Machine Learning,ML

Prof. Carl Gustaf Jansson
Prof. Henrik Boström
Prof. Fredrik Kilander
Department of Computer Science and Engineering
KTH Royal Institute of Technology, Sweden

Lecture 9
Scenarios for Concept Learning

So welcome to this fourth lecture of the second week of the course in machine learning so this week as you know we try to characterize learning problems, and on this lecture we are going to look at different scenarios for how learning can take place. So I will start with two basic distinctions, so the first very important distinction in machine learning is between supervised versus unsupervised Learning.
So in supervised learning all input data is pre-classified by the use of unique concept labels, so it may either be so that we only try to learn from data items, from one class but it could also multiple classes. And the goal of supervised learning is to based on this these pairs, the input data and the learn a concept definition which best approximates these relationships. So an optimal scenario will allow the algorithm to correctly determine a concept label for a new unseen data item that occurs. In contrast in unsupervised learning no data are classified, this means that the data set contains only whatever in an earlier lecture called Predictor Features or Predictor Variables and there are no constant labels. So unsupervised learning algorithms therefore have to it identify commonalities among all the data items and find structures in the data set, and then with that as a base, try to group the data items based on some kind of similarity measure. Unsupervised learning algorithms have to decide on optimal portfolio of concepts there could be many variants there that best match the data state at hand and then arrange groupings of subsets of the data set so that those groupings match the portfolio concepts. So these are the two main situations classically, a lot of work and I'm still so goes on in supervised learning where we really have a lot of data but the data are labeled and the only problem we have is to find reasonable abstractions that captures the character of the data. Of course when we're going to use machine learning in a wider range of settings and for example in self-driving cars we cannot rely on everything being classified all the time because we will have continuous data and so on. So unsupervised learning becomes much more important over time, one should realize of course that the unsupervised learning problem is a much more difficult problem than supervised one. Let's turn to the second distinction which is between what we call offline learning (Batch) learning and online or Incremental learning. So classically most of the work in machine learning has been done offline for with the use of static batches of data fully available at the start of the analysis but increasingly so, we are interested in looking at continuous flow of data items over time and being able to analyze those data when they come. And of course there are some middle

ground here the initials cases here are extremes, we could also consider to look at smaller batches of data that we still have a slower data but we partition that flow and handle each mini batch at a time. So there are all kinds of such very variants possible. The distinction being offline and online is relevant both for supervised and unsupervised learning so one can say this these two distinctions are orthogonal to each other. Yeah concerning offline learning we talked about the situation where the system is not operating in a real-time environment but handles pre-harvested the data in static and complete batch forms. Most traditional machine learning algorithm are well adapted to offline learning and the parallel access to the whole data sets give full flexibility in playing with the use of the data items in all kinds of variations to optimize the learning process. So in contrast Online learning is a learning scenario where data is processed in real time, in the full incremental fashion, input data incrementally, gradually and continuously used to inductively extend the existing model. Normally results of earlier learning phases are typically maintained and regarded are still being valid. Incremental algorithms are frequently applied to what we call data streams or big data and we see more and more of that examples are stock trend prediction, and user profiling of such data streams. Many traditional machine are learning albums inherently support incremental learning but may have to be adapted to facilitate this in practice. So to sum up the two distinctions traditionally, most machine learning was offline learning of supervised character while we're now moving into a time where we increasingly need techniques to handle the online unsupervised case. We are now going to go through a number of scenarios of learning of the supervised and unsupervised kind, of the online and on the offline kind. So we will start from the very most simple situation which we describe as learning a single concept offline from pre classified positive examples, and the picture you see is a way to depict this situation where you have objects of the same kind, all labelled in a consistent way and the tasks where this hand is to formulate a description of a concept that in an optimal way covers the data items considered. And there are no other elements around in this case. So now we turn to scenario two and actually scenario two is more or less scenario 1, the only thing we introduced there is something very important but also very problematic.

So we look at the first case but we acknowledge the possibility of the presence of noise. So noise is a fundamental underlying phenomenon that is present in all data sets, it's a distortion in the data, that is unwanted by the perceiver of the date. So noise is anything that is spurious and extraneous to the true data and typically the noise is due to a faulty capturing process. And as you as you see from the picture here this may result in that there are objects actually outside, objects that should be clearly inside the concept definition here are outside because they are noisy but there were also maybe erroneous data elements that are clearly within the frame but they are erroneous and shouldn't be there. So there are no of course any anything can happen with noise and any algorithm that we encounter need to defend itself towards noise that means they have to be always mechanisms to fight noise, there could never be any hundred percent guarantee but at least there must be defense mechanism, and as you should understand noise can occur in all scenarios to come in the same fashion that's well in this first very simple one. We now turn to scenario 3 and actually scenario 3 is related to the last scenario where we talked about noise. So here we will talk about Outlier. So an Outlier is a data item that is distant from other observations. And another may be due to natural but

extreme variation so it may be quite okay or it may indicate an experimental error or other noise. So of course outliers the handle outliers is a challenge and is of course crucial to distinguish between the measurement error cases and the cases where the population as a heavy tail skewed distribution. And as you understand when we have we are informed that we have the latter case the natural case, normally in the concept definitions becomes much more complex in order to see to that the definition is such that it also can covers these extreme outliers and the same comment goes here ask for a noise for any more complex scenario to be discussed when we can always talk about Outliers and way to handle them. So for scenario four we stick to the learning of single concept, yeah what we will discuss here is the role of Negative examples. So in the first scenario we only considered Positive examples but for many situations and many algorithms we may shield with faster convergence towards the appropriate concept definition if we feed the algorithms with a mix not only or positive but also negative examples. And of course negative examples can be made available in a variety of way, we may see later when we look at learning of multiple concepts that of course we will have examples of the different classes available in parallel classified so therefore we can use the examples of the other classes as negative examples for one of them. Yeah or it's also possible if you have some domain knowledge available that gives information of the character of examples we could also then artificially generate negative examples and feed that into the learning algorithm. So let's go on, let's talk about what we call your scenario five. So scenario file is actually an interesting variant of the use of negative examples which we refer to as Near Misses. So the question is when we have this situation where we have to do we want to use negative example to constrain the learning process to make it more efficient, oh it's not obvious what type of obvious we should use because there may be any alternatives and negative examples may be very far and differ considerably from the positive examples, which then doesn't help us much because there are too much flexibility still for the algorithm to determine the classification boundary in the light of negative examples that are very very far from that boundary. So the idea here is to use negative examples that differ from the learning concept in only a small number of significant ways. So as to picked it here in the small picture you see the negative examples kind of cling to the rim of the concepts and so these kinds of negative examples are very close we will refer to as Near Misses and as I said in the earlier slide for negative examples there are various ways of finding these not necessarily they may belong to the neighbouring concepts but they could be also sometimes artificially generated. So in scenario six that we were going to do now we will turn to another aspect actually of the dataset. So in we still talked about the same basic same kind of learning situation but in the earlier scenarios we have assumed no Internal Structure of the data set which means that all data items are been regarded as having the same status and importance and no structure or metric has been assumed among the data items in the set. And so because of one of that is because that they are all the same they can be handled in any order it doesn't matter in which order we treat the data in our analyses because they are all first class citizens and considered equal. So in contrast to that and we will now look into the situation where we really don't see them as equal that we actually first of all can say that some of the objects are like better or more typical than the others they are better representative from the concepts so there are better representatives of these concepts and less good members in the sense of typicality and of course then naively the more typical objects would be more advantageous to

start with in the learning process because you will start with those then then you get a kernel of a concept definition that is more stable than otherwise. Also in this situation we assume not only that objects are more or less typical but there is also some internal structure among the data and we can also talk about some measures of similarity between the various pairs of objects. And if we have such a similarity measure it also means that we can let the relation between the objects guide the order of considering the examples. So essentially the use of the would be to optimize the learning behaviour by the enhanced knowledge of the internal structure of the dataset. So let's turn to scenario 7, so the six scenario the last scenario implied a well-defined structure and similarity metric for the datasets with some tip typicality of objects actually this kind of situation open ups for a new scenario which is this one we're no longer we may need an explicit generalization our concept definition to be defined. So by Instance based learning or memory based learning we mean learning algorithms that instead of creating explicit generalizations, as the basis of making a prognosis in incoming cases, we just compare new problem instances with all the instances we already have and which we have already stored in our memory but of course in a structured way. In principle in this case of course the negative side of this is that when we have to get a new case we have to compare that with although one we match it against some abstract a general concept definition we have to compare it with all the into the store objects theoretically and this is computationally hard, however there are also advantages because the instance base model may typically have more it much more easy to adapt to the model to new data , so in this way one can say that the model is more agile because it can really change much faster with with the handling of new data and actually this kind of model then typically store each new instance and introduce it into their existing structure and in some cases and after some consideration one may also decide to throw away old instances because they may have been become obsolete. So in Scenario 8 we still keep two supervised learning but leave the offline case and go into the online case. So online learning is then a learning scenario where data is processed in real time in an incremental fashion. Input data are incrementally, gradually and continuously used to inductively extent than the existing model. So we still have one concept definition that we form and the picture here maybe is a little clumsy but you should interpret it that the sum of all the green circles are considered as the as the dataset so far, however we kind of harvest our data items in a time wise fashion and we treat in the data items of the data set when they come but, at least theoretically nothing is worse because it come earlier or later, they're in way equivalent in that sense, so when we treat new data items results were you learn entity maintained, so we know we can of course which was discussed in some of the earliest scenarios, decide to discard things for various reasons, discard data itself but not necessarily because of where they come into the flow o finput in this case. So actually the next scenario number nine focus exactly on this problem of what to do with data items that doesn't seem relevant anymore, and I mean we can say that in the continuous case we really can have more extreme situations where first of all the amount of data is gigantic this means that there are technical difficulties in storing and handling all the historical data. Also it can be so that this goes on all over a very long time which means that things that happened years ago still are not relevant by know. So in this extreme cases with a lot of data and very long time span it may be cleary so that you want to throw things away. And of course, when you throw things away you may also need to revise the concept definition because of course the concept

definition you have is based on all the analysis done so far so even if you analyze it some things the years ago it's still part of the definition but if you really think those old data is not relevant anymore then you have to revise that also the concept of the definition not only throwing away the data. Here you can see if you remember some of the earlier scenarios like instance based learning is that if you don't have a concept definition that's it's of course not a problem, so in that case you only have to have a throwaway so well that's one of the positive things with the instance-based learning if you really systematically want to tackle the situation with non-relevant data. So of course any algorithm that seriously and the online case must have some options for the discarding of objects and also rather in the definition of concept definition. So scenario 10, actually now moves from learning single concepts to multiple concepts and actually the scenario does not introduce many any big differences or surprises, I mean because as long as data are labelled, we in a way only multiply the problem and can handle it in parallel using the same kind of approaches so actually all aspects introduced in the scenario 1-9 are still relevant. So we still have to sound the handle noise, we still have to handle can must consider outliers, we can consider the negative examples for all the classes, we can look at near misses, internal structure is an issue, instance-based learning means an issue, online learning of course can take place also here and the question whether we should keep everything or discard or modify during the process is also still there. And so let's say that scenario 10 is like the complete supervised scenario with all its ingredients. So now we switch, we move from the supervised scenario to the unsupervised scenario. And now we look into the case where input data are not classified that means the input data only contains predictive features and no classification labels. So Unsupervised learning algorithms therefore have to identify commonalities and structures in the dataset and the group the input based on similarity. The main category of techniques that tackles the unsupervised case is called Cluster analysis. And maybe it's also it's not obvious to you that when we go into this realm, it becomes much more important to consider the issues of internal structure of the data set as brought up in some of the earlier scenarios, because if you don't really look into the structure of the data and similarity matrices among the data it will be very difficult to handle this this unsupervised situation. So therefore this kind of scenario of course in inherit most of the aspects in the earlier scenarios but in particular it is very important to consider the aspects of internal structure typicality, similarity matrices and so on of the data items of the data set. So even if we get back to these topics in coming weeks I added an extra slide here on cluster analysis, so which is the primary techniques one of the primary techniques to handle the unsupervised case. So actually cluster analysis then is the assignment of a set of observations into subsets called Clusters because when we start everything is unsorted, and so to see that the observations the data items within each subset are similar or more similar, than the distance is to objects to other clusters while observations drawn from different clusters are more dissimilar. So of course similarity metrics is a very important thing we have to have a possibility to measure the distances between data items. But then we will start to talk about a lot of topological things so actually it becomes very interesting to talk about the compactness or density of clusters, the degree of separation between them and so on. And so this is a whole new realm of its own. So actually I have put this slide here just to point at an important fact but maybe a trivial fact, is that when we now start to look at the potential groupings of unsorted data and look at clustering where we're trying to find optimal categories, it's not

only categories on one hand, so what was said on an earlier lecture about categories detractors is extremely relevant here, and of course any algorithm in the realm of unsupervised learning and clustering also mean potential mechanism not only to handle a flat category structure but also hierarchical structures. So in the end of this lecture now I take the liberty to rely on a little of learning by repetition. So in an earlier lecture I gave you the more or less this slide about the end-to-end process for concept learning and I think is very healthy to look at that now and then, because for most of this lecture we really have focused on the core data analysis phase again which is of course the core of machine learning, but we should never forget that the scenario we have for the analysis phase is very much dependent of course on all the earlier phases it's so much dependent on the kind of data we can acquire but we can manage this data how with how we can what theory we can establish, how we can engineer the features of the data items and so on. So putting everything we do in this context and have it always in the back of my mind is it's very important, because if we focus too much just on the core data analysis face, it is very difficult to keep this this big picture. So this is the end of the lecture so thanks for your attention this time now only remains one lecture for this week and this is the rather short tutorial for the assignments of the week thank you good bye