Data Mining

- Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules
- Common data mining tasks
  - Classification
  - Estimation
  - Prediction
  - Affinity Grouping
  - Clustering
  - Description

Knowledge Discovery

- Directed and Undirected KD
- Directed KD
  - Purpose: Explain value of some field in terms of all the others
  - Method: We select the target field based on some hypothesis about the data. We ask the algorithm to tell us how to predict or classify it
  - Similar to hypothesis testing (e.g., in regression modeling) in statistics
- Undirected KD
  - Purpose: Find patterns in the data that may be interesting
  - Method: clustering, affinity grouping
  - Closest to ideas of machine learning in artificial intelligence
- Comparison
  - UKD helps us to recognize relationships & DKD helps us to explain them

Classification

- Classifying observations into different categories given characteristics

Estimation

- Rules that explain how to estimate a value given characteristics

Prediction

- Rules that explain how to predict a future value or classification, given characteristics

Affinity Grouping

- Grouping by relations (not by characteristics)
Clustering

- Segmenting a diverse population into more similar groups
- In clustering, there are no pre-defined classes and no examples. Records are grouped together by some similarity measure.

Scientific Models

- **Mechanistic models**
  - Mechanism
  - Predictive power
  - Elegance
  - Consistency

- **Stochastic models**
  - Black box
  - Predictive power

---

Artificial Intelligence in Biosciences

- **Neural Networks (NN)**
- **Genetic Algorithms (GA)**
- **Formal Grammars (FG)**

Artificial Intelligence in Biosciences

- **Neural Networks (NN)**
- **Genetic Algorithms (GA)**
- **Formal Grammars (FG)**

---

**Neural Networks**

- interconnected assembly of simple processing elements (units or nodes)
- nodes functionality is similar to that of the animal neuron
- processing ability is stored in the inter-unit connection strengths (weights)
- weights are obtained by a process of adaptation to, or learning from, a set of training patterns
Neural Networks

Perceptron

Output layer

Input layer

\[ Y = \begin{cases} 1 & \text{if } \sum w_{ji} > \Theta \\ 0 & \text{otherwise} \end{cases} \]

Learning process: \( \Delta w_i = (T_p - Y_p)i_p \)

Artificial Intelligence in Biosciences

Neural Networks (NN)

Genetic Algorithms (GA)

Formal Grammars (FG)

Genetic Algorithms

Search or optimization methods using simulated evolution.
Population of potential solutions is subjected to natural selection, crossover, and mutation

choose initial population
evaluate each individual's fitness
repeat
    select individuals to reproduce
    mate pairs at random
    apply crossover operator
    apply mutation operator
    evaluate each individual's fitness
until terminating condition

Genetic Algorithms
Crossover

Parent A
Parent B
Child AB
Child BA

crossover point

Mutation

Genetic Algorithms Applications

Parents
10 00 01 00 10
10 00 00 01 11

Children
10 00 10 01 11
10 00 01 00 10

GA simulation of folding

Artificial Intelligence in Biosciences

Neural Networks (NN)
Genetic Algorithms (GA)
Formal Grammars (FG)

Membrane binding domain of Blood Coagulation Factor VIII (J.Moult)

Grammars and Language

gram·mar  n.
1. the study of the way the sentences of a language are constructed...
2. Generative Gram. a device, as a body of rules, whose output is all of the sentences that are permissible in a given language, while excluding all those that are not permissible.

Random House Unabridged Dictionary

Language Components

Semantics (meaning)
Syntax (structure, form)

Language Syntax

Alphabet
Primitive elements
Letters, phonemes

Vocabulary
Elements composed from the alphabet
Words, phrases, sentences,…

Grammar
Legal composition of vocabulary
Rules, operators
Semantics

Derived from syntax
Semantic content derived from vocabulary within a context
Vocabulary element has its own meanings
dictionary lookup
meanings depending on context

Time flies like an arrow
Fruit flies like a banana

Formal Grammars

formal grammar
a means for specifying the syntactic structure of natural language by a set of transformation functions

Chomsky hierarchy (for string grammars)
type 0: phrase structure
type 1: context sensitive
(type 2: context free (SCFG))
type 3: regular (Hidden Markov models)

Chomsky, Syntactic Structures (1957)

Markov Model (or Markov Chain)

A  T  C  T  A  G

Probability for each character based only on several preceding characters in the sequence
# of preceding characters = order of the Markov Model

Probability of a sequence


Hidden Markov Models

States -- well defined conditions
Edges -- transitions between the states

ATGAC
ATTAC
ACGAC
ACTAC

Each transition assigned a probability.

Probability of the sequence:
single path with the highest probability --- Viterbi path
sum of the probabilities over all paths -- Baum-Welch method

Hidden Markov Models

Adopted from Anders Krogh, 1998
Hidden Markov Model for Exon and Stop Codon (VEIL Algorithm)

A Markov state

Hidden Markov Model in Structural Analysis

A hidden Markov model consists of Markov states connected by directed transitions. Each state emits an output symbol, representing sequence or structure. There are four categories of emission symbols in our model: b, d, r, and c, corresponding to amino acid residues, three-state secondary structure, backbone angles (discretized into regions of phi-psi space) and structural context (e.g. hairpin versus diverging turn, middle versus end-strand), respectively.

Hidden Markov Model in Structural Analysis

HMM topology from merging of two motifs, the extended Type-I hairpin motif and the Serine hairpin.

Adopted from S. Salzberg, 1997

Adopted from C. Bystroff et al., 2000

Adopted from C. Bystroff et al, 2000

JMB, 301, 173