CSE/ECE 848 Introduction to Evolutionary Computation

Module 1 - Lecture 4 - Part 1

Machine Learning Basics

Wolfgang Banzhaf, CSE
John R. Koza Chair in Genetic Programming

ML

- High-level commonalities among ML Systems
  - Face similar task: How to learn from the experience in the environment?
- Implementation differences among ML systems
  - Many ML systems differ radically in how they learn from the environment

The Learning Domain

- Machine Learning is a process that starts with the identification of the learning domain and ends with testing and using the results of the learning
- ML systems are usually applied to a “learning domain”
- A learning domain is any problem or set of facts where the researcher is able to identify “features” of the domain that may be measured, and a result the researcher would like to predict
- Once the features (inputs) are chosen from the learning domain, they define the overall dimension of the environment that the ML system will experience and from which it will hopefully learn.

Training

- Training sets, training data
  - Researcher must choose specific past examples from the learning domain
  - Each example should contain data that represent one instance of the relationship between the chosen features (inputs) and the classes (outputs)
- Training
  - An ML system goes through the training set and attempts to learn from the examples

Testing/Generalization

- One way to appraise the quality of learning is to test the ability of the best solution of the ML system to predict outputs from a "test set"
- A test set is comprised of inputs and outputs from the same learning domain, but contains different examples
- The ability of the system to predict the outputs of the test set is often referred to as "generalization"

An Example: Iris Classification

- Data base of Iris data, with three classes of iris
- The goal in training is to take the sepal and petal measurements (the features) of the training set and learn to predict which of the three classes a particular iris belongs to

Input 1. Sepal length in cm.

Input 2. Sepal width in cm.

Input 3. Petal length in cm.

Input 4. Petal width in cm.

- Classic test, due to Fisher, 1936
- It then has to properly predict for a test set in a statistically significant manner

Major Issues in Machine Learning

- We can classify ML algorithms by how they answer the following four questions about the "how" of learning

1. How are solutions represented in the algorithm?
2. Is the learning supervised or unsupervised?
3. What search operators does the learning algorithm use to move in/through the solution space?
4. What type of search is conducted?

Representing the problem

- What is the Problem Representation?
  - The representation of the problem defines the space of candidate solutions an ML system can find for a particular problem

  - There are three different levels:

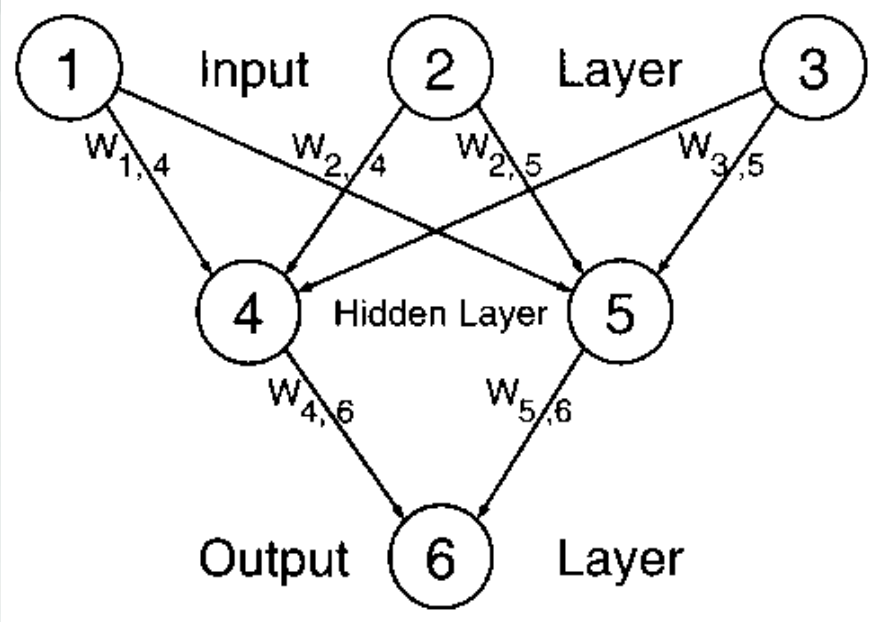
1. Representation of the input and output set
2. Representation of the set of concepts the computer may learn
3. Interpretation of the learned concepts as outputs

Boolean Representations

- Conjunctive Boolean Representation
  - The system uses the Boolean AND function to join features together into concepts and outputs
- Disjunctive Boolean Representation
  - The system uses the Boolean OR function to join features together into concepts and output

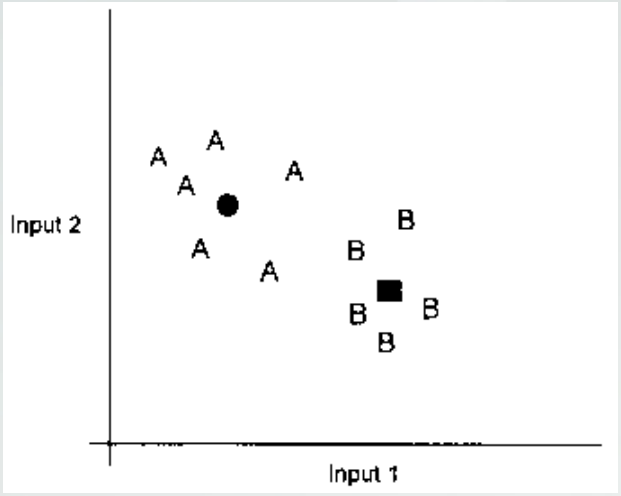
Threshold Representation

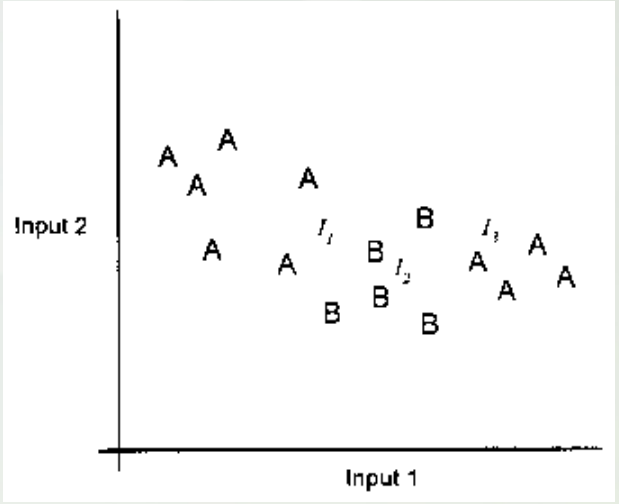
- Numeric threshold representation to represent a system
- A threshold unit may appear as an input, a concept, or an interpreter in a ML system
- Feedforward multilayer neural network



Case-based Representations

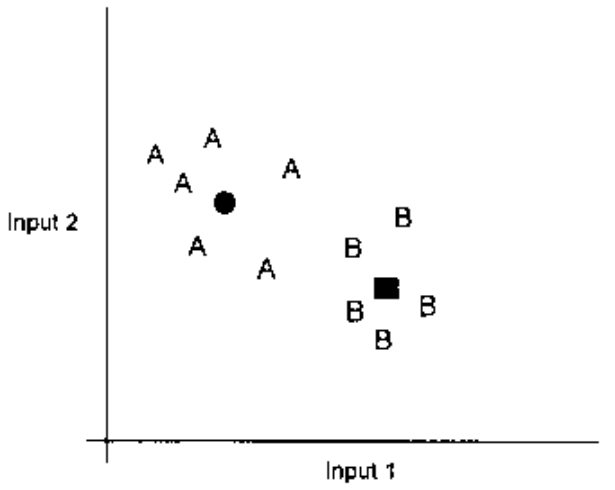
- Store training instances as representations of classes or store general descriptions of classes by averaging training instances in some way
- K-means nearest neighbour method
- Bayes/Parzen classification

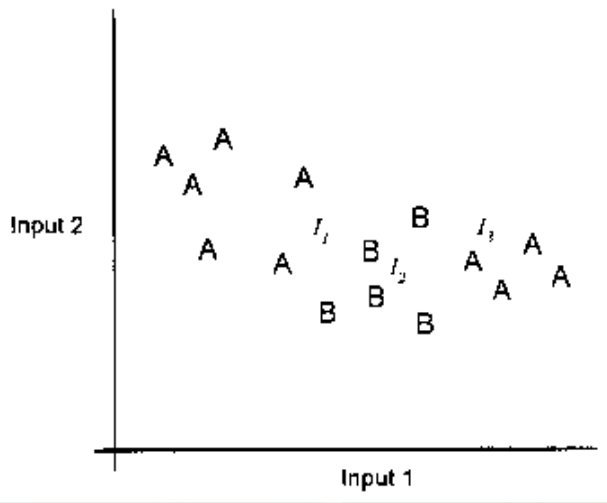




Linear Separability

- Domains that are “linearly separable” are the easiest to work with





Other Representations

- Tree Representations
  - Many problem space representations are based on decision trees (ID3)
  - The ID3 learning algorithm chooses the best feature for each new node by sorting the training set by the attributes of each potential feature

- Genetic Representations, here the GA
  - A genetic algorithm GA has a fixed length binary string
  - Each bit is assigned a meaning by the researcher
  - The bit string is the concept definition language for the GA and the meaning assigned to the bits is analogous to the interpreter

What kind of learning?

- Supervised learning
  - Each training instance is an input accompanied by the correct output
  - Many EC applications use supervised learning
- Unsupervised learning
  - The ML system is not told what the correct output is
  - Cluster algorithms are good examples of unsupervised learning
- Reinforcement learning
  - RL falls between supervised and unsupervised learning
  - Many fitness functions in EC are more complex than just comparing output to desired output