Recap – 3rd Lecture

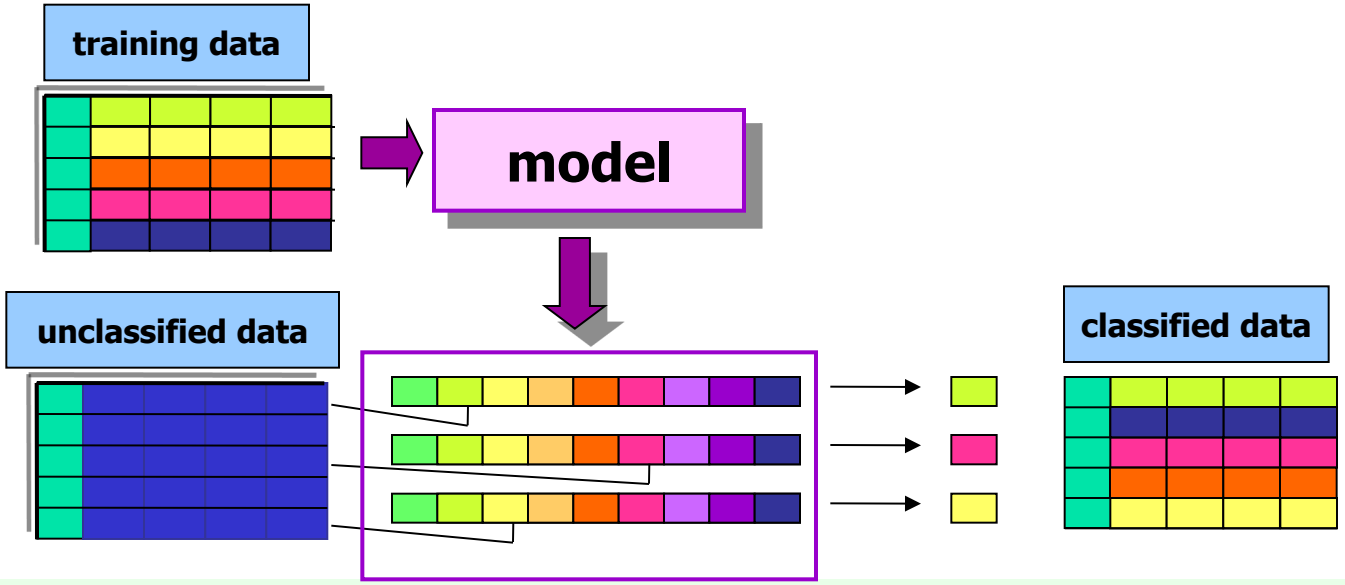
Data Base and Data Mining Group of Politecnico di Torino

Danilo Giordano

Politecnico di Torino

Classification

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



Evaluation of classification techniques

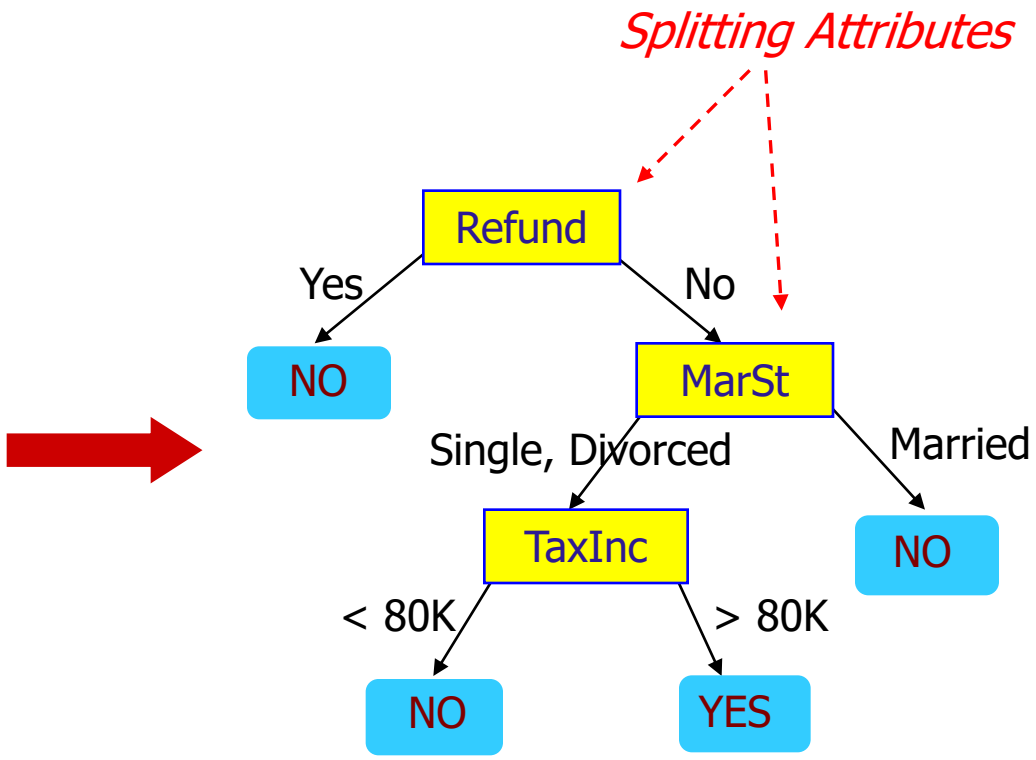
- Accuracy
  - quality of the prediction
- Efficiency
  - model building time
  - classification time
- Scalability
  - training set size
  - attribute number
- Robustness
  - noise, missing data
- Interpretability
  - model interpretability
  - model compactness

Example of decision tree

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

categorical categorical continuous class

Training Data

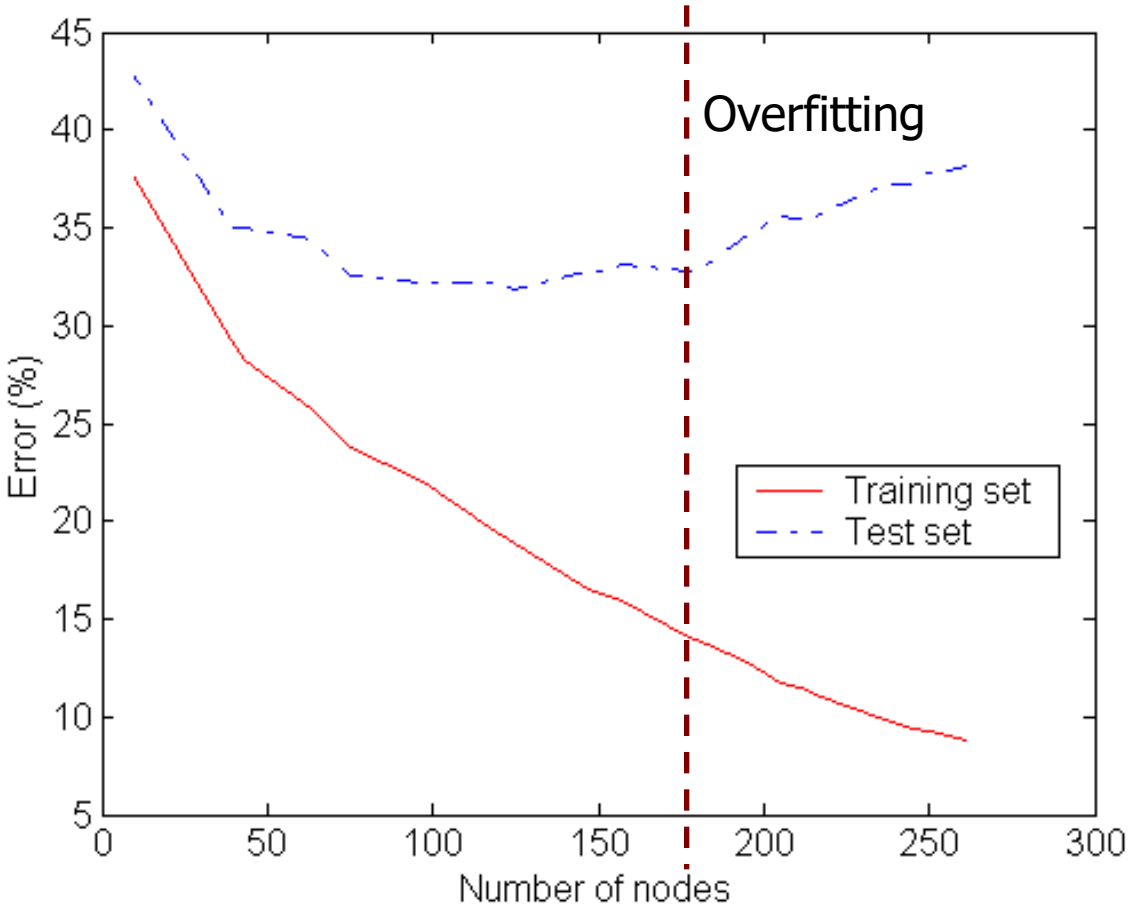


Model: Decision Tree

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

Underfitting and Overfitting



Underfitting: when model is too simple, both training and test errors are large

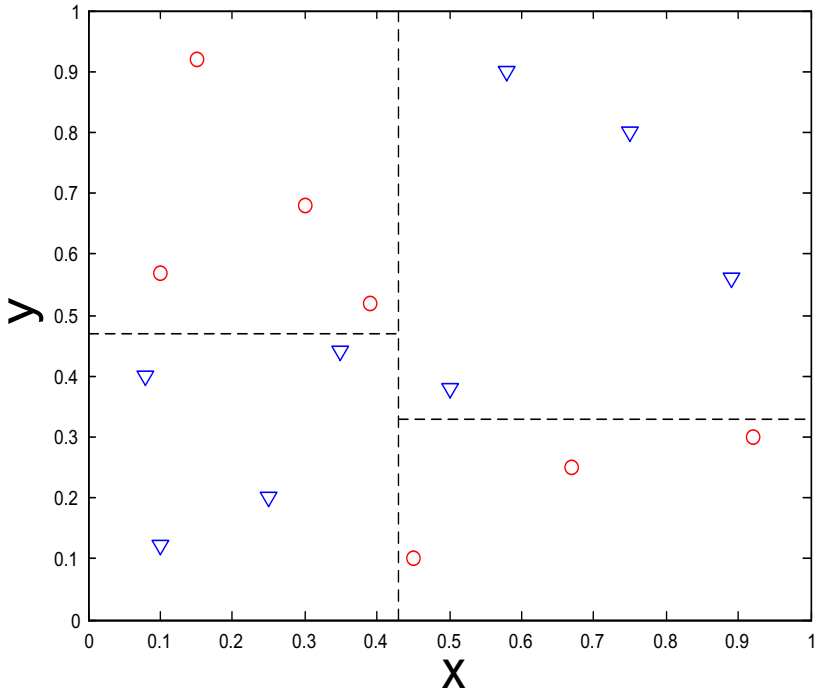
From: Tan,Steinbach, Kumar, Introduction to Data Mining, McGraw Hill 2006

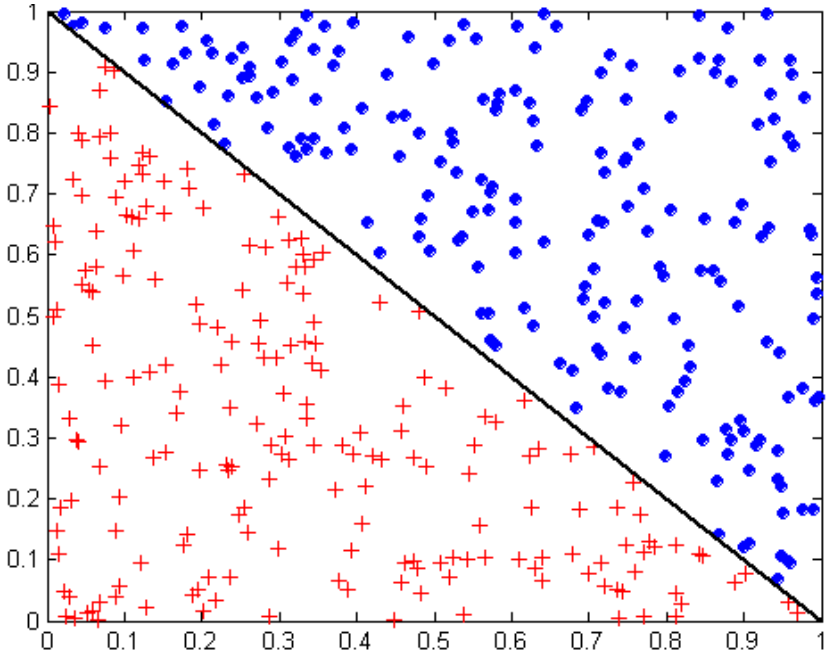
How to address overfitting

- Pre-Pruning (Early Stopping Rule)
  - Stop the algorithm before it becomes a fully-grown tree
- Post-pruning
  - Grow decision tree to its entirety
  - Trim the nodes of the decision tree in a bottom-up fashion

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

Decision boundary





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

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

Rule-based classifier

- Classify records by using a collection of "if…then…" rules
- Rule: (*Condition*) $\rightarrow$ *y*
  - (Taxable Income < 50K) $\wedge$ (Refund=Yes) $\rightarrow$ Cheat=No
- Characteristics of rules
- *Mutually exclusive* rules
  - Two rule conditions can't be true at the same time
- *Exhaustive* rules
  - Classifier rules account for every possible combination of attribute values

Associative classification

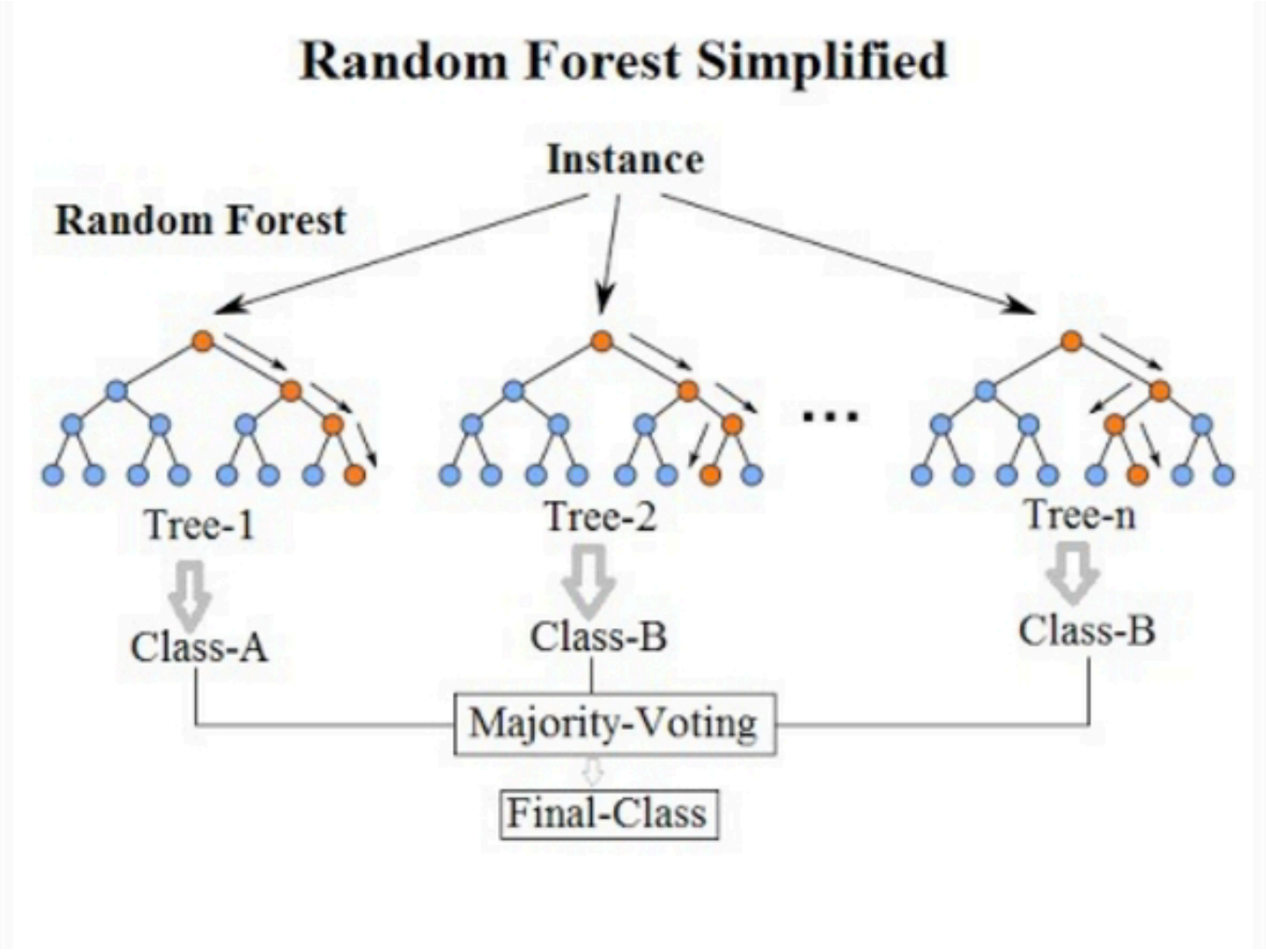
- Classify records by using a collection of "if…then…" rules
- Rule: (*Condition*) $\rightarrow y$
  - (Taxable Income < 50K) $\wedge$ (Refund=Yes) $\rightarrow$ Cheat=No
- Model generation – with permutation
  - Rule selection & sorting
    - based on support, confidence and correlation thresholds
  - Rule pruning
    - Database coverage: the training set is covered by selecting topmost rules according to previous sort

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

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

Random Forest Classifier



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

Random Forest Classifier

- Strong Points
  - High accuracy
  - Easily Scalable
  - Almost interpretable – Estimate important variable for classification
  - Handle Missing data
- Weak Points
  - May overfit with noisy datasets

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