MIT – AI course

(Machine Learning)

Lecture 005

Sandro Spina

Formal Languages and Automata

- What is a language ?
- A grammar generates a language
- Eg. S -> aSa | b
- But there are more complex languages
- There’s actually a hierarchy (Chomsky)

Chomsky's Language Hierarchy (in brief)

- *A Grammar is a quad-tuple <N, Sigma, S, P >*
  - N is a set of non-terminal symbols
  - Sigma is a set of terminal symbols
  - S is the set of initial non-terminals
  - P are the production rules

- Chomsky's hierarchy is defines in terms of restrictions on the production rules S

- In general we have (Si U N)* N (Si U N)* -> (Si U N)*

A counting Grammar

- *S -> 1S1 | +A*
- *A -> 1A1 | =*

- *Check out grammar which generates the language with strings have equal number of a’s and b’s*

- *Try it out !!!*

- *These are context-free grammar since the production rules are using the form N -> (Si U N)**

There are four classes

1. Regular Grammars
2. Context-free Grammars
3. Context-Sensitive Grammars
4. Unrestricted (any language accepted by a Turing machine)

- □ Each class can express (generate) fewer formal languages.
- □ 2,3 are used in parsers and compilers

Regular Grammars

- Simplest form of grammars …
- Production rules have the form
  - N -> Si N | Si
- Eg. S -> S10 | 0
- Equivalent to Finite State Automata

So we now know what a grammar is formally

- Regular Grammar
- Context-Free Grammar
- Stochastic Grammar
- Used as Transducer
- DFA / NDFA
- Regular Expressions
- Finite Representation !!!!!

What could a grammar represent ?

- Generates a language !!!

- But the language of what ???
  - Any sequence or syntactic structure
  - Examples:
    - Electrocardiogram
    - Chain-coded image contour
    - Banded chromosome grey-level sequence
    - Acoustic-phonetic sequence
    - Successive actions of an agent
    - Chess moves
    - Natural language sentence
    - Piece of music
    - DNA sequence …

- Syntactic structure is properly represented by Grammars and Automata.

What is learning ?

- Learning OF what ??!!??
- Learning FROM what ??!!??
- How do humans/animals learn ???
  - Learning by observing patterns
  - Learning by querying a teacher !! Sometimes
  - GI started when researchers started questioning how children were able to **learn** their natural language simply by being exposed to it.
- A few years of casual contact with the ambient language suffices fir the infant to master a very complex grammatical system. The child’s learning mechanism is apparently built to acquire any human language !!

How do we learn ?

- Generic Paradigm:
  - Learning can be conceived of as a *game* between Nature and a scientist.
  - First, a *class* of “possible worlds”, is specified in advance; the class is known to both players of the game.
  - Nature is conceived as choosing one member from the class, to be the actual world; her choice is initially unknown to the scientist.
  - Nature then provides a series of clues about the actual reality. These clues constitute the data upon which the scientist will base his *hypothesis*.
  - Each time nature provides a new clue, the scientist may produce a new hypothesis.
  - The scientist wins the game if his hypothesis ultimately becomes stable and accurate !!

- Different paradigms formalize this picture in different ways, resulting in different games.

A Simple Paradigm

- Call a set of +ve integers “describable” just in case it can be uniquely described using an English expression. Ex. {2,4,6,8, …} in one such set since it is uniquely described by the expression “all positive, even integers”.

- Let us now focus on a subset of these realities, namely the sub collection *C,* which contains all sets that consist of every +ve integer with a sole exception. For example the set {1,3,4,5,6,…} is uniquely described by “all positive integers except for 2.”

- In this paradigm I will play the role of nature and you the scientists !! I will select a member of the class *C,* and you must discover the set that we have in mind.

- Clues about the chosen set will be provided one element at a time. For example I might say 2,3,5,4,7,6,9,8,… Each time a number is presented one by one you may announce a conjecture about the set chosen from *C* at the beginning of the game. You win if you make only a finite number of conjectures, and the last one is correct.

A Simple Paradigm (Guessing rules)

- Guessing Rule: Suppose that S is the set of numbers that have been presented so far. Let m be the least +ve integer that is not a member of S. (S is finite, so such a number certainly exists !!). Emit the conjecture “all positive integers except for m”
  - FACT: No matter which set was chosen from C at the start of the game, and no matter what list was made from that set, consistent application of this guessing rule is a winning strategy; you will win in all cases.
- BUT what happens if we modify the game by adding the set of all positive integers to the initial collection *C*. The guessing rule is no longer guaranteed to succeed at the game. You would change your conjecture infinitely often, and hence never produce a last, accurate conjecture.

Identification

- Identification of functions …
- Identification of languages …

GI – A Formal Definition

- GI addresses the following problem
  - Given a finite set of strings that belong to some unknown formal language *L*, and possibly a finite set of strings that do not belong to *L*, we require a learning algorithm that infers *L*.
- The problem can be applied to any **language class.** We will focus our attention of the class of regular languages.

How possible is that ?

- No superfinite class of languages (that class which contains at least one infinite language ex.regular languages) can be identified in the limit from positive examples only. Gold *BAD NEWS*

- Any enumerable class of recursive language (context-sensitive and below) can be identified in the limit if the learner is presented with both positive and negative examples. Gold *GOOD NEWS but task is proved to be NP-HARD*

Theoretical Implications

- More formally, automata learning can be expressed as the following decision problem:
  - *Given an integer n and two disjoint sets of strings I+ and I- over a finite alphabet Σ, does there exist a DFA consistent with I+ and I- that has a number of states less than or equal to n?*
  - Gold shows that this problem is NP-Complete, so the best we can hope for is to output a DFA consistent with the training set which has the least number of states m, which still is greater than n.

Inference as Search

- We want to *search* (in an infinite search space of DFAs) for the target grammar.
- Grammatical Inference is all about searching for the most appropriate hypothesis !!
- Of course we can never be 100% sure whether our search has been carried out correctly or not !!

Search through what ?

- We search through a *lattice (a directed acyclic graph)* of DFAs.
- We keep on moving along our search path until we determine what our hypothesis will be.
- We'll later see how this set can actually become a partially order set using a partial order operator on it.

How do we search ?

- We search by looking at this partially order set (poset) and move from one DFA to the next.
- As we move along this transition relation we will get grammars which are more general !!
- A grammar A is more general than a grammar B when the language generated by B is a subset of the language generated by A.

How do we train ?

- Given an initial training set we construct an automata which is the most specific automata with respect to the training set.
- Then, we generalize this very specific automata by traversing the poset until we cannot create automata which are more generic and are still consistent with the training set.

Passive Learning …

- Why passive ?? Because there is no interaction between the learning process and the external world.

- The only information that is available to the learner is a **finite** set of strings from which to extract (infer) the target grammar

… And Active Learning

- Why active ?? Because there is interaction between the learning process and the external world.

- An oracle or teacher is used to correct the hypothesis of the learner. Different types of queries are possible. Moreover oracles may produce counterexamples on the learners hypothesis in order to aid the learner in correcting the hypothesis.

Applications of GI

- Bioinformatics
- Machine Translation
- Pattern Recognition

EDSM Algorithm for GI

- Training Set
- APTA
- Transition Trees
- Merging of States
- Compatibility of States
- From APTA -> hypothesis

- Learning = going from something specific (APTA) to something generic (hypothesis)