Data Mining

CATEGORIZATION OF CLASSIFICATION ALGGORRITHMS

Prof. Dr. Hikmat Ullah Khan
Department of Information Technology

UNIVERSITY OF SARGODHA, SARGODHA

Classification of Classification Algo

- Probability based Classifier
  - Naive Bayes
  - Gaussian Naïve Bayes
  - Multinomial Naïve Bayes
- Decision Tree Based Classifier
  - C4.5/5.0
  - J48
  - Classification and Regression Tree (CART)
  - Iterative Dichotomiser 3 (ID3)
  - Chi-Squared Automatic Interaction Detection (CHAID)
  - Decision Stump
  - Conditional Decision Tree

Classification of Classification Algo

- Bayesian networks
  - Bayesian Network (BN)
  - Bayesian Belief Network (BNN)
- Neural networks
  - Perceptron
  - Back-Propagation
  - Hopfield Network
  - Radial Basis Function Network (RBFN)
- Instance based Classifier
  - K-Nearest Neighbor
  - Self-Organizing Map (SOM)
  - Learning Vector Quantization (LVQ)
  - Locally Weighted Learning (LWL)

Classification of Classification Algo

- Kernel based Classifier
  - SVM
  - Variations of SVM
- Ensemble
  - Random Forest
  - Gradient Boosting Machine (GBM)
  - Boosting
  - Bootstrapped Aggregation (Bagging)
  - AdaBoost
  - Stacked Generalization
- Genetic algorithms
- Fuzzy classification

Naïve Bayes

- Advantages:
  - Easy to implement
  - Efficient
  - Good results obtained in many applications
  - Good results in Text data
- Disadvantages
  - Interconnection between Features can not be learnt
  - Feature /Dimensions Reduction is an issue

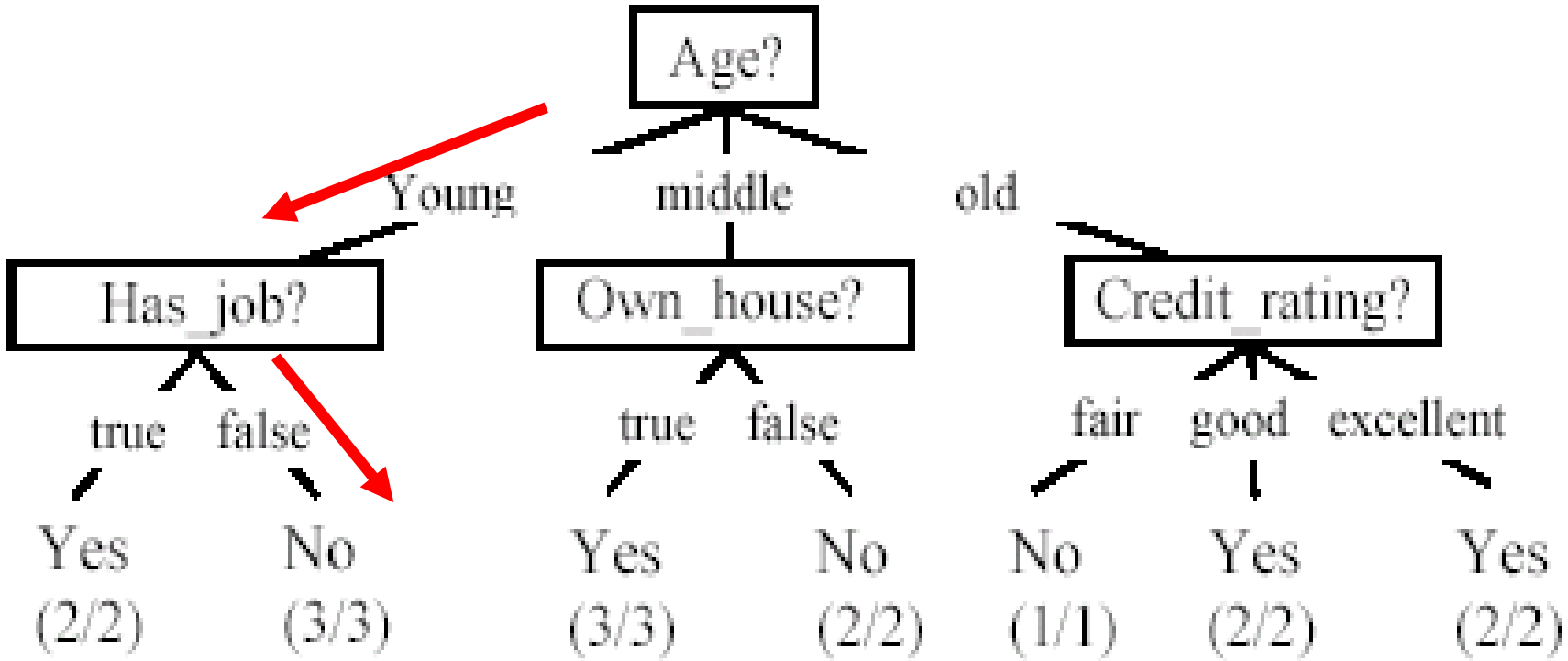
Decision Tree

- Widely used techniques
  - accuracy is competitive with other methods
  - efficient.
- classification model is a tree, called decision tree.
- Advantages
  - Easy to interpret and explain
  - Fast and scalable,
- Disadvantages
  - don’t support **online learning,** so you have to rebuild your tree when new examples come on
  - Another disadvantage is that they easily overfit

Use the decision tree

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

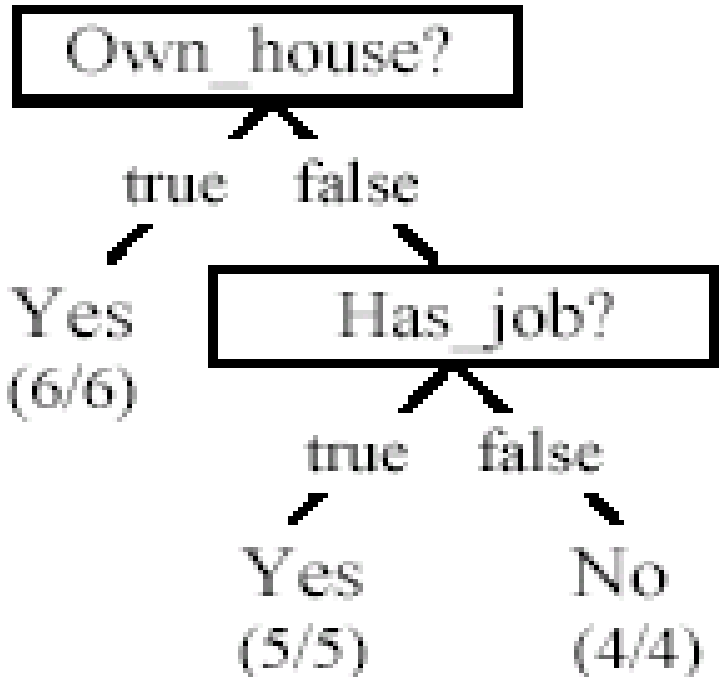
No



From a decision tree to a set of rules

A decision tree can be converted to a set of rules

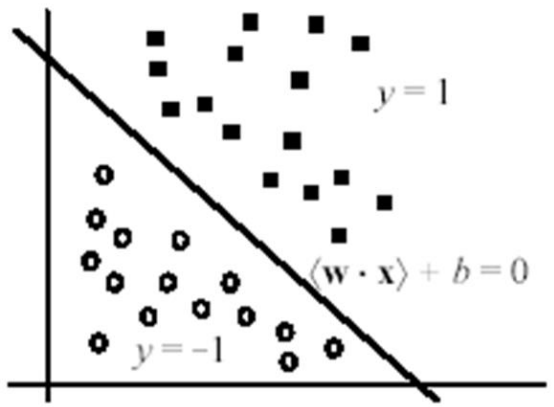
Each path from the root to a leaf is a rule.

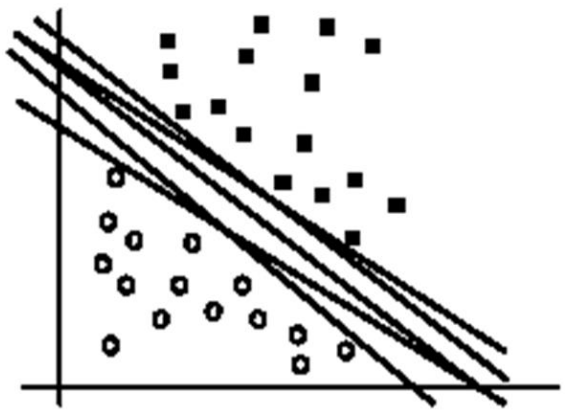


Own_house = true → Class =Yes [sup=6/15, conf=6/6]
Own_house = false, Has_job = true → Class = Yes [sup=5/15, conf=5/5]
Own_house = false, Has_job = false → Class = No [sup=4/15, conf=4/4]

SVM (Support Vector Machine)

- SVMs are linear classifiers that find a hyperplane to separate two class of data, positive and negative.
- Kernel functions are used for nonlinear separation.
- More accurate than most other methods in applications, especially for high dimensional data.
- It is perhaps the best cla





SVM

- Advantages:
  - More accurate than most other methods in applications, especially for high dimensional data.
  - Very Efficient
  - Perhaps the best classifier for text classification.
- Disadvantages
  - Memory-intensive,
  - hard to interpret, (Hyperplane is hard to understand by human users. Commonly used in applications that do not required human understanding.)
  - SVM works only in a real-valued space. For a categorical attribute, we need to convert its categorical values to numeric values.

k-Nearest Neighbor Classification (kNN)

- Unlike all the previous learning methods, kNN does not build model from the training data.
- For given D data values, To classify a test instance *d*, define *k*-neighborhood *P* as *k* nearest neighbors of *d*

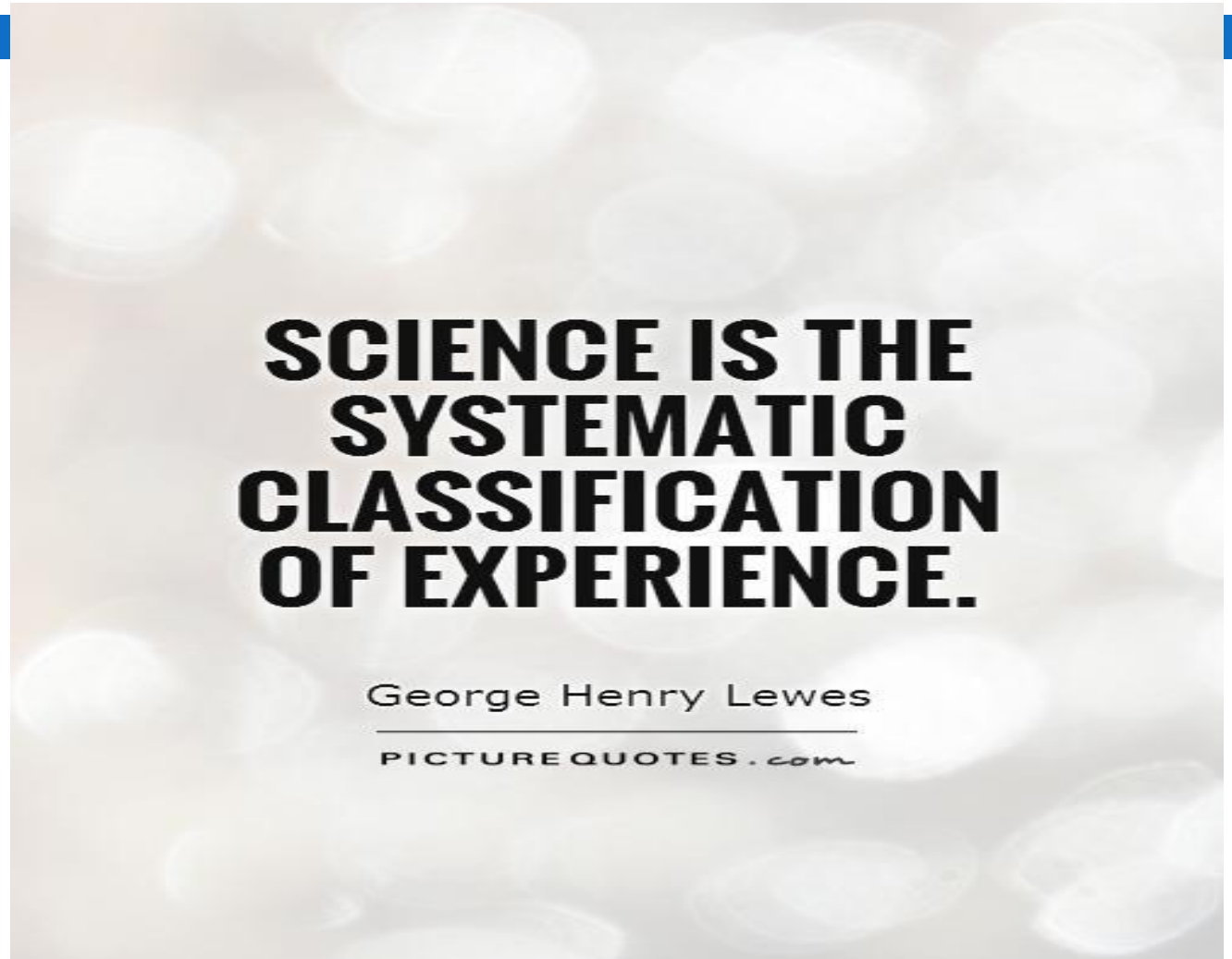
Algorithm kNN(D, d, k)

1 Compute the distance between d and every example in D;
2 Choose the k examples in D that are nearest to d, denote the set by P ($\subseteq D$);
3 Assign d the class that is the most frequent class in P (or the majority class);

KNN

- Advantages:
  - kNN can deal with complex and arbitrary decision boundaries.
  - accuracy of kNN is good in many cases as accurate as other methods.
- Disadvantages
  - kNN is slow at the classification time
  - kNN does not produce an understandable model

Algorithms	Problem Type	Training speed	Prediction speed	Amount of parameter tuning needed (excluding feature selection)
KNN	Either	Lower	Fast	Depends on n
Linear regression	Regression	Lower	Fast	Fast
Logistic regression	Classification	Lower	Fast	Fast
Naive Bayes	Classification	Lower	Fast (excluding feature extraction)	Fast
Decision trees	Either	Lower	Fast	Fast
Random Forests	Either	Higher	Slow	Moderate
AdaBoost	Either	Higher	Slow	Fast
Neural networks	Either	Higher	Slow	Fast


SCIENCE IS THE SYSTEMATIC CLASSIFICATION OF EXPERIENCE.
George Henry Lewes
PICTUREQUOTES.com