Changelog: 2 Nov 2016, 7 Nov 2016

Lecture 1, Part II: What is machine learning?

The required reading for this part of the lecture is the handout and chapter 1 of the weka book. Optionally, you can also watch the introduction by Andrew Ng at Stanford University. The link is on the timetable. Andrew NG is a guru in ML. Andrew is very young and he is one of the founders of Coursera, an online platform for Massive Open Online Courses (MOOCs).

Outline: first definitions of the discipline, then the difference with statistics. We will continue by pointing out the difference between data and information. All in all, we will see what is meant by ML methods.

ML is a subfield of computer science and it is highly interdisciplinary. ML can be used in any possible field not only linguistics. Knowing ML is an asset for your life, even if you do not become a computational linguist.

Machine learning has been studied from a variety of perspectives, sometimes under different names. It is also not uncommon for a machine learning method or technique to have been studied under different names in the fields of artificial intelligence (AI) and statistics, for instance. Although machine learning techniques such as neural nets have been around since the 50's, the term "machine learning", as it is used today, originated within the AI community in the late 70's to designate a number of techniques designed to automate the process of knowledge acquisition.

Definitions of ML abound. In this slides you can read three definitions, Read: the first one seems to be very intuitive. Read: the second a bit more formal. Read: The third one quite mysterious. The first definition gives you the intuition of Ml: The core idea of ML is to create intelligent systems that "learn" from examples, learn from data, learn from experience". without being explicitly programmed. Sounds quite advanced, and quite revolutionary. In which way is it different from the traditional approach?

To solve a problem on a computer, we need an algorithm. An algorithm is a sequence of instructions that should be carried out to transform the input to output. For example, one can devise an algorithm for sorting. The input is a set of numbers and the output is their ordered list. For the same task, there may be various algorithms and we may be interested in finding the most efficient one, requiring the least number of instructions or memory or both.

For some tasks, however, we do not have an algorithm—for example, to tell spam emails from legitimate emails. We can easily compile thousands of example messages some of which we know to be spam and what we want is to "learn" what constitutes spam from them. We know what the input is: an email document that in the simplest case is a file of characters. We know what the output should be: yes/no label indicating whether the message is spam or not. But we do not know how to transform the input to the output. What is considered spam changes over time and from individual to individual.

The mission of ML (if you to use the word "mission") is to construct a good a useful approximation of a process. We may not be able to identify the process completely, but we believe we can construct a good and useful approximation. That approximation may not explain everything, but may still be able to account for some part of the data. We believe that though identifying the complete process may not be possible, we can still detect certain patterns or regularities.

There is no need to learn to sort numbers, we already have algorithms for that; but there are many applications for which we do not have an algorithm but do have example data. Other examples: predict customer behaviour (people do not go to supermarkets and buy things at random: when they buy beer, they tend to buy chips; those who buy ice cream in summer also buy glögg in winter, etc.). There are patterns in data, and ML is about using mathematical models to identify meaningful patterns relying on example data or past experience.

What we lack in knowledge, we make up for in data. In other words, we would like the computer (machine) to extract automatically the algorithm for this task. Such patterns may help us understand the process, or we can use those patterns to make predictions: Assuming that the future, at least the near future, will not be much different from the past when the sample data was collected, the future predictions can also be expected to be right.

We use ML to solve complex problems where much data is available: Learning to recognize spoken words (speech recognition), learning to identify linguistic forms and structures (tagging, parsing, named entity recognition, etc.). learning to drive autonomous vehicles (trains at the airport, e.g. at Stansted airport in the UK); learning to classify new astronomical structures (black holes), learning to play games (chess, backgammon, etc. ML can be applied to everything. It is not something specific to linguistics, but we will use ML MAINLY on linguistic data, because all data has its own specificity.

ML is multidisciplinary. On the screen you can see an excerpt from Tom Mitchell's book where he points out in which way other disciplines merge into ML. For instance, you see that Bayes' theorem is important within ML because it is the basis for calculating probabilities of hypotheses. Statistics is important within ML because it helps us characterize errors that occur when estimating the accuracy of a hypothesis on a limited sample of data. From Information theory, ML borrows the measure of entropy. Go back there, ie to the Scalable Learning platform, if you need to refresh your knowledge. AI has strongly influenced ML especially as for problem solving, ML uses prior knowledge in the form of training data to guide learning. Philosophy has also a bearing on ML: in all ML books and manual you will read about the Occam's razor. Occam was a monk who lived in the Middle Ages. Occam's razor is a principle attributed to the 14th century logician and Franciscan friar William of Ockham. So, Occam's razor is a principle from philosophy. It basically states that the simplest hypothesis is the best. In simple words this means, if you have 2 competing models, choose the simplest one. Well… in short ML is a kind of crucible, or melting pot, a container where many other sciences are merged together to solve complex problems.

Machine learning uses the theory of statistics in building mathematical models, because the core task is making inference from a sample. Historically, the two ML and Statistics have had rather different traditions. If forced to point to a single difference of emphasis, it might be that statistics has been more concerned with testing hypotheses, whereas machine learning has been more concerned with formulating the process of generalization as a search through possible hypotheses. But this is a crude oversimplification: statistics is far more than hypothesis testing, and many machine learning techniques do not involve any searching at all. In the past, very similar methods have developed in parallel in machine learning and statistics. One is decision tree induction. Four statisticians (Breiman et al. 1984) published a book on Classification and regression trees in the mid-1980s, and throughout the 1970s and early 1980s a prominent machine learning researcher, J. Ross Quinlan, was developing a system for inferring classification trees from examples. These two independent projects produced quite similar methods for generating trees from examples, and the researchers only became aware of one another's work much later. A second area in which similar methods have arisen involves the use of nearest-neighbor methods for classification. These are standard statistical techniques that have been extensively adapted by machine learning researchers, both to improve classification performance and to make the procedure more efficient computationally. We will examine both decision tree induction and nearest-neighbor methods later. But now the two perspectives have converged. The techniques we will examine in this book incorporate a great deal of statistical thinking. From the beginning, when constructing and refining the initial example set, standard statistical methods apply: visualization of data, selection of attributes, discarding outliers, and so on. Statistical tests are used to validate machine learning models and to evaluate machine learning algorithms. In this course statistics is highly involved.

Society produces huge amounts of data from many sources: business, science, medicine, economics, geography, environment, sports, … Potentially valuable resource. However, raw data is useless: need techniques to automatically extract information from it. We can make a theoretical distinction between data, ie recorded facts and information, ie patterns underlying the data.

Extracting information from data. This information can be implicit, previously unknown, and potentially useful. What is needed is programs that detect patterns and regularities in the data. Then we get patterns that can be Strong patterns ⇒ good predictions. Easy to say. But there often problems because most patterns are not interesting, patterns may be inexact, or data may be garbled or missing

There are many algorithms for acquiring structural descriptions from examples. Structural descriptions represent patterns explicitly. Structural descriptions can be rules they can be trees, or they can take other forms. The important thing to stress here is that structural descriptions can be used to predict outcome in new situation or they can be used to understand and explain how prediction is derived. These methods originate from artificial intelligence, statistics, information theory, etc.

Let's say that our problem is to give recommendations about contact lenses. We need an ML algorithm that automatically issues contact lenses recommendations. We can represent the contact lenses problem in the form of a dataset similar to what you can see

on the slides. In general instances in a dataset are characterized by the values of the features (called attributes), that measure different aspects of the instances. Instance means example. Each row in the dataset represent and example or instance. Each instance is characterized by 5 attributes, namely the age of the patient, the spectacle prescription, the third columns indicates whether the patient is astigmatic, the fourth relates to the rate of tear production, which is important to lubricate contact lenses, and the final column shows which kind of lenses to prescribe, either hard, soft or none.

The complete set of rules correctly classifies all the examples (all the instances) that are listed in the dataset are on the slides. These rules are complete and deterministic: they give a unique prescription for every example. Usually this is not the case. Sometimes there are situations in which no rule applies; other times more than one rule may apply, resulting in conflicting recommendations. Sometimes probabilities or weights may be associated with the rules themselves to indicate that some are more important, or more reliable than others. For this contact lenses problem, which has very narrow scope, this set of rules is comprehensive. This set of rules represents a traditional way to tackle the problem as it is in this point in time. If add more instances later, we do not know whether this set of rules will still be effective. We said at the beginning that The core idea of ML is to create intelligent systems that ”learn from examples, learn from data, learn from experience”. Therefore, one may wonder if there is a more concise set of rules that can describe the same problem and at the same time make predictions on future examples, that are not yet listed in the dataset just now. The answer is yes. We live in a dynamic world where data and information are continuously updated. So we can use ML models to make predictions and to guess prospect cases, and to describe and gain knowledge from data.

You can see a structural description for the contact lens data in the form of a decision tree on the slides. This representation is more concise and can be visualized more easily. The tree calls first for a test on tear production rate, and the first two branches correspond to the two possible outcomes. If tear production rate is reduced (the left branch), the outcome is none. If it is normal (the right branch), a second test is made, this time on astigmatism. Eventually, whatever the outcome of the tests, a leaf of the tree is reached that dictates the contact lens recommendation for that case. This is an example of a machine learning model (a MLalgorithm) (the Decision Tree method that we will study later) that learns from the data. It does not only account for the current data, but it learns from data and can sort out future examples that are not yet in the dataset.

To the same problem, to the same dataset we could apply another ML model to get a different description, interpretation and prediction of the same data. Maybe a better interpretation, maybe a worse interpretation? We will learn later how to evaluate di outcome, the performance of different ML methods applied to the same data.

Definition 2 revisited: ”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**.” Mitchell (1997: 2) Contact lenses recommendation problem: Task T: the kind of contact lenses to prescribe Performance measure P: correct diagnoses Training experience E: a database of correct diagnosis. So also definition 2 explains in a concise way what ML does. There also many other good

definitions. They might differ a lot, but they all stress the importance of data, or past experience in the learning process.

Definition 3: Machine learning is programming computers to optimize a performance criterion using example data or past experience. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience.

The model may be ***predictive*** to make predictions in the future, or ***descriptive*** to gain knowledge from data, or both." In the case of a decision tree, parameters are things such as the number of branches of the tree, how to select the top nodes, etc. In the case of linear models, parameters weights and bias and so on. We will learn these things later. At this stage, the important thing to memorize is that we apply mathematical/statistical models that represent data and learn from data in different ways.

ML and Data Mining: many different opinions on how to characterize these two discipline. Here we will say that Machine Learning focuses on designing algorithms that can learn from and make predictions on the data. We said that the design of ML algorithms is influenced by many other disciplines, like statistics, probability, and the like. Application of machine learning methods to large databases is called data mining. The analogy is that a large volume of earth and raw material is extracted from a mine, which when processed leads to a small amount of very precious material; similarly, in data mining, a large volume of data is processed to construct a simple model with valuable use for example, having high predictive accuracy. So basically, these two fields are very overlapping for our purpose, because we are going to use a datamining workbench, called Weka, that implements all the ML algorithms that we are going to learn in this course.

ML & LT: ML within LT is pervasive. Nowadays, I honestly do not know whether you can become a computational linguist without competence in ML. ML is basically used in all types of applications, and it usually have the best performance over other approaches.

--- the end----