GE77
COMPUTATIONAL LINGUISTICS

Athanasios N. Karasimos
akarasimos@gmail.com

BA in Linguistics | National and Kapodistrian University of Athens
Lecture 11 | Wed 30 May 2018
MACHINE LEARNING

The battle between Unsupervised and Supervised Techniques
**Machine learning** is a field of computer science that gives computers the ability to learn without being explicitly programmed.

Machine learning is closely related to computational statistics, which also focuses on prediction-making through the use of computers (relation to mathematical optimization).

Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis (*unsupervised learning*).

Machine learning can be *unsupervised* and *supervised*. 
MACHINE LEARNING: DEFINITION

01
Mitchell (1997): "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 E."

02
This follows Turing’s question "Can machines think?", which is replaced with the question "Can machines do what we (as thinking entities) can do?".

03
In Turing's proposal the various characteristics that could be possessed by a thinking machine and the various implications in constructing one are exposed.
• Machine learning tasks are typically classified into two broad categories, depending on whether there is a learning "feedback" available to a learning system:

• Supervised learning: The computer is presented with example inputs and their desired outputs, given by a «supervisor», and the goal is to learn a general rule that maps inputs to outputs.

• Semi-supervised learning: the computer is given only an incomplete training signal: a training set with some (often many) of the target outputs missing.

• Active learning: the computer can only obtain training labels for a limited set of instances, and also has to optimize its choice of objects to acquire labels for.

• Reinforcement learning: training data is given only as feedback to the programs actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

• Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).
• Another categorization of machine learning tasks arises when one considers the desired output of a machine-learned system:
  
  • In **classification**, inputs are divided into two or more classes, and the learner must produce a model that assigns unseen inputs to one or more (multi-label classification) of these classes. This is typically tackled in a supervised way (i.e. spam vs. non-spam emails).
  
  • In **regression**, also a supervised problem, the outputs are continuous rather than discrete.
  
  • In **clustering**, a set of inputs is to be divided into groups. Unlike in classification, the groups are not known beforehand, making this typically an unsupervised task.
  
  • In **density estimation** it finds the distribution of inputs in some space.
  
  • In **dimensionality reduction** it simplifies inputs by mapping them into a lower-dimensional space (docs with similar tasks).
Supervised Language Learning
Against
Supervised Language Learning
MACHINE LEARNS SYNTAX & MORPHOLOGY

01 Comparing [Machine] Syntactic and Morphological Learning (Penn Treebank)

02 Unsupervised Syntax Learning

03 Unsupervised Morphology Learning

04 Supervised Morphology Learning

05 Lightly (Un)Supervised Morphology Learning
• Inspired by older linguistic branches (language acquisition and psycholinguistics)

• Independent Models of Natural Language Learning

• Precursors of UML
  • Pacak & Pratt (1976),
  • Rumelhart & McClelland (1986)
  • Koch, Küstner & Rüdiger (1989)
  • Wothke & Schmidt (1992)

• Goldsmith (2001): Gold-standard approach

UNSUPERVISED MORPHOLOGY LEARNING

- First approach: identifying the boundaries of the morphemes and categorizing the stems, suffixes and prefixes (Harris 1955, 1967 and Hafer & Weiss 1974).
- Third approach: exploiting the model of phonological relations between pairs of associated words, (Dzeroski & Erjavec 1997).
- Fourth Approach: Minimum-Length Description (Goldsmith 2001).
GOLDSMITH’S (2001) APPROACH

- Linguistica: implementation of this model
- Analysis in a huge corpus of unannotated corpora.
- The aim is word segmentations in a way that approaches the analysis of a real morphologist.
- Create signatures
  - a group of affixes (either prefixes or suffixes) associated with a given set of roots or themes.
  - NULL.ed.ing.s + jump, laugh, walk, move, prove
  - e.ed.ing, NULL.s, NULL.ing.s, NULL.er.est.ly
GOLDSMITH’S (2001) APPROACH

• Signature Architecture
  Stem1
  Stem2  Suffix1
  Stem3  Suffix2
  Stem4

• Problems:
  • Absence of formulas
  • Management of allomorphy
  • Apply phonological rules
GOLDSMITH’S (2001) APPROACH

\[
\sum_{w=t+f}[w](\log \text{prob}(\sigma(w)) + \log \text{prob}(t) + \log \text{prob}(f | \sigma(w)))
\]

\[
P(w = w_{1,i} + w_{i+1,l}) = \frac{1}{\sum_{j=1}^{l-1} H(w_{1,i}, w_{i+1,l})} e^{-H(w_{1,i}, w_{i+1,l})}
\]

\[
H(w_{1,i}, w_{i+1,l}) = -i \log(\text{freq}(\text{stem} = w_{1,i})) + (l - i) \log(\text{freq}(\text{suffix} = w_{i+1,l}))
\]

\[
\sum_{t \in T} (\log(26) * \text{length}(t) + \log \frac{[W]}{[t]})
\]

\[
\sum_{w \in W} \left[ \log \frac{[W]}{[\sigma(w)]} + \log \frac{[\sigma[w]]}{[\text{stem}(w)]} + \log \frac{[\sigma[w]]}{[\text{suffix}(w) \in \sigma(w)]]} \right]
\]
Creating and evaluating signatures

Create Candidate Signatures

Firstly, the system generates a few candidate signatures (joining elements) and then evaluate the candidates, so the system decides which are the real signatures.

This method begins to create a list of affixes (mainly suffixes), a reverse dictionary (by typing from the right of the words) and builds sets of possible suffixes of up-to-six characters in length (for example, -ivity > #cret#ivity, where # signs the boundaries). Then he uses an algorithm that weighs/calculates all possible suffixes to detect the actual suffixes and then groups them into signatures.
After the signatures generation, Goldsmith proposes an evaluation metric based on Rissanen’s (1989) Minimal Length Theory, where the best proposal for signatures is the most compact description of the specific language.

The normal signatures emerged from the final dismantling of the candidates generation together with the related stems >> the proposal of the UML model for the word segmentation of the language into subjects and suffixes;

However, it evaluates all the signature in order to keep the real ones, that analyzes correctly the words. Goldsmith (based on MLD) uses an evaluation metric to enable a more structured and condensed description of the morphology.
GOLDSMITH’S (2001) APPROACH

• Linguistica’s Results

<table>
<thead>
<tr>
<th>Categories</th>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>82.9%</td>
<td>83.3%</td>
</tr>
<tr>
<td>Wrong Analysis</td>
<td>5.2%</td>
<td>6.1%</td>
</tr>
<tr>
<td>Spurious Analysis</td>
<td>8.3%</td>
<td>6.4%</td>
</tr>
<tr>
<td>Failed to analyze</td>
<td>3.6%</td>
<td>4.2%</td>
</tr>
</tbody>
</table>
GOLDSMITH’S (2001) APPROACH

Criteria of the UML Models

A UMLM does not accept any morphological and phonological rule, does not include pre-created dictionary/vocabulary and obviously does not use any advantage from any (more specific morphological) theory or theoretical framework.

The complexity of the fusional morphological languages

The intense combination of productive affixes.

The presence and participation of allomorphy.
GOLDSMITH’S (2001) APPROACH

- Testing Greek Corpora
  - Corpus 1 (55,897 tokens) εφ. Μακεδονία
  - Corpus 2 (30,907 tokens) Targeted word list
  - Corpus 3 (281,821 tokens) Σκήπτρο του Φοίνικα (2 books)

<table>
<thead>
<tr>
<th>Analyses</th>
<th>Corpus 1 «Μακεδονία»</th>
<th>Corpus 2 «Στοχευμένο»</th>
<th>Corpus 3 «Σκήπτρο x2»</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>07,31%</td>
<td>23,41% (*27,72%)</td>
<td>11,49%</td>
</tr>
<tr>
<td>Wrong analyses</td>
<td>05,22%</td>
<td>42,30% (*40,01%)</td>
<td>49,62%</td>
</tr>
<tr>
<td>Failed analyses</td>
<td>86,38%</td>
<td>29,76% (*26,21%)</td>
<td>31,44%</td>
</tr>
<tr>
<td>Spurious Analyses</td>
<td>01,09%</td>
<td>04,53% (*06,06%)</td>
<td>07,45</td>
</tr>
</tbody>
</table>
GOLDSMITH’S (2001) APPROACH

**Results**

- **Correct Analyses:** nominal Inflectional Class without allomorph + Verbal Present

- **Wrong Analyses:**
  - Merging two affixes (αντικατα-, -τζηδες)
  - Stem as part of suffixes (αιμα-τα, παπα-δες, αγαπ-ησσα)
  - Failed similar stems (βηματ-α/ βηματ- ακι || αιμ-ατα/ αιμ-ατακι)
  - No-detection of linking elements (φωτο-βολος)
  - No allomorphy detection (παιδι~ παιδ!)
Approaches to Supervised Morphology Learning

- Rule-based models
- Stochastic models
- Connectic models

**Basic idea:** it is the extraction of some generalized standards/rules/behaviors from a training data set.

- The relationship between the input and output results presented in a set of examples >> therefore, the algorithm learned from the training data, to predict what will be requested from the new input.
More Specific Approaches to Supervised Morphological Learning

- Maximum Entropy
- Memory-based Learning
- Transformation-based Learning
Ratnaparkhi (1997): Maximum entropy theory is a clear way for researchers to combine data / findings for data modeling; at the same time, it points out that it is independent of computational analysis and can be applied seamlessly to other linguistic issues.

It represents accurately the behavior of a random processing, where such a model is the method of estimating the dependent probability that with contextual data X will give the extracted Y.
MAXIMUM ENTROPY

- A set of $X$ elements related to the past of events (i.e. preceding words, word tagging, morphological data)
- A set of data $Y$ relating to the future of the events (i.e. the word under consideration, the combination of characteristics, the relationship between morphological data)
- An indicative number of features describing the relationship between elements $X$ and $Y$.

$$p^* = \arg \max_{p \in P} H(p)$$
<table>
<thead>
<tr>
<th>ID</th>
<th>Form</th>
<th>Lemma</th>
<th>POS</th>
<th>MS FEATS</th>
<th>Head</th>
<th>Rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lo</td>
<td>Lo</td>
<td>RD</td>
<td>gen=M</td>
<td>num=S</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>scampato</td>
<td>scampato</td>
<td>A</td>
<td>gen=M</td>
<td>num=S</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>pericolo</td>
<td>pericolo</td>
<td>S</td>
<td>gen=M</td>
<td>num=S</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>scatena</td>
<td>scatenare</td>
<td>V</td>
<td>num=S</td>
<td>per=3</td>
<td>mod=I</td>
</tr>
<tr>
<td>5</td>
<td>la</td>
<td>la</td>
<td>RD</td>
<td>gen=F</td>
<td>num=S</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>squadra</td>
<td>squadra</td>
<td>S</td>
<td>gen=F</td>
<td>num=S</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>PU</td>
<td></td>
<td>6</td>
<td>punc</td>
</tr>
</tbody>
</table>
The Memory-Based Learning Theory >> decisions about new data are based on reuse of stored past experiences/data.

The prediction for the output is the result of some attributes of the input data made by identifying data in the memory, matching a model to these data in order to make predictions based on the model.
 MEMORY-BASED LEARNING

• A Memory-Based Learning model consists of four components:
  • a distance metric,
  • the number of nearby neighbours,
  • a weighting function and a
  • a model

Example
• KYMA ~ KYMATA
• BHMA ~ ???
• Keuleers & Daelemans (2007) aim to guess the ordinal order of each input as well as plural types of approximate models stored in the model memory.
Main Idea: to start the model with simple solutions to the problem, to implement some transformations constantly, so that they grow to the benefit of the system, chosen and applied to the problem.

The algorithm stops when the selected transformations do not further modify the data or there are no other transformations to select.
ALLOMANTIS’ EXPERIMENT

- AlloMantIS: An AMIS prediction algorithm analyzer
- Nominal test model (2755 derivatives)
- First attempt (86.49%), 2nd attempt (91.43%)
- Changing the syllabic number for improvement
- Training Corpus: Inflectional words, Test Corpus: derivational words
### ALLOMANTIS’ EXPERIMENT

<table>
<thead>
<tr>
<th>AC3</th>
<th>Positive affection weights</th>
<th>Negative affection weights</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Syllable2_τρης</td>
<td>5.41E+02</td>
</tr>
<tr>
<td></td>
<td>Syllable3_ρη</td>
<td>1.37E+02</td>
</tr>
<tr>
<td></td>
<td>Syllable1_μεγ</td>
<td>7.62E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable2_δης</td>
<td>5.31E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable3_πης</td>
<td>4.28E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable2_υτης</td>
<td>3.54E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable3_δης</td>
<td>2.87E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable3_φης</td>
<td>2.58E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable4_χη</td>
<td>2.19E+01</td>
</tr>
<tr>
<td></td>
<td>Syllable1_ζη</td>
<td>2.05E+01</td>
</tr>
</tbody>
</table>
• Results

![AlloMantIS Results](image-url)
ASSIGNMENT 2

• Paper Review
• 450 to 600 words
• 5 chosen topic from Machine Learning
• 1 topic from Unsupervised Learning
• 3 topic from specific Supervised Learning Model
• 1 general topic from Supervised Learning
• Due to: **Wednesday 20/6/2018**
READINGS


• **ΚΑΡΑΣΙΜΟΣ Αθανάσιος** 2011. Υπολογιστική Επεξεργασία της Αλλομορφίας στην Παραγωγή Λέξεων της Νέας Ελληνικής (κεφάλαιο 4ο). Διδακτορική Διατριβή. Πάτρα: Πανεπιστήμιο Πατρών.