Artificial Intelligence

Inductive Learning Algorithms

Outline

- Supervised Learning — Classification
  - Training Dataset Format
  - Information based learning
  - Distance based learning
  - Probability based learning
  - Deep learning
- Association — Discover Association Rules
- Unsupervised Learning — Cluster Analysis

Data Design

- Data collection is formed as a table.
- Each row in the table represents one instance of the prediction subject—the phrase one-row-per-subject is often used to describe this structure.
- Each row is composed of a number of attributes/features that capture the basic characteristics of an instance.
- An attribute/feature is a property or characteristic of an instance that may vary, either from one instance to another or from one time to another.
- One of the attributes is designated as the target feature. The rest of the attributes are descriptive features.

Information Based Learning

- Information based machine learning algorithms try to build predictive models using only the most informative features.
- In this context an informative feature is a descriptive feature whose values split the instances in the dataset into homogeneous sets with respect to the target feature value.
- Model Representation:
    - Expert systems
    - Decision trees

Decision Tree

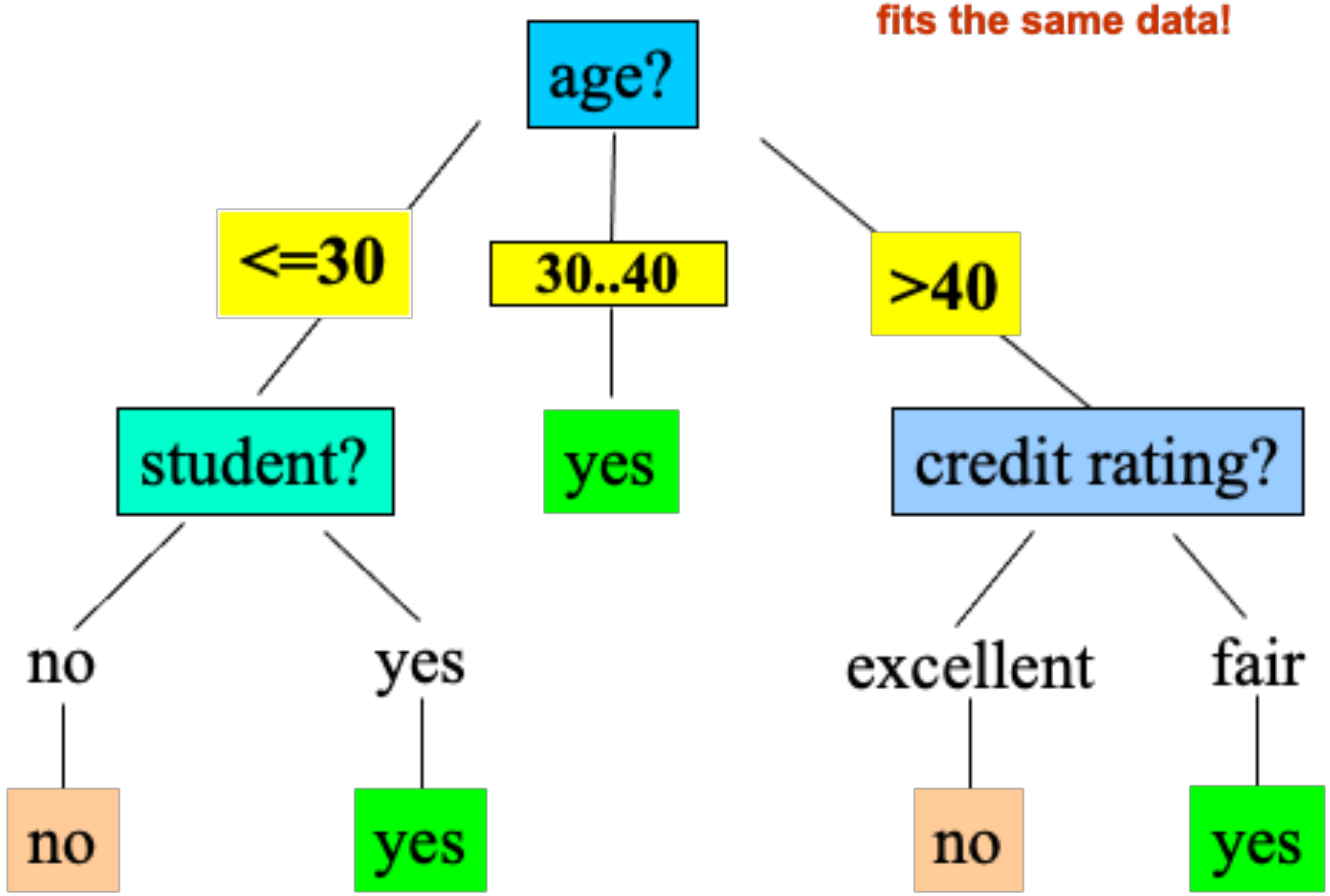
- A decision tree consists of:
    - a root node (or starting node),
    - interior nodes,
    - and leaf nodes (or terminating nodes).
- Each of the non-leaf nodes (root and interior) in the tree specifies a test to be carried out on one of the query's descriptive features.
- Each of the leaf nodes specifies a predicted classification for the query.

An Example of Training Dataset

Age	Income	Student	Credit_rating	Buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

Sample Decision Tree

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



Advantages and Limitations

- Simple to understand and interpret
- Uses a white box model
- Performs well with large datasets

- Prone to overfitting
- Not suitable for some concepts, such as XOR
- The problem of learning an optimal decision tree is known to be NP-complete.

Select the Attribute

- Use greedy algorithms
- Apply to building decision tree:
  - In each step, choose the attribute that seems to be the “best”
  - “best” -- the attribute that most likely splits the dataset into pure sets with respect to the target feature
  - Result: shallower trees
- Computational metric of the purity of a set — Entropy

Entropy (I)

- Claude Shannon's entropy model defines a computational measure of the impurity of the elements of a set.
- An easy way to understand the entropy of a set is to think in terms of the uncertainty associated with guessing the result if you were to make a random selection from the set.
- Entropy is related to the probability of an outcome:
    - High probability — Low entropy
    - Low probability — High entropy

Entropy (II)

- Shannon's model of entropy is a weighted sum of the logs of the probabilities of each of the possible outcomes when we make a random selection from a set.

- Entropy at a given node t:
  $Entropy(t) = -\Sigma_j p(j|t) log(p(j|t))$
  (NOTE: p(j | t) is the relative frequency of class j at node t).

  - Measures homogeneity of a node.

  - Maximum (log n_c) when records are equally distributed among all classes, implying least information

  - Minimum (0.0) when all records belong to one class, implying most information

Information Gain

- Information Gain:

$$GAIN_{split} = Entropy(t) - (\Sigma_{i=1}^{k} \frac{n_i}{n} Entropy(i))$$

  - Parent Node t is split into k partitions
  - n_i is number of records in partition i

- Information gain measures reduction in entropy achieved because of the split.
- Greedy algorithm (such as ID3) chooses the split that achieves the most reduction.
- Disadvantage: tends to prefer splits that result in large number of partitions, each being small but pure.

Practical Issues with Decision Trees

- Overfitting --- splitting the data on an irrelevant feature
  - Pre-pruning
  - Post-pruning
- Under-fitting
- Missing value in training data
- Missing value in query

Similarity Based Learning

- The fundamentals of similarity-based learning are:
  - Feature space
    - An abstract n-dimensional space that is created by taking each of the descriptive features in a training data set to be the axes of a reference space and each instance in the dataset is mapped to a point in the feature space based on the values of its descriptive features.
  - Similarity metrics
    - Measures the similarity between two instances according to a feature space.

Metric

- Mathematically, a metric must conform to the following four criteria:
  - Non-negativity: metric(a, b) >= 0
  - Identity: metric(a, b) = 0 <==> a = b
  - Symmetry: metric(a, b) = metric(b, a)
  - Triangular Inequality:
    metric(a, b) <= (metric(a, c) + metric(c, b)
    Where metric(a, b) is a function that returns the distance (or dissimilarity) between two instances a and b.

Common Metric

- Hamming (Manhattan) distance (p = 1)
- Euclidean distance (p = 2)
- Minkowski distance in a feature space with m descriptive features:

$$Minkowski(a, b) = (\Sigma_{i=1}^{m} abs(a[i] - b[i])^{p})^{\frac{1}{p}}$$

- The larger the value of p, the more emphasis is placed on the features with large differences in values because there differences are raised to the power of p.

The Nearest Neighbour Algorithm

- Require: set of training instances
  and a query to be classified

- Algorithm:

  - Iterate across the instances and find the instance that is shortest distance from the query position in the feature space.

  - Make a prediction for the query equal to the value of the target feature of the nearest neighbour.

Advantages vs Disadvantages

- It is a instance-based learning algorithm
    - Store training examples and delay the processing ("lazy evaluation") until a new instance must be classified.
- It is easy to add new data items into the training dataset to update the model.

- Supervised machine learning is based on the stationarity assumption which states that the data doesn't change – remains stationary – over time.
- In the context of classification, supervised machine learning creates models that distinguish between the classes that are present in the dataset they are induced from.
- So if a classification model is trained to distinguish between lions, frogs and ducks, the model will classify a query as being either a lion, a frog or a duck; even if the query is actually an elephant.

Probability Based Learning

- We can use estimates of likelihoods to determine the most likely prediction that should be made.
- More importantly, we revise these predictions based on data we collect and whenever extra evidence becomes available.
- Bayes' Theorem

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- Example:
  A patient has tested positive for a serious disease. The test is 99% accurate. However, the disease is extremely rare, striking only 1 in 10,000 people. What is the actual probability that the patient has the disease?

Advantages vs Disadvantages

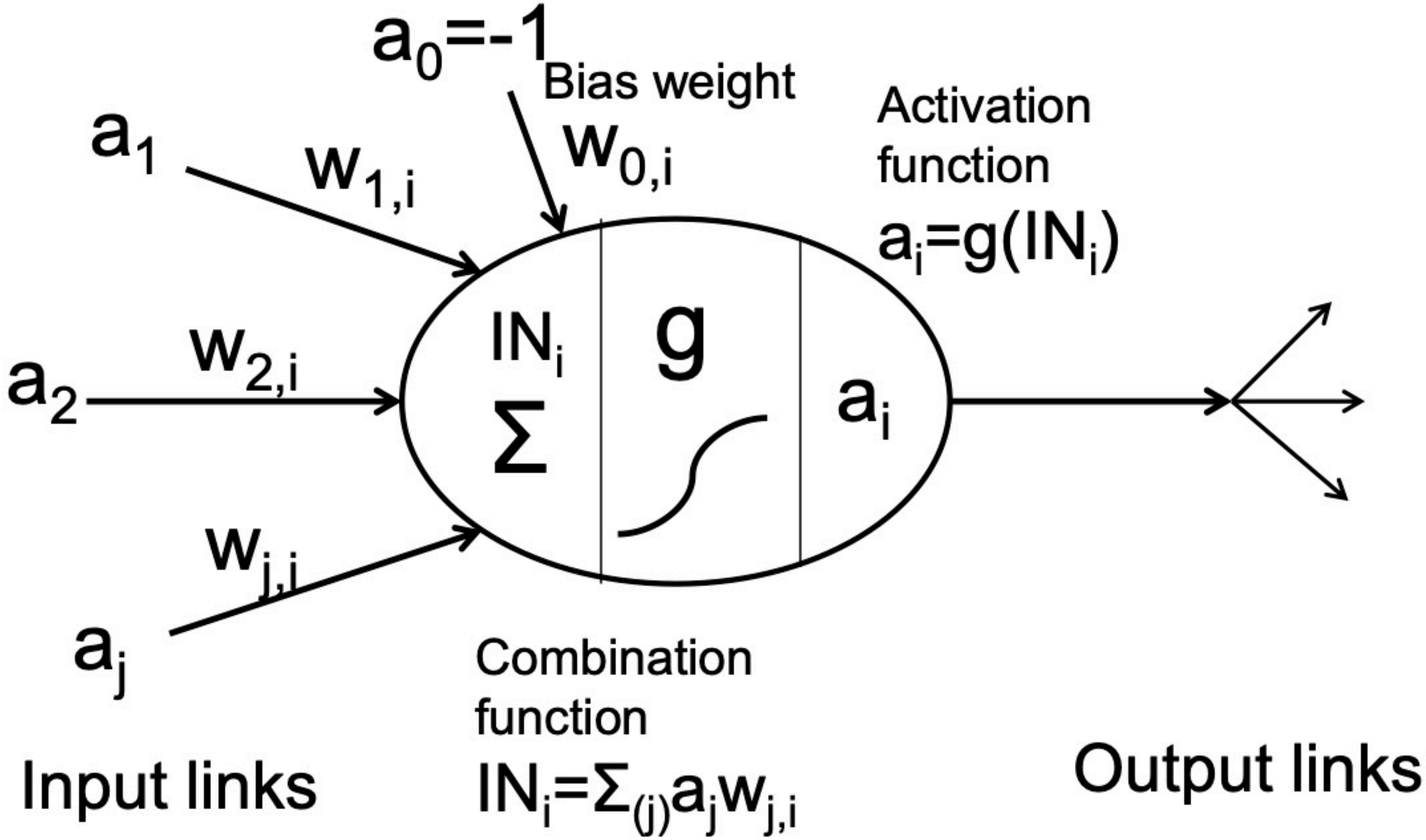
- Incremental
- Probabilistic prediction

- Practical difficulty — require initial knowledge of many probabilities, significant computational cost
- If dataset is not large enough, model is over-fitting to the training data.

Deep Learning — Artificial Neural Network

- Simulate human brain
- Typical human brain has
  - 10^11 neurons of 20+ types,
  - 10^14 synapses
  - 1ms to 10ms cycle time
- Signals are noisy “spike trains” of electrical potential.

Modelling a Neuron



Modelling Neuron Networks

- Consists nodes and edges.
- Node takes input and triggers other nodes through connections.
- Each node has an activation function to decide whether to fire up.
- Each edge not only permits to transfer the value, but also has a weight.
- Artificial neural network simulates the brain.
- Artificial neural network is abstract and media independent. We can use parallel circuits or execute a program on a serial processor.

Forward Application and Back-propagation Learning

- Forward Application:
  Feed forward propagation of input pattern signals through network to result outputs

- Back-propagation Learning:
  computes error signal, propagates the error backwards through network, adjusting the weight where the actual and desired output values are different

Advantages vs Disadvantages

- Very powerful — With sigmoidal activation functions, it can be shown that a three-layer ANN can approximate any continuous function to arbitrary accuracy.

- Learning is simply adjusting edge weights

- Overfitting — Memorizing training data instead of learning knowledge.

- An ANN is a blackbox.

Association and Training Dataset Format

- Discover Association Rules:
  - Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrences of other items in the transaction
- Training Dataset Format
  - A large set of transactions
  - Each transaction is a list of items
  - An itemset is a collection of one or more items
- Association Rule — An implication expression of the form X → Y, where X and Y are itemsets
  - Rule form: “Body → Head [support, confidence]”
  - Example: buys(x, “diapers”) → buys(x, “beers”) [0.5%, 60%]

Terminologies

- Support count (σ)
  - Frequency of occurrence of an itemset; E.g. σ({Milk, Bread,Diaper}) = 2
- Support
  - Fraction of transactions that contain an itemset; E.g. σ({Milk, Bread, Diaper}) = 2/5
- Frequent Itemset
  - An itemset whose support is greater than or equal to a min-support threshold
- Support (S) of an association rule X → Y
  - Fraction of transactions that contain both X and Y
- Confidence (C) of an association rule X → Y
  - Measures how often items in Y appear in transactions that contain X
- Interest (I)
  - The interest of an association rule X → Y is the absolute value of the amount by which the confidence differs from the probability of Y

Association Algorithms

- Brute-force approach — computationally prohibitive
  - List all possible association rules
  - Compute the support and confidence for each rule
  - Prune rules that fail the min-support and min-confidence thresholds
- Two-step approach:
  - Frequent Itemset Generation — Generate all itemsets whose support is greater than min-support
    - Brute-force algorithm — Computationally expensive
    - A-Priori Algorithm and FP Growth Algorithm
  - Rule Generation — Generate high confidence rules from each frequent itemset, where each rule is a binary partitioning of a frequent itemset

Cluster Analysis

- Cluster: a collection of data objects
    - Similar to one another within the same cluster
    - Dissimilar to the objects in other clusters
- Cluster analysis
    - Grouping a set of data objects into clusters
    - Intra-cluster distances are minimized
    - Inter-cluster distances are maximized
- Clustering is unsupervised classification: no predefined class labels

Common Clustering Algorithms

- Partitioning algorithms
  - K-means and its variants
- Hierarchy algorithms
- Density-based algorithms
- Grid-based algorithms
- Model-based algorithms

Summary

- The main objective of inductive learning:
  to capture the relationships among data's features from observing the behaviour of a large collection of data objects.
- A model learned by induction is not guaranteed to be correct.
- Learning can't occur unless the learning process is biased in some way.
- There is not one best approach that always outperforms the others in learning in general and in machine learning in particular.
- Key tasks in building an inductive learning process
  - Become situationally fluent so that we can converse with experts in the application domain
  - Collect as much relevant data as possible
  - Explore the data to understand it correctly
  - Spend time cleaning and organizing the data
  - Think hard about the best ways to represent features
  - Spend time designing the evaluation process correctly