CMPT-882: Statistical Learning of Natural Language

Lecture #1

Anoop Sarkar
anoop_sarkar@sfu.ca
http://www.sfu.ca/~acs03

Plan for the course

- http://www.sfu.ca/~acs03/cmpt882_fall2002.html
  - Course description and outline of topics to be covered
  - Most papers can be downloaded from the web
  - Those that are pre-web will be handed out in class

Objectives of the course

- How can we learn to process natural language text?
- How much human supervision is needed for the learning process?
- Emphasis on basic algorithms that produce state of the art results on tasks involving natural language text

Objectives of the course

- Performance of each algorithm we study will be evaluated on human labelled data usually with comparisons between multiple learning algorithms

- We will cover corpus-based methods using both completely human labelled and partially labelled data

- Instead of particular tasks we will look at algorithms that can be applied to several tasks

Objectives of the course

- Compare knowledge-rich approaches which use a lot of human supervision to knowledge-poor techniques which try to bootstrap information

- The course will compare different properties of various machine learning algorithms

Objectives of the course

- For example, comparisons between:
  - Generative vs. Discriminative models
  - Probabilistic vs. Non-probabilistic corpus-based methods
  - Bootstrapping between single vs. multiple learners

NLP: what's it good for?

- Humans use language to communicate with each other
- Perhaps we can build programs that can be more useful to us by "listening" in . . .
- The AI challenge: language will have to play a large part in any mimic
- Lots of data available: newswire, the web

NLP: what's it good for?

- Many useful applications in text: from spam detection to information extraction from large collections
- Many useful applications in speech: transcription, speaker identification – this course: text only
- NLP provides a challenging testbed for ML algorithms (high dimensionality, complex classification, sparse data)
- Cognitive Science: sentence processing

NLP on text: some common applications

- Information Retrieval
- Named entity recognition – TE task, MUC eval
- Information Extraction – flipdog, TR task, MUC eval

NLP on text: some common applications

- Summarization
- Document Classification (spam detection, search engine IR, . . .)
  – TDT eval
- Machine Translation – bleu

NLP on text: some common applications

- Cross Language Information Retrieval
- Language Understanding: *what's the output?* – parseval, Communicator
- Language Generation: *what's the input?* – bleu

NLP on text: some common applications

- Question Answering
- Knowledge Acquisition (building a dictionary, thesaurus, . . . automatically)
- Improving Speech Recognition (better language models)

  $\arg\max_{w_i} Pr(w_i \mid w_0, \ldots, w_{i-1})$

NLP on text: some common applications

- Spelling correction, accent restoration, correcting speech recognized or OCR text
- Dialog Systems (call centres)
- Multi-modal dialog systems (AT&T, smartkom)

NLP on text: some uncommon applications

- Plagarism Detection
- Automatic evaluation of test essays – ETS
- Author identification

  Frederick Mosteller and David Wallace, *Inference and Disputed Authorship: The Federalist Papers*, Addison Wesley, 1964
- NLP for biological sequences (finding genes, predicting protein folding characteristics)

NLP on text: some common fundamental tasks

- Tokenization (not as easy as it might seem)
- Sequence Analysis:
  - Part of Speech tagging (is *can* a noun or a verb?)
  - Chunking
- Parsing and full utterance understanding (the Holy Grail)

NLP on text: some common fundamental tasks

- Word-sense disambiguation – senseval
- Co-reference resolution
- Discourse models

NLP on text: some common fundamental tasks

- Language modelling
- Parallel Corpus alignment
- Noisy channel models for MT

Machine Learning for NLP

- Zipf's Law: pervasive in language
- Capturing the tail through automatic acquisition of information
- Generalization captured through hand-written rules or via hand-selected models which learn generalizations
- Human supervision is better stored as labelled data rather than as rules embedded within a particular implementation
- Experts provide model selection criteria and relevant templates for feature extraction, learning algorithm does the rest

Machine Learning for NLP

- Supervision is expensive, labelled data makes it reusable
- Machine learning provides means of exploring lowering amounts of human supervision
- In this course, we will explore algorithms for learning from labelled data
- In addition, we will look at techniques that minimize amount of human supervision

Machine Learning for NLP

- Different types of machine learning:
  - Theoretical analysis of learning: recursion theory, PAC learning, VC dimension theory, large margin classifiers
  - Applied machine learning for particular problems of language data: sparseness, high dimensionality, branching processes (decision trees/lists, HMMs, PCFGs, . . .)
- In this course, we will concentrate on the latter but keep in mind the importance of theoretical analysis

Supervised Decision List Learning

- A class of applications for decision list learning
  - Accent/Capitalization restoration
    AIDS/aids Bush/bush
    French: cote $\rightarrow$ côte (coast) / côté (side)
    Spanish: sabana $\rightarrow$ sabana (grassland) / sábana (bedsheet)
  - Word sense disambiguation
    plant $\rightarrow$ living/factory
    plant growth vs. nuclear plant

Supervised Decision List Learning

- Input to the learning algorithm:
  1. n labelled training examples of the form $(\mathbf{x}_i, y_i)$
  2. $\mathbf{x}_i$ is the input without the correct label
  3. y_i is the label for that example i
  4. y_i is a member of $\mathcal{Y} = \{1, \ldots, k\}$
  5. $\mathbf{x}_i$ is a list of m_i features: $\{x_{i_1}, x_{i_2}, \ldots, x_{i_{m_i}}\}$
  6. each x_{i_j} is a member of $\mathcal{X}$, the set of possible features

Supervised Decision List Learning

- Output of the learning algorithm:

  1. function $h : \mathcal{X} \times \mathcal{Y} \to [0, 1]$
  2. $h(x, y)$ is an estimate of the conditional probability $p(y \mid x)$
  3. remember: $\widehat{p}(y \mid x) = \frac{\widehat{p}(x,y)}{\widehat{p}(x)}$ and $\widehat{p}(x) = \sum_{y \in \mathcal{Y}} \widehat{p}(x, y)$
  4. h defines a decision list of rules $x \to y$ sorted by weight $h(x, y)$
  5. $h(x, y) = \frac{c(x,y)+\alpha}{c(x)+k\alpha}$
  6. α avoids zero counts, k is the number of labels

Supervised Decision List Learning

- Features are extracted using the following templates (one sense per collocation)

  1. Word immediately to the right ($+1$ W)
  2. Word immediately to the left (-1 W)
  3. Word found in $+-k \ldots j$ word window, e.g. ($+-2 \ldots 10$)
  4. Pair of words at offsets -2 and -1
  5. Pair of words at offsets -1 and $+1$
  6. Pair of words at offsets $+1$ and $+2$

Supervised Decision List Learning: living or factory

```
 training_VBG new_JJ Ukrainian_JJ     plant    _NN operators_NNS to_TO re
who_WP are_VBP leaving_VBG the_DT     plant    s_NNS in_IN Ukraine_NNP ar
CC safety_NN procedures_NNS at_IN     plant    s_NNS in_IN both_DT countr
t_IN the_DT Orange_NNP County_NNP     plant    _NN ._.
Z closing_VBG three_CD missile_NN     plant    s_NNS in_IN southern_JJ Ca
_IN the_DT whole_JJ Chernobyl_NNP     plant    _NN in_IN 1991_CD ,_, five
IN a_DT hill_NN ,_, gardeners_NNS     plant    _NN begonias_NNS ,_, makir
$_$ 200_CD million_CD printing_NN     plant    _NN in_IN Brooklyn_NNP ,_,
of_IN incompletely_JJ oxidated_JJ     plant    _NN and_CC animal_NN sedim
whenever_WRB you_PRP eat_VBP a_DT     plant    _NN ._. ''_''
n_IN return_NN for_IN a_DT new_JJ     plant    _NN near_IN Tuscaloosa_NNP
T carmaker_NN could_MD finance_VB     plant    _NN construction_NN with_I
n_IN return_NN for_IN a_DT new_JJ     plant    _NN near_IN Tuscaloosa_NNP
```

Supervised Decision List Learning

- For this example, $y = \{\texttt{living}, \texttt{factory}\}$
- $h(\texttt{e}nv_i, y) = \frac{c(env_i,y)+\alpha}{c(y)+2\alpha}$
- Also $p(y \mid x)$ can be estimated using the log-likelihood ratio
- $h(\texttt{e}nv_i, y) = \texttt{abs}(\texttt{log}(\frac{p(\texttt{living}|env_i)}{p(\texttt{factory}|env_i)}))$

Supervised Decision List Learning

- Test example with features $\mathbf{x}$ is then presented to the decision list
- Output is $f(\mathbf{x}) = \underset{x \in \mathcal{X}, y \in \mathcal{Y}}{\text{arg max}} \; h(x, y)$
- Implicit search for single best feature x in the list $\mathbf{x}$
- Decision lists provide an `if-then-else` statement

For Wednesday

- Read Unsupervised Word Sense Disambiguation Rivaling Supervised Methods David Yarowsky (1995).

- http://www.cis.upenn.edu/~anoop/courses/cmpt882_fall2002.html

- http://www.sfu.ca/~anoop/cmpt882_fall2002.html