26:198:722 Expert Systems

- Knowledge representation
- Knowledge acquisition
- Machine learning
- ID3 & C4.5

Knowledge Representation

Recall:

- Knowledge engineering
  - Knowledge acquisition
    - Knowledge elicitation
  - Knowledge representation
    - Production rules
    - Semantic networks
    - Frames

Knowledge Representation

- Representation is more than just encoding (encrypting)
- Coding preserves structural ambiguity
- Communication assumes prior knowledge
- Representation implies organization

Knowledge Representation

- Representation
  - A set of syntactic and semantic conventions that make it possible to describe things (Winston)
- Description
  - makes use of the conventions of a representation to describe some particular thing
- Syntax v. semantics

Knowledge Representation

- STRIPS
  - Predicate-argument expressions
    - at (robot, roomA)
  - World models
  - Operator tables
    - push (X, Y, Z)
      - Preconditions at (robot, Y), at (X, Y)
      - Delete list at (robot, Y), at (X, Y)
      - Add list at (robot, Z), at (X, Z)

Knowledge Representation

- STRIPS
  - maintained lists of goals
  - selected goal to work on next
  - searched for applicable operators
  - matched goals against formulas in add lists
  - set up preconditions as sub-goals
  - used means-end analysis

Knowledge Representation

- STRIPS - lessons
  - Heuristic search
  - Uniform representation
  - Problem reduction
- Procedural semantics

Knowledge Representation

- MYCIN
  - Assists physicians who are not experts in the field of antibiotics in treating blood infections
  - Consists of
    - Knowledge base
    - Dynamic patient database
    - Consultation program
    - Explanation program
    - Knowledge acquisition program

Knowledge Representation

- MYCIN
  - Production rules
    - Premises
      - Conjunctions of conditions
    - Actions
      - Conclusions or instructions
  - Patient information stored in context tree
  - Certainty factors for uncertain reasoning
  - Backward chaining control structure (based on AND/OR tree)

Knowledge Representation

- MYCIN
  - Evaluation
    - Panel of experts approved 72% of recommendations
    - Good as experts
    - Better than non-experts
    - Knowledge base incomplete (400 rules)
    - Required more computing power than available in hospitals
    - Doctors did not like the user interface

Knowledge Acquisition

- Stages
  - Identification
  - Conceptualization
  - Formalization
  - Implementation
  - Testing
- KADS
- Ontological analysis

Knowledge Acquisition

- Expert system shells
  - EMYCIN
  - TEIRESIAS
    - Rule models (meta-rules)
    - Schemas for data types
    - Domain-specific knowledge
    - Representation-specific knowledge
    - Representation-independent knowledge
    - Explain-Test-Review

Knowledge Acquisition

- Methods and tools
  - Structured interview
  - Unstructured interview
  - Case studies
    - Retrospective v. observational
    - Familiar v. unfamiliar
  - Concurrent protocols
    - Verbalization, “thinking aloud”
  - Tape recording
  - Video recording

Knowledge Acquisition

- Methods and tools
  - Automated knowledge acquisition
    - Domain models
    - Graphical interfaces
    - Visual programming language

Knowledge Acquisition

- Different types of knowledge
  - Procedural knowledge
    - Rules, strategies, agendas, procedures
  - Declarative knowledge
    - Concepts, objects, facts
  - Meta-knowledge
    - Knowledge about other types of knowledge and how to use them
  - Structural knowledge
    - Rules sets, concept relationships, concept to object relationships

Knowledge Acquisition

- Sources of knowledge
  - Experts
  - End-users
  - Multiple experts (panels)
  - Reports
  - Books
  - Regulations
  - Guidelines

Knowledge Acquisition

- Major difficulties with elicitation
  - Expert may
    - be unaware of the knowledge used
    - be unable to verbalize the knowledge used
    - provide irrelevant knowledge
    - provide incomplete knowledge
    - provide incorrect knowledge
    - provide inconsistent knowledge

Knowledge Acquisition

- "The more competent domain experts become, the less able they are to describe the knowledge they used to solve problems"
(Waterman)

Knowledge Acquisition

- Detailed guidelines for conducting structured and unstructured interviews and both retrospective and observational case studies are given in Durkin (Chapter 17)

Knowledge Acquisition

- Technique Capabilities

	Interviews		Case Studies			
			Retrospective		Observational	
Knowledge	Unstructured	Structured	Familiar	Unfamiliar	Familiar	Unfamiliar
Facts	Poor	Good	Fair	Average	Good	Excellent
Concepts	Excellent	Excellent	Average	Average	Good	Good
Objects	Good	Excellent	Average	Average	Good	Good
Rules	Fair	Average	Average	Average	Good	Excellent
Strategies	Average	Average	Good	Good	Excellent	Excellent
Heuristics	Fair	Average	Excellent	Good	Good	Poor
Structures	Fair	Excellent	Average	Average	Average	Average

Knowledge Acquisition

- Analyzing the knowledge collected
  - Producing transcripts
  - Interpreting transcripts
    - Chunking
  - Analyzing transcripts
    - Knowledge dictionaries
    - Graphical techniques
      - Cognitive maps
      - Inference networks
      - Flowcharts
      - Decision trees

Machine Learning

- Rote learning
- Supervised learning
  - Induction
    - Concept learning
    - Descriptive generalization
- Unsupervised learning

Machine Learning

- META-DENDRAL
  - RULEMOD
    - Removing redundancy
    - Merging rules
    - Making rules more specific
    - Making rules more general
    - Selecting final rules

Machine Learning

- META-DENDRAL
  - Version spaces
    - Partial ordering
    - Boundary sets
    - Candidate elimination algorithm

    - Monotonic, non-heuristic
    - Results independent of order of presentation
    - Each training instance examine only once
    - Discarded hypotheses never reconsidered
    - Learning is properly incremental

Machine Learning

- Decision trees and production rules
  - Decision trees are an alternative way of structuring rules
  - Efficient algorithms exist for constructing decision trees
  - There is a whole family of such learning systems:
    - CLS (1966)
    - ID3 (1979)
    - ACLS (1982)
    - ASSISTANT (1984)
    - IND (1990)
    - C4.5 (1993) - and C5.0
  - Decision trees can be converted to rules later

Machine Learning

- Entropy
  - Let X be a variable with states x_1 - - - x_n
  - Define the entropy of X by

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{p}\left(x_i\right)\log_2\left(\mathrm{p}\left(x_i\right)\right)$$

  - N.B. $\log_2\left(x\right) = \frac{\log_{10}\left(x\right)}{\log_{10}\left(2\right)} = \frac{\ln\left(x\right)}{\ln\left(2\right)}$

Machine Learning

- Entropy
  - Consider flipping a perfect coin:

    e.g., $n = 2$

    $X : x_1 , x_2$

    $\mathrm{p}(x_1) = \mathrm{p}(x_2) = 1/2$

Machine Learning

- Entropy

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{p}(x_i)\log_2\left(\mathrm{p}(x_i)\right)$$

$$= -\left[\frac{1}{2}\log_2\left(\frac{1}{2}\right) + \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]$$

$$= -\left[\frac{1}{2}(-1) + \frac{1}{2}(-1)\right] = 1$$

Machine Learning

- Entropy
  - Consider n equiprobable outcomes

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{p}(x_i)\log_2\left(\mathrm{p}(x_i)\right)$$

$$= -\sum_{i=1}^{n} \frac{1}{n}\log_2\left(\frac{1}{n}\right)$$

$$= \sum_{i=1}^{n} \frac{1}{n}\log_2(n) = \log_2(n)$$

Machine Learning

- Entropy
  - Consider flipping a totally biased coin:

    e.g., $n = 2$

    $X : x_1 , x_2$

    $p(x_1) = 1 \qquad p(x_2) = 0$

Machine Learning

- Entropy

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{p}(x_i)\log_2\left(\mathrm{p}(x_i)\right)$$

$$= -\left[\log_2(1) + 0\log_2(0)\right]$$

$$= -\left[0 + 0\log_2(0)\right] = 0$$

(by L'Hopital's rule)

Machine Learning

- Entropy
  - Entropy is a measure of chaos or disorder
  - $\mathrm{H}(X)$ is maximum for equiprobable outcomes

Machine Learning

- Entropy
  - X: x_1 - - - x_m and Y: y_1 - - - y_n be two variables

$$\mathrm{H}(X,Y) = -\sum_{i=1}^{m}\sum_{j=1}^{n}\mathrm{p}\left(x_i, y_j\right)\log_2\left(\mathrm{p}\left(x_i, y_j\right)\right)$$

  - If X and Y are independent

$$\mathrm{H}(X,Y) = \mathrm{H}(X) + \mathrm{H}(Y)$$

Machine Learning

- Conditional Entropy
  - Partial conditional entropy of Y given X is in state x_i :

$$\mathrm{H}(Y|x_i) = -\sum_{j=1}^{n} \mathrm{p}\left(y_j|x_i\right)\log_2\left(\mathrm{p}\left(y_j|x_i\right)\right)$$

  - Full conditional entropy of Y given X

$$\mathrm{H}(Y|X) = \sum_{i=1}^{m} \mathrm{p}\left(x_i\right)\cdot \mathrm{H}(Y|x_i)$$

Machine Learning

- Binary Logarithms

1	0.0000
2	1.0000
3	1.5850
4	2.0000
5	2.3219
6	2.5850
7	2.8074
8	3.0000

Machine Learning

- ID3
  - Builds a decision tree first, then rules
  - Given a set of attributes, and a decision, recursively selects attributes to be the root of the tree based on Information Gain:

    H(decision) - H(decision | attribute)
  - Favors attributes with many outcomes
  - Is not guaranteed to find the simplest decision tree
  - Is not incremental

Machine Learning

- C4.5
  - Selects attributes based on Information gain ratio:

    (H(decision) - H(decision | attribute)) / H(attribute)
  - Uses pruning heuristics to simplify decision trees
    - to simplify
    - to reduce dependence on training set
  - Tunes the resulting rule(s)

Machine Learning

- C4.5 rule tuning
  - Derive initial rules by enumerating paths through the decision tree
  - Generalize the rules by possibly deleting unnecessary conditions
  - Group rules according to target classes and delete any that do not contribute to overall performance on the class
  - Order the sets of rules for the target classes and choose a default class

Machine Learning

- Rule tuning
  - Rule tuning may be useful for rules derived by a variety of other means besides C4.5
    - Evaluate the contribution of individual rules
    - Evaluate the performance of the rule set as a whole

Machine Learning

- A data set for classification (Quinlan)

	-------------Attributes-------------			---Decision---
	<u>Height</u>	<u>Hair</u>	<u>Eyes</u>	<u>Attractiveness</u>
1	short	blond	blue	+
2	tall	blond	brown	-
3	tall	red	blue	+
4	short	dark	blue	-
5	tall	dark	blue	-
6	tall	blond	blue	+
7	tall	dark	brown	-
8	short	blond	brown	-

Machine Learning

- A data set for classification (Quinlan)
  - H(decision) = H(Attractiveness)

$$= -\frac{3}{8}\log_2\left(\frac{3}{8}\right)-\frac{5}{8}\log_2\left(\frac{5}{8}\right)=0.955$$

Machine Learning

- A data set for classification (Quinlan)
  - Height:
    - short: 1, 4, 8 p(+|short) = 1/3 p(-|short) = 2/3
    - tall: 2, 3, 5, 6, 7 p(+|tall) = 2/5 p(-|tall) = 3/5
  - H(decision|attribute) = H(Attractiveness|Height) =

$$\frac{3}{8}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right)-\frac{2}{3}\log_2\left(\frac{2}{3}\right)\right]+\frac{5}{8}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]=0.951$$

  - Information gain = $0.955 - 0.951 = 0.004$

Machine Learning

- A data set for classification (Quinlan)
  - Hair:
    - blond: 1, 2, 6, 8 — p(+|blond) = 2/4 — p(-|blond) = 2/4
    - red: 3 — p(+|red) = 1/1 — p(-|red) = 0/1
    - dark: 4, 5, 7 — p(+|dark) = 0/3 — p(-|dark) = 3/3
  - H(decision|attribute) = H(Attractiveness|Hair) =

$$\frac{4}{8}[1]+\frac{1}{8}[0]+\frac{3}{8}[0]=0.500$$

  - Information gain = $0.955 - 0.500 = 0.455$

Machine Learning

- A data set for classification (Quinlan)
  - Eyes:
    - blue: 1, 3, 4, 5, 6 p(+|blue) = 3/5 p(-|blue) = 2/5
    - brown: 2, 7, 8 p(+|brown) = 0/3 p(-|brown) = 3/3
  - H(decision|attribute) = H(Attractiveness|Eyes) =

$$\frac{5}{8}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]+\frac{3}{8}[0]=0.607$$

  - Information gain = $0.955 - 0.607 = 0.348$

Machine Learning

- A data set for classification (Quinlan)
  - Hence Hair is chosen as the best choice for the root of the tree
  - Now we recursively repeat this process for the (three) resulting branches
  - In this case, the branches for Hair: red and Hair: dark are already completely classified, and we need to work only on the sub-table for Hair: blond

Machine Learning

- A data set for classification (Quinlan)

	---------Attributes----------		---Decision---
	Height	Eyes	Attractiveness
1	short	blue	+
2	tall	brown	-
6	tall	blue	+
8	short	brown	-

- H(decision) = H(Attractiveness)

$$= -\frac{2}{4}\log_2\left(\frac{2}{4}\right)-\frac{2}{4}\log_2\left(\frac{2}{4}\right)=1$$

Machine Learning

- A data set for classification (Quinlan)
  - Height:
    - short: 1, 8 p(+|short) = 1/2 p(-|short) = 1/2
    - tall: 2, 6 p(+|tall) = 1/2 p(-|tall) = 1/2
  - H(decision|attribute) = H(Attractiveness|Height) =

$$\frac{2}{4}[1]+\frac{2}{4}[1]=1$$

  - Information gain = 1-1=0

Machine Learning

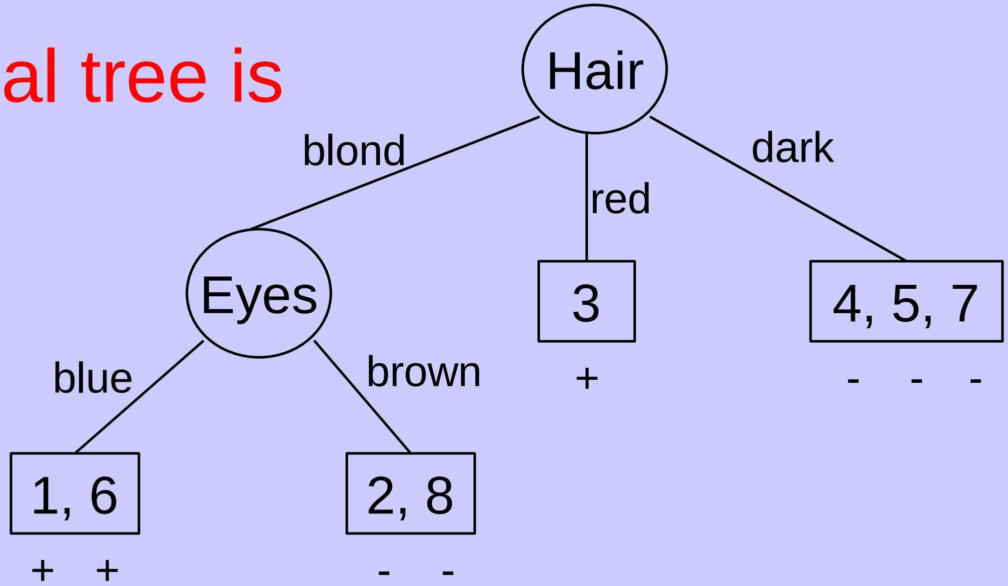
- A data set for classification (Quinlan)
  - Eyes:
    - blue: 1, 6 p(+|blue) = 2/2 p(-|blue) = 0/2
    - brown: 2, 8 p(+|brown) = 0/2 p(-|brown) = 2/2
  - H(decision|attribute) = H(Attractiveness|Eyes) =

$$\frac{2}{4}[0]+\frac{2}{4}[0]=0$$

  - Information gain = 1-0=1

Machine Learning

- A data set for classification (Quinlan)
  - Hence Eyes is chosen as the best root of this subtree
  - The final tree is



Machine Learning

- A data set for classification (Quinlan)
  - We may now build rules from this decision tree
    - R1: (Hair, dark) --> (Attractiveness, -)
    - R2: (Hair, red) --> (Attractiveness, +)
    - R3: (Hair, blond) & (Eyes, blue) --> (Attractiveness, +)
    - R4: (Hair, blond) & (Eyes, brown) --> (Attractiveness, -)
  - Note that height is irrelevant

Machine Learning

- A data set for classification (Quinlan)
  - Dropping conditions from rules
    - Rules 1 and 2 have only one condition
    - Rule 3: neither condition can be dropped (case 5 needs the first condition and case 2 needs the second condition)
    - Rule 4: we can drop the first condition
    - R4': (Eyes, brown) --> (Attractiveness, -)

Machine Learning

- A data set for classification (Quinlan)
  - Dropping conditions from rules
    - Linear
      - Scan rule left to right
      - Try to drop conditions one at a time
      - If possible, drop for good
      - Iterate (n conditions, n attempts)
    - Exponential
      - Scan rule left to right
      - Try to drop conditions one at a time
      - Then try to drop pairs, triples,etc. (n conditions, 2^{n}-2 attempts)

Machine Learning

- A data set for classification (Quinlan)
  - Now consider Information gain ratio
  - For initial root of tree we already know
  - H(decision) = H(Attractiveness)

$$= -\frac{3}{8}\log_2\left(\frac{3}{8}\right)-\frac{5}{8}\log_2\left(\frac{5}{8}\right)=0.955$$

Machine Learning

- A data set for classification (Quinlan)
  - H(decision|attribute) = H(Attractiveness|Height) =

$$\frac{3}{8}\left[-\frac{1}{3}\log_2\left(\frac{1}{3}\right)-\frac{2}{3}\log_2\left(\frac{2}{3}\right)\right]+\frac{5}{8}\left[-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right)\right]=0.951$$

  - H(decision|attribute) = H(Attractiveness|Hair) =

$$\frac{4}{8}[1]+\frac{1}{8}[0]+\frac{3}{8}[0]=0.500$$

  - H(decision|attribute) = H(Attractiveness|Eyes) =

$$\frac{5}{8}\left[-\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)\right]+\frac{3}{8}[0]=0.607$$

Machine Learning

- A data set for classification (Quinlan)
  - H(attribute) = H(Height) =

$$-\frac{3}{8}\log_2\left(\frac{3}{8}\right)-\frac{5}{8}\log_2\left(\frac{5}{8}\right)=0.955$$

  - H(attribute) = H(Hair) =

$$-\frac{4}{8}\log_2\left(\frac{4}{8}\right)-\frac{1}{8}\log_2\left(\frac{1}{8}\right)-\frac{3}{8}\log_2\left(\frac{3}{8}\right)=1.406$$

  - H(attribute) = H(Eyes) =

$$-\frac{5}{8}\log_2\left(\frac{5}{8}\right)-\frac{3}{8}\log_2\left(\frac{3}{8}\right)=0.955$$

Machine Learning

- A data set for classification (Quinlan)
  - Hence the Information gain ratios are
    - Height: 0.004
    - Hair: 0.324
    - Eyes: 0.364
  - By this criterion, Eyes is chosen as the best root available
  - The branch for Eyes: brown is already completely classified, and we need to work only on the sub-table for Eyes: blue

Machine Learning

- A data set for classification (Quinlan)

	----Attributes----		---Decision---
	<u>Height</u>	<u>Hair</u>	<u>Attractiveness</u>
1	short	blond	+
3	tall	red	+
4	short	dark	-
5	tall	dark	-
6	tall	blond	+

- H(decision) = H(Attractiveness)

$$= -\frac{3}{5}\log_2\left(\frac{3}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)=0.971$$

Machine Learning

- A data set for classification (Quinlan)
  - Height:
    - short: 1, 4 p(+|short) = 1/2 p(-|short) = 1/2
    - tall: 3, 5, 6 p(+|tall) = 2/3 p(-|tall) = 1/3
  - H(decision|attribute) = H(Attractiveness|Height) =

$$\frac{2}{5}\left[-\frac{1}{2}\log_2\left(\frac{1}{2}\right)-\frac{1}{2}\log_2\left(\frac{1}{2}\right)\right]+\frac{3}{5}\left[-\frac{2}{3}\log_2\left(\frac{2}{3}\right)-\frac{1}{3}\log_2\left(\frac{1}{3}\right)\right]=0.951$$

  - H(Height) = $-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{3}{5}\log_2\left(\frac{3}{5}\right)=0.971$

Machine Learning

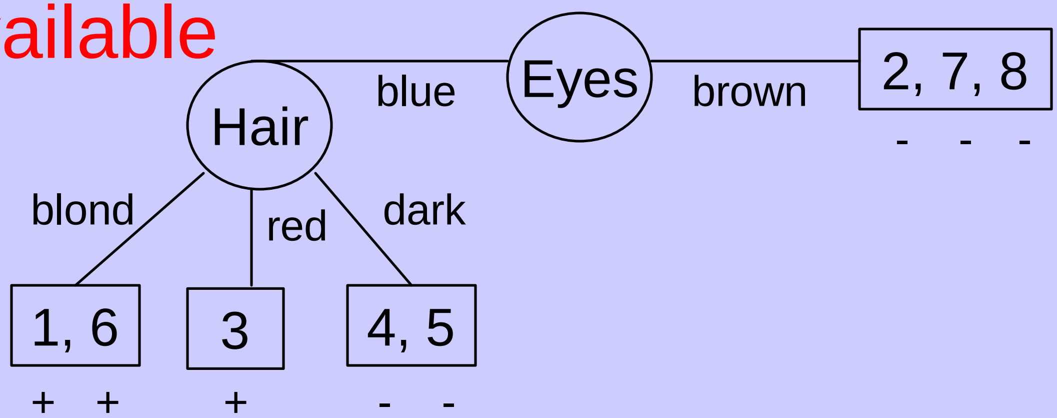
- A data set for classification (Quinlan)
  - Hair:
    - blond: 1, 6 — p(+|blond) = 2/2 — p(-|blond) = 0/2
    - red: 3 — p(+|red) = 1/1 — p(-|red) = 0/1
    - dark: 4,5 — p(+|dark) = 0/2 — p(-|dark) = 2/2
  - H(decision|attribute) = H(Attractiveness|Hair) =

$$\frac{2}{5}[0]+\frac{1}{5}[0]+\frac{2}{5}[0]=0$$

  - H(Hair) = $-\frac{2}{5}\log_2\left(\frac{2}{5}\right)-\frac{1}{5}\log_2\left(\frac{1}{5}\right)-\frac{2}{5}\log_2\left(\frac{2}{5}\right)=0.793$

Machine Learning

- A data set for classification (Quinlan)
  - Hence the Information gain ratios are
    - Height: 0.021
    - Hair: 1.224
  - By this criterion, Hair is chosen as the best root available



Machine Learning

- A data set for classification (Quinlan)
  - We may now build rules from this decision tree
    - R1: (Eyes, brown) --> (Attractiveness, -)
    - R2: (Eyes, blue) & (Hair, blond) --> (Attractiveness, +)
    - R3: (Eyes, blue) & (Hair, red) --> (Attractiveness, +)
    - R4: (Eyes, blue) & (Hair, dark) --> (Attractiveness, -)
  - These are different rules
  - Note that after dropping conditions, however, they are the same - this is NOT generally true

Machine Learning

- ID3 & C4.5
  - What if too many cases?
    - Windowing
  - What if the data is incomplete?
  - What if the data is inconsistent?
  - What if the data is continuous?
    - Binarization
    - Discretization
  - Incremental algorithms?
  - Pruning?