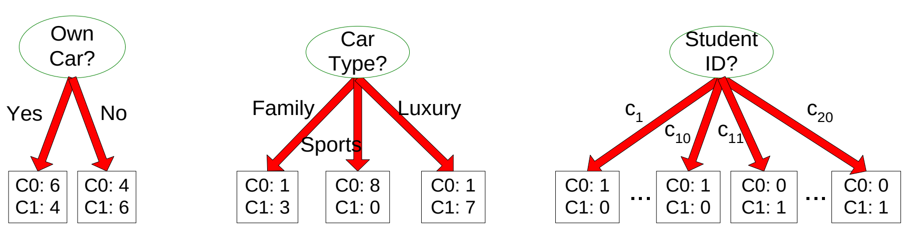
(ii) Multi-way split

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



Which attribute (test condition) is the best?

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

24

Selection of the best attribute

- Attributes 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

DBMG

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

DBMG

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

DBMG

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

27

Associative classification

DBMG

Data Base and Data Mining Group of Politecnico di Torino

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

DBMG 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

DBMG 30

Neural networks

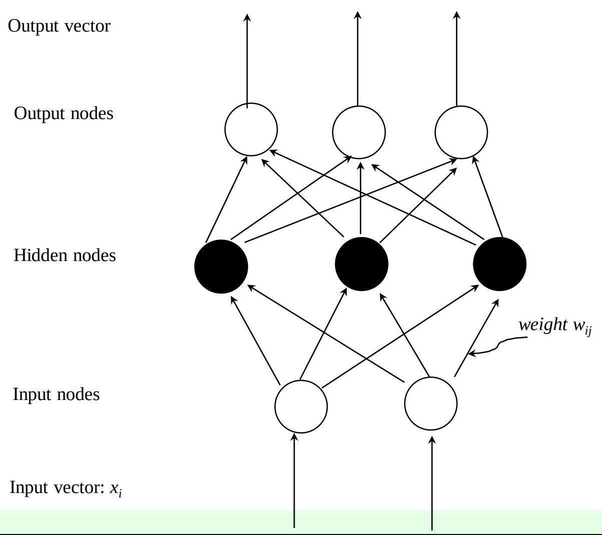
Data Base and Data Mining Group of Politecnico di Torino

Politecnico di Torino

Neural networks

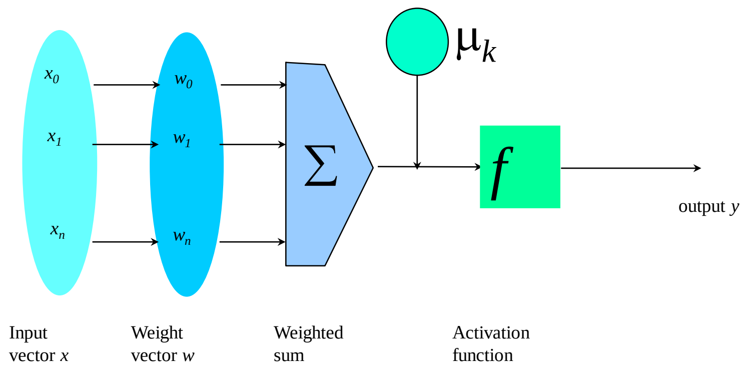
- Inspired to the structure of the human brain
  - Neurons as elaboration units
  - Synapses as connection network

Structure of a neural network



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

Structure of a neuron



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

Construction of the neural network

- For each node, definition of
  - set of weights
  - offset value

  providing the highest accuracy on the training data
- Iterative approach on training data instances

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

Bayesian Classification

Data Base and Data Mining Group of Politecnico di Torino

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
  - $X = <x_1,\ldots,x_k>$ 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

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

DMBG 39

Bayesian classification

- How to estimate $P(X|C)$, i.e. $P(x_1,\ldots,x_k|C)$?
- Naïve hypothesis

$$P(x_1,\ldots,x_k|C) = P(x_1|C)\, P(x_2|C) \ldots P(x_k|C)$$

  - *statistical independence* of attributes $x_1,\ldots,x_k$
  - 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

DMBG 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

DMBG

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

41

Bayesian classification: Example

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

P(p) = 9/14

P(n) = 5/14

DMBG

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

42

Bayesian classification: Example

- Data to be labeled
  X = <rain, hot, high, false>
- For class p
  P(X|p)·P(p) =
  = P(rain|p)·P(hot|p)·P(high|p)·P(false|p)·P(p)
  = 3/9·2/9·3/9·6/9·9/14 = 0.010582
- For class n
  P(X|n)·P(n) =
  = P(rain|n)·P(hot|n)·P(high|n)·P(false|n)·P(n)
  = 2/5·2/5·4/5·2/5·5/14 = 0.018286

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

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

Holdout

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

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

Metrics for model evaluation

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

	PREDICTED CLASS	
ACTUAL 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

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

Accuracy

- For a binary classifier

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

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

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

DBMG 51

Limitations of accuracy

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

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

  - 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

DBMG 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

DBMG 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

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

DBMG 54

Class specific measures

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

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

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

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

DBMG

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

55