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 12
Decision Trees

Welcome to the second lecture of the third week of the course in machine learning this lecture will be about Decision Trees. So let's talk about decision trees so on some general characteristics of this representation so actually decision trees is something that been developed in decision analysis where tree can be used to visually and explicitly represent decision alternatives and this is making of options in various situations and typically a decision tree is drawn upside down with its root at the top. Nodes in the decision represents features, edges represent feature values or feature intervals, which of course the Feature value and Feature Intervals they can embody decision options. Leaves represent either discrete values typically than the discrete values and situations or classes. But it can also be that the leaves are present continuous outcomes, so this discrete case we talk more or less about classification of entities and situation while in the second case we talk about regression situations. In the decision-making scenario decision trees are typically used in a purely normative or prescriptive mode that we can actually set up a tree that is the guide for how to behave and the decision trees then simply stipulated, so if you get a simple example to the right you say this is a manual for how to act so you wake up in the morning you look if it's raining, okay it is it's raining, if it's not raining it don't do anything. Then in the other case you may say always windy it's raining but it we yeah but it's not windy if it's not windy I can bring an umbrella and so on and so on. So it's a little manual for her to decide yeah it's a slightly different setting in machine learning because in machine learning when we do work with decision trees we essentially want to build up at a relevant tree, from data sets collected in your domain, so is not predefined it's not prescriptive the tree should be more or less true to the data considered. So the tree challenge here is to design an optimal tree so that the tree makes the best fit of the considered data items and has the best predictive performance for new data items.

So the interdisciplinary sources of inspiration for this Representation as has already been mentioned is decision-making theory in economy in business, which have for a long time used simple models of this kind. It has also been said is that these models are normally defined and they are prescriptive. But there is an interesting parallel of course because we will see when we build try to build this kind of trees from data sets and with the aim of having a true picture close to the data set, of course we hope that even if a person or some persons just this define this tree, the way they do it is indirectly and informally based on earlier experiences from their site this is the normal case. So at some point learning takes place however it has done informally among the people involved or we do it formally in the form of machine learning. So the core component and the core problem solving for the representation is as follows. So when you construct the tree, you construct it upside down, the root to the top, if you do this manually you intuitively choose some order of features and the features you represent as notes and from each level from each node you branch the tree by looking at the values so of course the edge is represent on every level feature values or featured intervals. And finally when you come to a leaf that leaf represents the discrete or continuous outcome. So that's the build up of the tree the use of the tree is also straightforward you start from the tree the top and evaluate the values of the features in the given order eventually up with a leaf with a unique outcome. If we now turn to learning for this representation which means that we want to build up a decision tree for a realistic domain with many features and many data items.

So then we consider two kinds of decision tree analysis, one we can call classifications Tree analysis where the leaves are the classes of the categories that we want to define or regression tree analysis when the leaves are intervals or real numbers. The challenge is to design a tree that that this tree optimizes the fit of the considered data attempts and the predicted performance for still unseen data types minimal prediction error basis is aimed at and it's not trivial because we may have many features to look at and it's not crystal clear for most domains which is the most important or discriminatory feature to use first because of course the idea is the to start building the tree using the features that that best discriminates among the data and there are several approaches that can be combined in addition to learning techniques so one kind of technique is a kind of proactive which means that at every point we evaluate the whole dataset with respect to a potential selection of our feature and we analyse the discriminatory power of that feature using different kind of approaches. So and typically the way we evaluate is by looking if some kind of information theoretic measures on the whole situation and their approach is referred to as information gain beta of some an entropy concept also approaches called Gini impurity measure, but the purpose in general in all these cases is to judge the discrimination power of the feature. Another kind of measure to take is to first contract the tree but then at a later stage in the process proven it and there of course also different techniques and criteria for what is a good way to prove. A third approach is to actually several trees instead of just one in parallel, so this means that through the process the design process we would we generate not just a tree but a forest, and then further on in the process evaluate which is the most optimal tree of within the forest. One criteria that is always important to have in the back of ones head in these matters is the very well-known principle called Occam's razor but of course if there is a choice of structure you should always prefer the simplest choice or the simplest excellent tree in this case.

So if we look at what happens to the data set when we fabricate preliminary tree we can observe that what always happened is that if we have a population at the start with so in this example we have like

14 data items in in the root and then we make the first split, use one feature and we split up in in three possibilities which are the values of that feature. So then what happens is that by doing that we also in a way split the data set or the populate data set into three boxes. And then in the second step when we make the second split using some other feature we make a split again and of course it should always hold that every step is displayed and the sum of the items in the box is created should sum up to data set which was 14, so as you see here on for the first plate and we divided up in three boxes and when we take the second step and we don't go further with the third option we go oh yeah that's split further the two first option but altogether the five from resulting leaves in that case partitions the data set as a passive whole. Another perspective on what's going on in this process is to compare a preliminary tree with some other graphical depiction of the feature space, so if we in this very simple case have to two features and we in a way split with respect to one breaking point for each of these example illustrates how the three constructed partitions a two-dimensional depiction of this little features set. So as you can see here there is one quarter of the area represents one option and three quarters of the area that represents the other option or outcome.

So I will end this lecture by showing you a few examples so the intention with the first example to show you like a complete picture where you get a whole data set with a number of data items and then a corresponding tree, whether that tree is optimal we are not going to discuss at this point we will come back to those issues next week when we look into a variety of algorithm that has been designed for this kind of problem. So this example is related to the analysis of election outcomes in the u.s. essentially the features in this example are the outcomes in various states, so this means that the ordering of features is based on the importance of those outcomes for the final results and as you see this has been done here in such a way that the tree is reasonably well behaved in the sense that one of the alternatives the what is depicted as blue here are kind of more dense to the to the left and the other alternative the red outcomes are sorted to the right. So I give you two more examples here I won't say much about them just a few comments the left one is a classification example and the right one is a regression example. So in the in the left one the classification is essentially in two outcomes a child gets a Christmas gift or a child doesn't get Christmas gift those are the two only outcomes on this and of course as you see the outcome can depend on certain behaviours of the child and I are no further comments on the regression tree example, taken you see here how what do is illustrate it earlier that the population of the or the data items in the data set are on every level distributed among the leaf nodes at that point. So this was the end of this lecture thanks for your intention the next lecture will be on the topic of Bayesian belief networks thank you good bye