Connectionist and Statistical Language Processing

Lecture 1: Introduction

Matthew W Crocker

Computerlinguistik
Universität des Saarlandes

Goals of the Course

- How can natural languages be learned?
  - How do *people* learn language: *cognitive* computational models
  - How can *machines* learn language: *useful* computational models
- Part I: Connectionist models of human language learning
  - Emphasis of cognitively plausible model of human development
  - Text:
    - McLeod et al. (1998). *Introduction of connectionist modelling of cognitive processes.* UK:OUP.
  - Software: `tlearn` neural network simulator (**Win**/Mac/Linux)
- Part II: Machine learning of natural language
  - Various machine learning methods, evaluation, comparison
  - Texts:
    - Mitchell (1997). *Machine Learning.* New York: McGraw-Hill.
    - Witten & Frank (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.* San Diego, CA: Morgan-Kaufmann.
  - Software: `Weka` machine learning environment (**Java**)

Cognitive models of language learning

- Goal: account for human language learning behaviour
  - Performance: people successfully learn language
  - Development: how does learning take place over time
  - Pathologies: model “errors” or “weaknesses” of human performance
- What must people learn?
  - What is innate (*nature*)?
  - What it learned from experience (*nurture*)?
- Methods: Cognitive Plausibility
  - Psychologically (neurologically) plausible learning *mechanisms*
    - Connectionist? Symbolic?
  - Plausible learning *environment* (what are children exposed to)

Connectionist language learning: contents

- Connectionist Information Processing
  - Simple connectionist models and their properties: The perceptron
  - Multi-layer perceptrons: feed-forward networks and internal representations
  - The encoding problem: Localist and distributed representations
  - Generalisation, association, and translational invariance
- Connectionist Models of Language
  - Modelling acquisition of the English Past-Tense and reading aloud
  - Processing sequences: Simple recurrent networks
  - Modelling acquisition of hierarchical syntactic knowledge

Machine learning of natural language

- Goal: learn to automatically process language from experience
  - Performance: process language as accurately as possible
- What is learned:
  - Learn some output representation from some input
- Methods:
  - Statistical, symbolic, connectionist
  - Determined by performance and efficiency
  - Evaluation: performance on new (unseen data)

- From Witten and Frank (2000:6):

  "Things learn when they change their behavior in a way that makes them perform better in future."
- From Mitchell (1997:2):

  "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience P."

Machine learning techniques: contents

- Evaluation
  - Held-out estimation, cross validation, precision & recall, task-based
- Decision trees
  - Classification problems, attribute selection, inductive bias,overfitting
- Linear regression
  - Least squares estimation, predicted values
- Bayesian learning
  - Bayes theorem, maximum likelihood, naive Bayes classifier
- Clustering
  - Hierarchical vs. non-hierarchical clustering, k-means clustering

Recommended background

- Students should have completed the Vordiplom:
  - Ideally: <u>Mathematische Grundlagen III: Statistische Methoden</u>
- Mathematical background:
  - General facility for math and statistics is helpful, but specifics will be covered in the lectures:
    - Basic statistics
    - Linear algerbra (vectors and matrixes)
    - Differentiation (Analysis)
- No programming required:
  - Connectionist simulations will be done using the `tlearn` simulator
  - Machine learning techniques will be investigates using `Weka`

Structure of the Course

- 2 meetings each week:
  - Vorlesung (Di 14-16, room 2.11):
  - Übung (Do 14:30-16, CIP Raum):
    - Lab session: to work through simulation exercises (CIP Raum)
- Course review: 13 February
- <u>Course mark will be based entirely on the final Klausur</u>
  - Klausur: 18 Feb. 2002 @ 14:00 - 16:00
  - In order to take the Klausur, students must successfully complete the tutorial exercises (e.g. simulations)
    - These can mostly be completed during the lab/tutorial sessions
    - **No more than 1 tutorial session/exercise can be missed**
- Responsible for all material covered in lectures, exercises, assigned readings

Overview

- The problem of language acquisition and language processing
  - What constitutes a theory of language?
  - How does acquisition take place?
- Two (and a half) solutions:
  - Rationalist: symbolic/rule-based, theory-driven
  - Empiricist: *tabula rasa*, data-driven
  - Connectionist: Distributed information processing and learning mechanism
- The connectionist perspective
  - Connectionism and the brain
  - The appeal of connectionist modelling
- Basics of connectionist information processing

- Reference for lecture:
  - McLeod et al. (1998). Chapter 1: Introduction of connectionist modelling of cognitive processes. OUP, UK.

The Problem: Natural Language is ...

- Complex
  - Consider the formal systems of rules and representations required:
    - Phonology, Morphology,Syntax, Semantics, Discourse and Pragmatics
  - Interaction with reasoning, perception, context, task, ...
- Ambiguous
  - Speech:
    - "Recognise speech" *versus* "wreck a nice beach"
  - Language:
    - Lexical: "Time flies like an arrow"
    - Structural: "Sabine broke the window with her brother"
    - Semantic: "Every woman loves a man"
- Noisy
  - Speech: signal never the same, difference speakers, overlap, slips, restarts
  - Language: ungrammaticalities, unknown words

Theories of Language

- Theories of the human language faculty:
  - Knowledge: What is the nature of our knowledge of language?
    - Rules and Representations
    - Symbolic versus Distributed
    - Explicit versus Implicit
  - Acquisition: Where does knowledge come from?
    - Is some knowledge innate?
    - What linguistic knowledge learned, and how?
  - Use: How do people use knowledge to process new input?
    - What mechanisms do people use in applying existing linguistic knowledge to the interpretation of novel input?
- Connectionist models of human language:
  - Addresses these issues simultaneously

The "Traditional" Perspective

- Modern cognitive modelling has been heavily influenced by available theories of computation.
  - This has put an emphasis on digital, symbolic, logic-based accounts.
- Symbolic processing:
  - Emphasis on explicit rules and representations
  - Computation: digital computers, logic based
  - Language: Chomskian theory
    - Criticism of statistical approaches
  - General problem solving: Newell and Simon

Connectionist Information Processing

- Connectionist models of information processing can become complex, but the idea is based on simple neuronal processing in the brain:
  - Basic computational operation involves one neuron passing information related to the sum of the signals reaching it to other neurons
  - Learning involves changing the strength of the connections between neurons, and thus the influence they have upon each other
  - Cognitive processes involve the use large numbers of neurons to perform many of these basic computations in parallel
  - Information about an input signal or memory of past events, is distributed across many neurons and connections
- Terms: connectionism, parallel distributed processing, neural networks, neurocomputing

The "Connectionist" Perspective

- Rumelhart and McClelland (1987, p. 196):
  - *"... implicit knowledge of language may be stored among simple processing units organized into networks. While the behaviour of such networks may be describable (at least approximately) as conforming to some system of rules, we suggest that an account of the fine structure of the phenomena of language and language acquisition can be best formulated in models that make reference to the characteristics of the underlying networks"*
- Key ideas:
  - Neurologically based (but not true models of the brain)
  - Distributed, implicit representations
  - Dense connectivity
  - Communication of "real values" not "symbols"
  - Representations and processing are the same
  - Learning: supervised and unsupervised

Central issues

- All accept the fact that human language, like other cognitive and perceptual faculties, are realised (or implemented) in the neural-tissues of the brain.
- Big Questions:
  - What is the right computational level at which to develop our theories?
  - Can connectionism fully replace symbolic accounts or should it be viewed as complementary?
  - Is there a clear boundary between connectionist and symbolic computation in the brain, or does symbol/rule-like behaviour emerge gradually?
  - What kinds of cognitive function require connectionist explanation, and what are best suited to symbolic accounts?
  - Can statistical models be used to augment symbolic approaches with some of the relevant characteristics of the distributed/neural architecture?

Theory development

- **Neuroscience:** developing more faithful and accurate models of information processing in the brain.
- **Cognitive science:** modelling human behaviour as observed in
  - Linguistics: explaining linguistic knowledge
  - Psycholinguistics: modelling how knowledge is used
  - Cognitive psychology: psychological studies of human behaviour
- **Symbolic perspective:** develop a symbolic account, and then provide a connectionist implementation
- **Intermediate:** the possibility of connectionist implementation, and (relevant) properties of connectionist networks should inform theory development
- **Radical connectionist:** connectionism is so fundamentally different, it challenges the very foundations of symbolic theories.

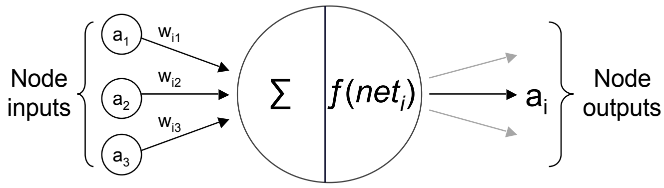
Properties of Connectionist Networks

- Learning
  - There is usually no predetermined (innate) knowledge of language, but ...
    - Input/output representation are often specified
    - The architecture of the network may be “suited” to a particular task
    - The learning mechanism and parameters provide degrees of freedom
  - Learning is takes place in direct response to experience
- Generalisation
  - Networks are able to learn generalisations not just by rote
  - More efficient representation of information
  - Novel inputs can be processed
- Representation
  - Learned automatically, and typically distributed

Properties continued

- Rules versus exceptions
  - Single mechanism to explain both general rules and also exceptions
- Graded:
  - Can often give a useful output to new, partial, noisy input
  - Damage is distributed, and some performance is still possible:
    - Modelling of brain damage and neurological disorders in possible
- Frequency effects
  - Model response time behaviours where high frequency inputs are recognised faster than low frequency ones

Basic Structure of Nodes



- A node can be characterised as follows:
  - Input connections representing the flow of activation from other nodes or some external source
  - Each input connection has its own *weight*, which determines how much influence that input has on the node
  - A node *i* has an output activation $a_i = f(net_i)$ which is a function of the weighted sum of its input activations, *net*.

- The net input is determined as follows: $net_i = \sum_j w_{ij} a_j$

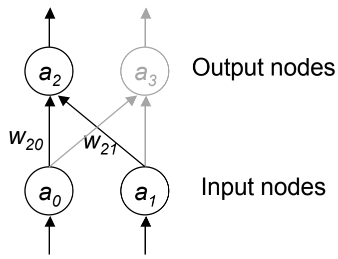
An example

- A two-layer feed-forward network:

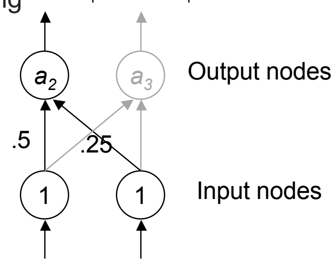
  $$net_i = \sum_j w_{ij} a_j$$

- So the net input for a_2 is:

  $$\text{net input } a_2 = w_{20} \cdot a_0 + w_{21} \cdot a_1$$



- Consider the network with the following inputs and weights:

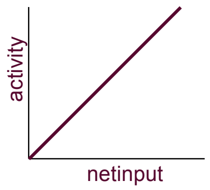


- The net input for node a_2 is:
  - 1 x .5 + 1 x .25 = 0.75

Activation functions

- The activation function determines the activation a_i for node *i* from the net input (net_i) to the node: $f(net_i)$
- Linear activation function
  - (McCulloch-Pitts neurode, perceptron)
  - Identity: the $a_i = net_i$

$$\boxed{\begin{array}{l} f(net_i) = net_i \\ f(0.75) = 0.75 \end{array}}$$



- Other activation functions:
  - Threshold
  - Binary threshold
  - Sigmoid function

About weights

- Node *j* influences a unit *i* by passing information about its activity level.
- The degree of influence it has is determined by the weight connecting node *j* to node *i*.
  - A smaller weights correspond to reduced influence of one particular node on another
- Weights can be either positive or negative
  - Positive weights contribute activation to the net input
  - Negative weights lead to a reduction of the net input activation
  - Brain: *excitatory* versus *inhibitory* connections

Übung

- Consider the following network:
  - Assume a linear activation function

- Determine the network outputs for the following inputs:
  - 0 0
  - 1 0
  - 0 1
  - 1 1