ΓΕ77
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: DEFINITION

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: TYPES I

- 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).

MACHINE LEARNING: TYPES II

- 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).

MACHINE LANGUAGE LEARNING

Supervised Language Learning
Against
Supervised Language Learning

MACHINE LEARNS SYNTAX & MORPHOLOGY

01
Comparing [Machine] Syntactic and Morphological Learning(βλ.

02
Unsupervised Syntax Learning

03
Unsupervised Morphology Learning

04
Supervised Morphology Learning

05
Lightly (Un)Supervised Morphology Learning

UNSUPERVISED 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
- Yarowsky & Wicentowski (2001), Schone & Jurafsky (2001), Creutz & Lagus (2002) και Johnson & Martin (2003)

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)
- Second approach: bigrams and trigrams, which are part of the part of the morphemes (cf. Janssen 1992, Klenk 1992 and Flenner 1994, 1995).
- 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

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

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

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

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

$$\sum_{w \in W}[w][\ \log\frac{[W]}{[\sigma[\mathrm{w}]]} + \log\frac{[\sigma[\mathrm{w}]]}{[\mathrm{stem(w)}]} + \log\frac{[\sigma[\mathrm{w}]]}{[\mathrm{suffix(w)} \in \sigma(\mathrm{w})]}\ ]$$

GOLDSMITH'S (2001) APPROACH

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.

GOLDSMITH'S (2001) APPROACH

Evaluating 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

Categories	English	French
Good	82,9%	83,3%
Wrong Analysis	5,2%	6,1%
Spurious Analysis	8,3%	6,4%
Failed to analyze	3,6%	4,2%

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
  - Corpus1 (55.897 tokens) εφ. *Μακεδονία*
  - Corpus2 (30.907 tokens) *Targeted word list*
  - Corpus3 (281.821 tokens) *Σκήπτρο του Φοίνικα* (2 books)

Analyses	**Corpus 1 «Μακεδονία»**	**Corpus 2 «Στοχευμένο»**	**Corpus 3 «Σκήπτρο x2»**
Good	07,31%	23,41% (*27,72%)	11,49%
Wrong analyses	05,22%	42,30% (*40,01%)	49,62%
Failed analyses	86,38%	29,76% (*26,21%)	31,44%
Spurious Analyses	01,09%	04,53% (*06,06%)	07,45

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-ησα)
    - Failed similar stems (βηματ-α/ βηματ- ακι || αιμ-ατα/ αιμ-ατακι)
    - No-detection of linking elements (φωτο-βολος)
    - No allomorphy detection (παιδι~ παιδ!)

SUPERVISED MORPHOLOGY LEARNING

- Approaches to Supervised Morphology Learning
  - Rule-based models
  - Stochastic models
  - Connectic models
- <u>Basic idea</u>: 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.

SUPERVISED MORPHOLOGY LEARNING

- More Specific Approaches to Supervised Morphological Learning
  - Maximum Entropy
  - Memory-based Learning
  - Transformation-based Learning

MAXIMUM ENTROPY

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.

$$p^* = \arg\max_{p \in P} H(p)$$

MAXIMUM ENTROPY

ID	Form	Lemma	POS	MS FEATS	Head	Rel
1	Lo	Lo	RD	gen=M|num=S	3	det
2	scampato	scampato	A	gen=M|num=S	3	mod
3	pericolo	pericolo	S	gen=M|num=S	4	sogg
4	scatena	scatenare	V	num=S|per=3|mod=I|tmp=P	0	ROOT
5	la	la	RD	gen=F|num=S	6	det
6	squadra	squadra	S	gen=F|num=S	4	ogg_d
7	.	.	PU	_	6	punc

- Dell'Orletta et al. (2007)
  - Detecting Subjects and Objects in Italian and Czech

MEMORY-BASED LEARNING

01

The Memory-Based Learning Theory >> decisions about new data are based on reuse of stored past experiences/data.

02

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 & Daelmans (2007) aim to guess the ordinal order of each input as well as plural types of approximate models stored in the model memory.

TRANSFORMATION-BASED LEARNING

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

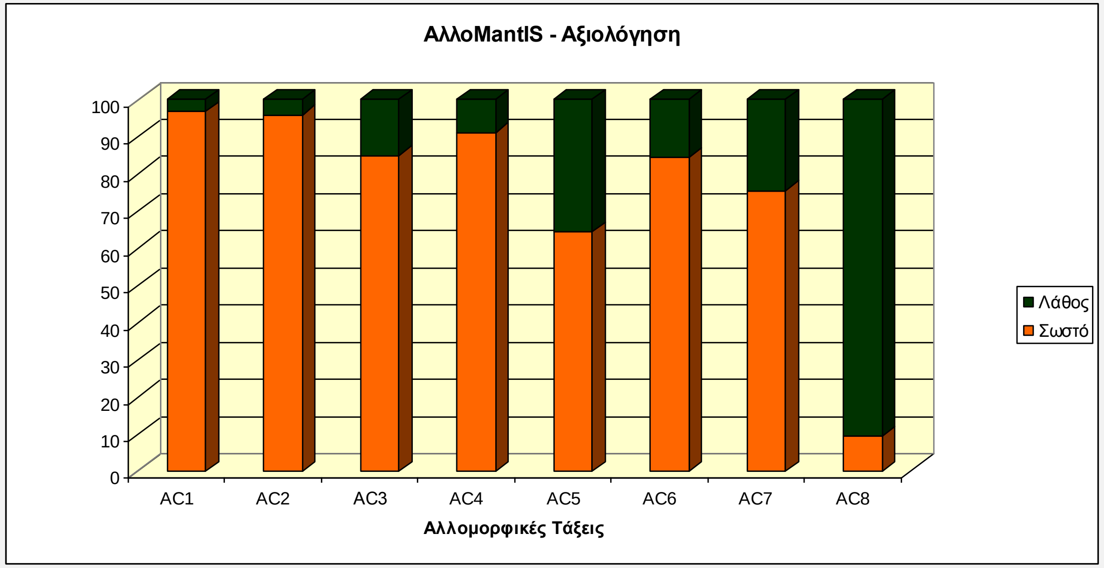
- AlloMantIS: An AMIS prediction algorithm analyzer
- Nominal test model (2755 derivatives)
- First attempt (86,49%), 2nd attempt (91,43%)
- Changing the syllabic number for improvement
- Training Corpus: Inflectional words, Test Corpus: derivational words

ALLOMANTIS' EXPERIMENT

AC3	Positive affection weights		Negative affection weights	
	Syllable2_τρης	5,41E+02	Character3_θ	1,96E-01
	Syllable3_ρη	1,37E+02	Syllable2_τα	1,78E-01
	Syllable1_μεγ	7,62E+01	Syllable2_να	1,64E-01
	Syllable2_δης	5,31E+01	Syllable3_δα	1,24E-01
	Syllable3_πης	4,28E+01	Character2_σ	1,09E-01
	Syllable2_ντης	3,54E+01	Stress_antipenultimate	9,18E-02
	Syllable3_δης	2,87E+01	Syllable2_μα	7,96E-02
	Syllable3_φης	2,58E+01	Syllable3_τα	7,05E-02
	Syllable4_χη	2,19E+01	Origin_italian	2,43E-02
	Syllable1_ζη	**2,05E+01**	**Origin_turkish**	**1,98E-02**

ALLOMANTIS' EXPERIMENT

- Results



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

- GOLDSMITH John 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* **27**(2), pp. 153-198.
- DELL' ORLETTA Felice, LENCI Alessandro, MONTEMAGNI Simonetta & PIRRELLI Vito 2007. Corpus-based modeling of grammar variation. In A. Sansò (ed.) *Language Resources and Linguistics Theory*, pp. 38-55.[Materiali Linguistici 59]. Milano: Franco Angeli.
- KEULEERS Emmanuel & DAELEMANS Walter 2007. Memory-based Learning of Inflectional Morphology: A methodological case study. *Lingue e Linguagio* **VI**.2, pp. 151-174. Bologna: Il Mulino.
- FLORIAN Radu & NGAI Grace 2001. Multidimensional transformational-based learning. In *Proceedings of the 5th Conference on Computational Natural Language Learning*, pp. 1-8.
- ΚΑΡΑΣΙΜΟΣ Αθανάσιος 2011. *Υπολογιστική Επεξεργασία της Αλλομορφίας στην Παραγωγή Λέξεων της Νέας Ελληνικής (κεφάλαιο 4ο)*. Διδακτορική διατριβή. Πάτρα: Πανεπιστήμιο Πατρών.