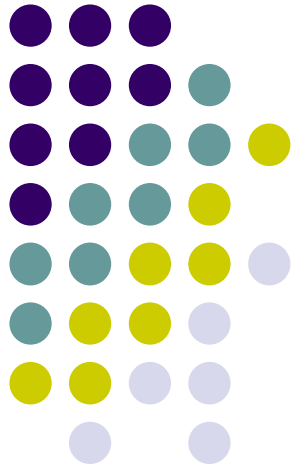
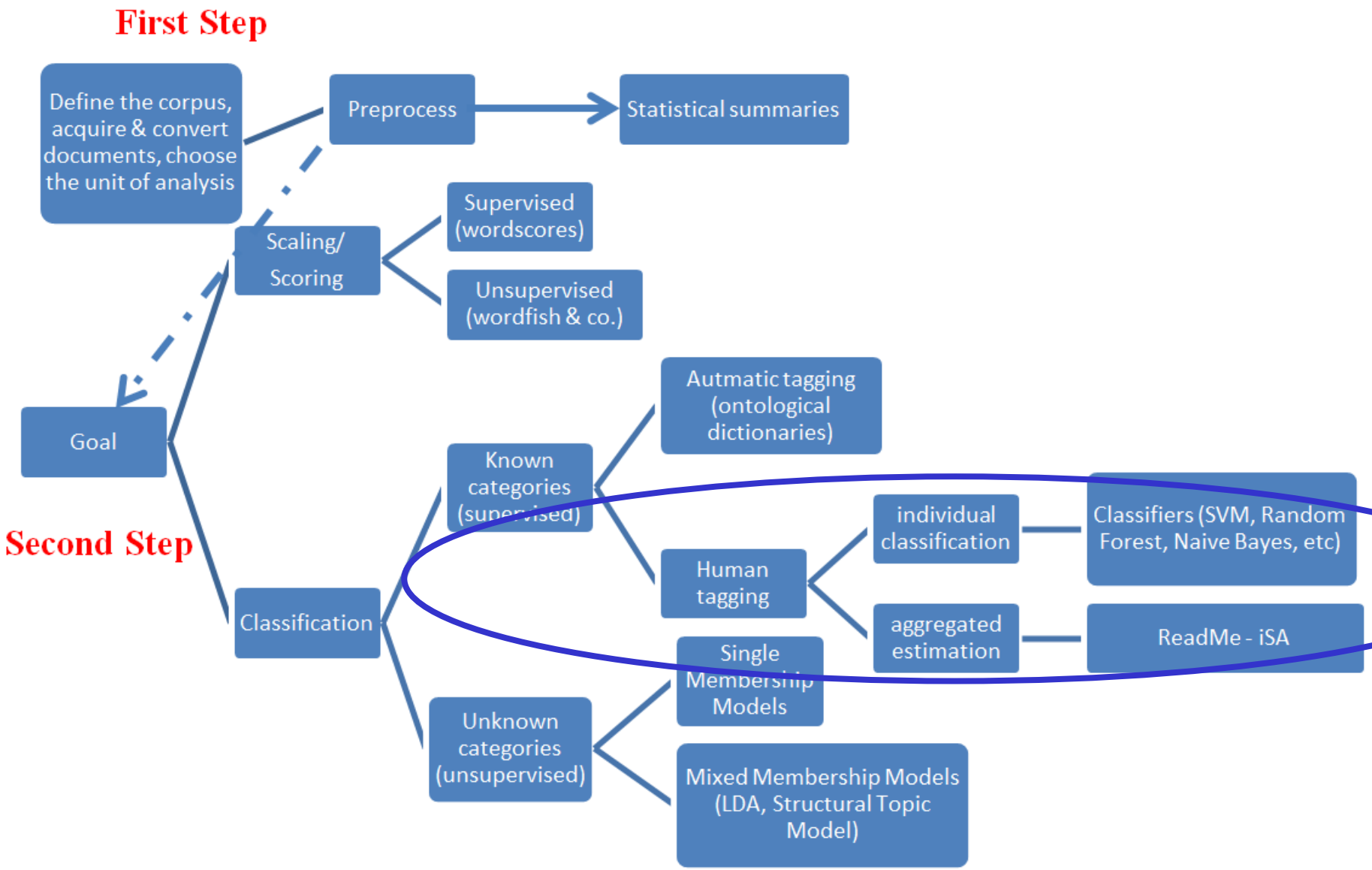
Big Data Analytics

Lecture 6 – part 2

Supervised classification methods with human tagging: an introduction



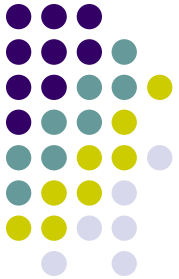
Our Course Map



References

- ✓ Grimmer, Justin, and Stewart, Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267-297
- ✓ Curini, Luigi, and Robert Fahey (2020). Sentiment Analysis and Social Media. In Luigi Curini and Robert Franzese (eds.), *SAGE Handbook of Research Methods is Political Science & International Relations*, London, Sage, chapter 29
- ✓ Barberá, Pablo et al. (2020) Automated Text Classification of News Articles: A Practical Guide, *Political Analysis*, DOI: 10.1017/pan.2020.8

Supervised Learning (classification) Methods



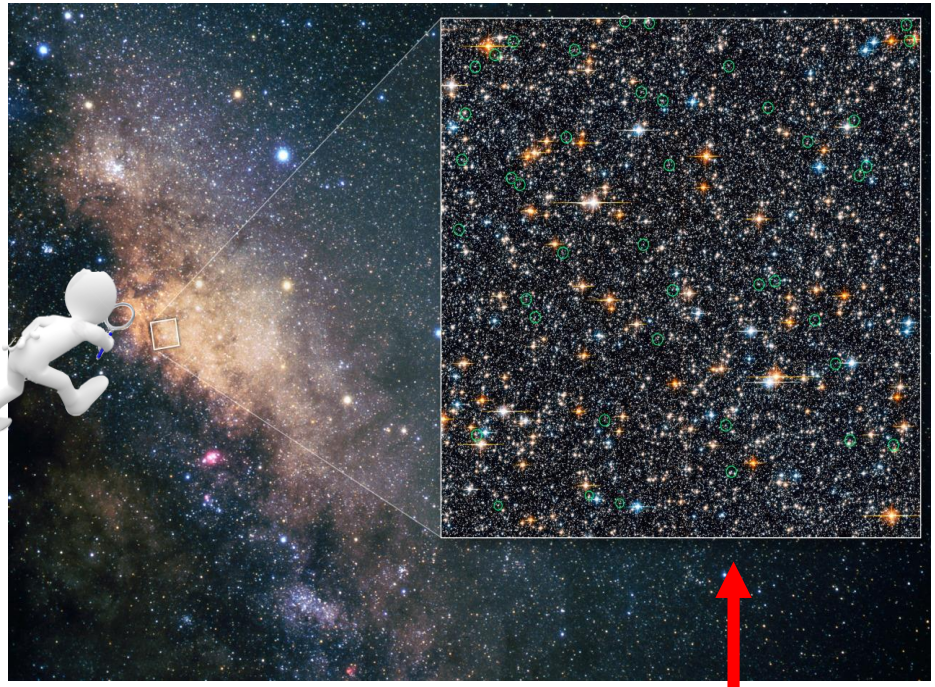
The idea of supervised learning is simple: human coders categorize a set of documents (the "**training-set**" or "**labelled-set**") by hand into a set of pre-defined categories (such as positive, negative, neutral for example)

The algorithm "learns" how to sort the documents into categories using the **training set and words**

Then, it classifies the remaining set of document not classified by hand (the "**test-set**" or "**unlabelled set**") using the characteristics (i.e., words) of the unread documents to place them into the categories

A four-step procedure

1. Data preparation: separating the training set from the test set in the corpus

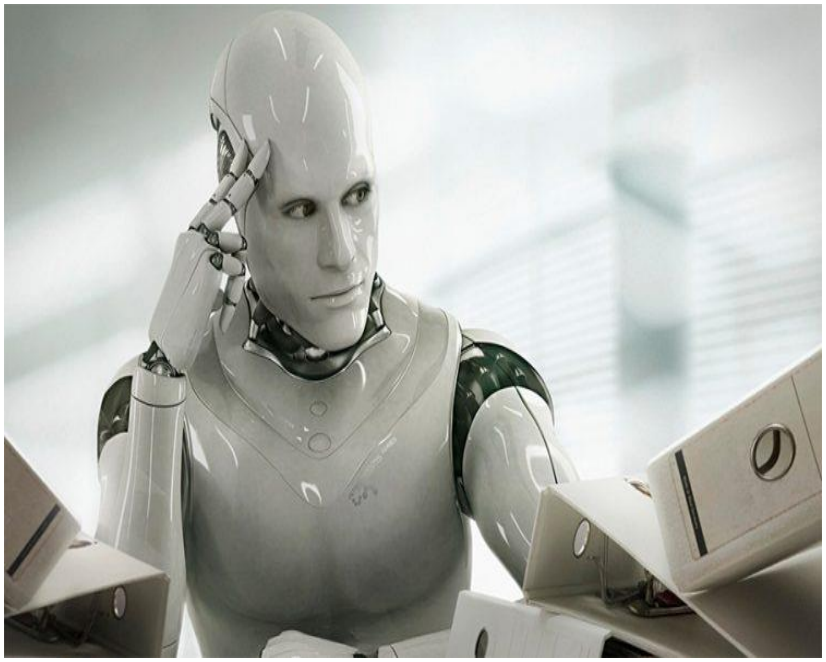


the training set

2. Human classification
of the training set on a base of a list of pre-defined categories

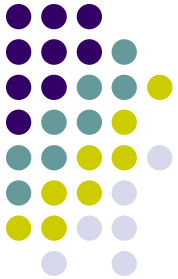
A four-steps procedure

3. Cogito ergo sum! The algorithm learns from the human classification done in the training set



4. Let's classify! The well-educated algorithm is now ready to classify all the texts in the test-set

Supervised Learning (classification) Methods



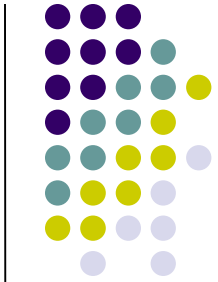
Despite the fact that the methods to do supervised classification are diverse, they share a common structure that usefully unifies the methods

Suppose there are N_{train} documents (i=1,…, N_{train}) in our training set and that we have pre-defined K categories (k=1,…,K) for our classification, such as positive, negative, neutral in the case of a sentiment analysis

Each document i's category is then represented by $Y_i \in (C_1, \ldots, C_K)$ and the entire training-set is represented as $\mathbf{Y}_{train}=(Y_1,\ldots,Y_{Ntrain})$

$\mathbf{W}_{train}$ is the term-document matrix for N_{train}

Supervised Learning (classification) Methods



Each supervised learning algorithm assumes that there is some (unobserved) function that describes the (true) relationship between the words and the labels in the training-set:

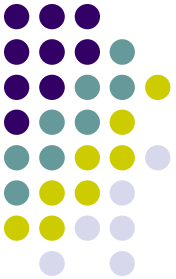
$$\boldsymbol{Y}_{train} = \mathrm{f}(\boldsymbol{W}_{train})$$

Each algorithm attempts to learn this relationship by estimating the "true" function *f* with $\hat{f}$ (the **classification function**)

$\hat{f}$ is then used to infer properties of the test set (the unlabeled set), $\widehat{\boldsymbol{Y}_{test}}$ using the test set's words $\boldsymbol{W}_{test}$:

$$\widehat{\boldsymbol{Y}_{test}} = \hat{f}(\boldsymbol{W}_{test})$$

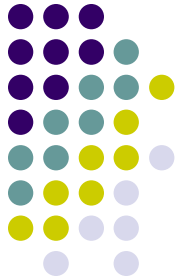
Supervised Learning (classification) Methods



Summing up…

Supervised learning models **share the same goal**: learn the potentially complicated relationships that relate (combinations of) features x to the outcome of interest y in general, using information available in the set of observations for which the pair (x; y) is fully observed (i.e., in the training-set)

Supervised Learning (classification) Methods

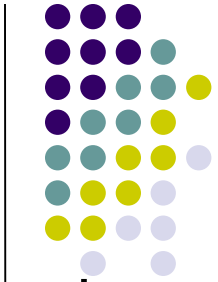


We have two broad classes of Supervised Learning Methods:

a) those who aim to classify the **individual documents** into categories

b) those who aim to measure the **proportion of documents** in each category

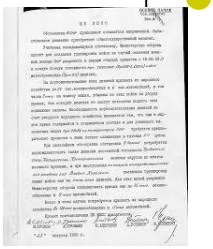
Applying a supervised learning model

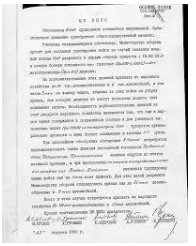


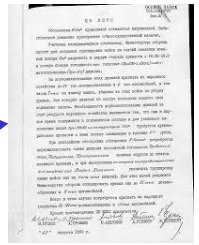
Individual approaches estimate the category of each text in a new corpus of texts (test-set) aiming at *minimizing the probability of error in individual class assignment*…

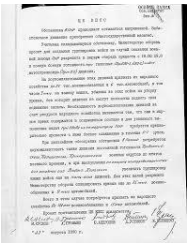
…while **aggregated approaches** estimate the class proportions into the new corpus of texts (test-set), rather than assembling the results of several individual classifications, aiming at *minimizing the error between the estimated proportions and true proportions*

Supervised learning: individual classification



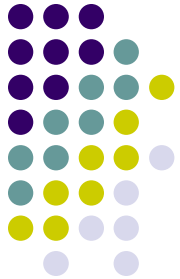






Supervised learning: proportional classification

Applying a supervised learning model



There is an important **theoretical (and statistical) difference** between individual and proportional classification:

- ✓ for some social science application, only the proportion of documents in a category is needed, not the categories of each individual document
- ✓ The opposite happens in some other application

Once again, there is no “a best method” out there. It depends on your research topic (remember the 4 rules!)

In our course, we will discuss mainly about individual classification approaches. Hopefully we will have time to discuss about proportional classification algorithm as well later on