Pivotal

Manish Tripathi <mtripathi@pivotal.io>

Epilogue: Thursday Thunder- Bagging, Random Forests, Ensemble

Manish Tripathi <mtripathi@pivotal.io> Mon, Mar 21, 2016 at 1:14 AM
To: PDS-ALL <PDS-ALL@pivotal.io>

Hi

I was doing my laundry a while back and was listening to the podcast of (TheTalkingMachines) on my phone in the laundry room when in one of the talk someone mentioned about Bagging and Random Forests.

I realized couple of team members have asked the same and I don't want to feel guilty by not answering it. So here is an Epilogue to the Thunders. As usual will miss sending these. :-|

Off.. You again came back to haunt?.

No. Just to end it with some answers to unfinished questions. You won't be bothered after this. Sorry 😕

Ok. So what is Bagging and Random Forests?.

Well, lets talk about one Ensemble (Bagging) and one special case of Bagging- Random Forests.

Can you make it simple?. These terms look very complex and scary.

Let me try from whatever little knowledge I have gained*.* This will be long so pay attention.

Ok.

So here is a situation*.*

Suppose you want to plan a vacation but your are very indecisive on the choice of place to visit this time.

So you ask recommendations of places to visit from Vatsan to see if he thinks you will like it.

Before answering your questions Vatsan needs to figure out what places you like so he asks you the list of places you have gone in the past and whether you liked it or not (**i.e you give him a labelled training set**).

Now when you ask him if he thinks you will like place X or not, he plays a game of 20 questions with you by asking you questions like -*"Do you like domestic or international locations"* or *"Do you like beaches or mountains"* and so on. He asks questions which give him more information first (**i.e he is maximizing the information gain of each question**) and gives you a Yes or No answer in the end for a choice of place.

Vatsan is **your decision tree for places recommendations.**

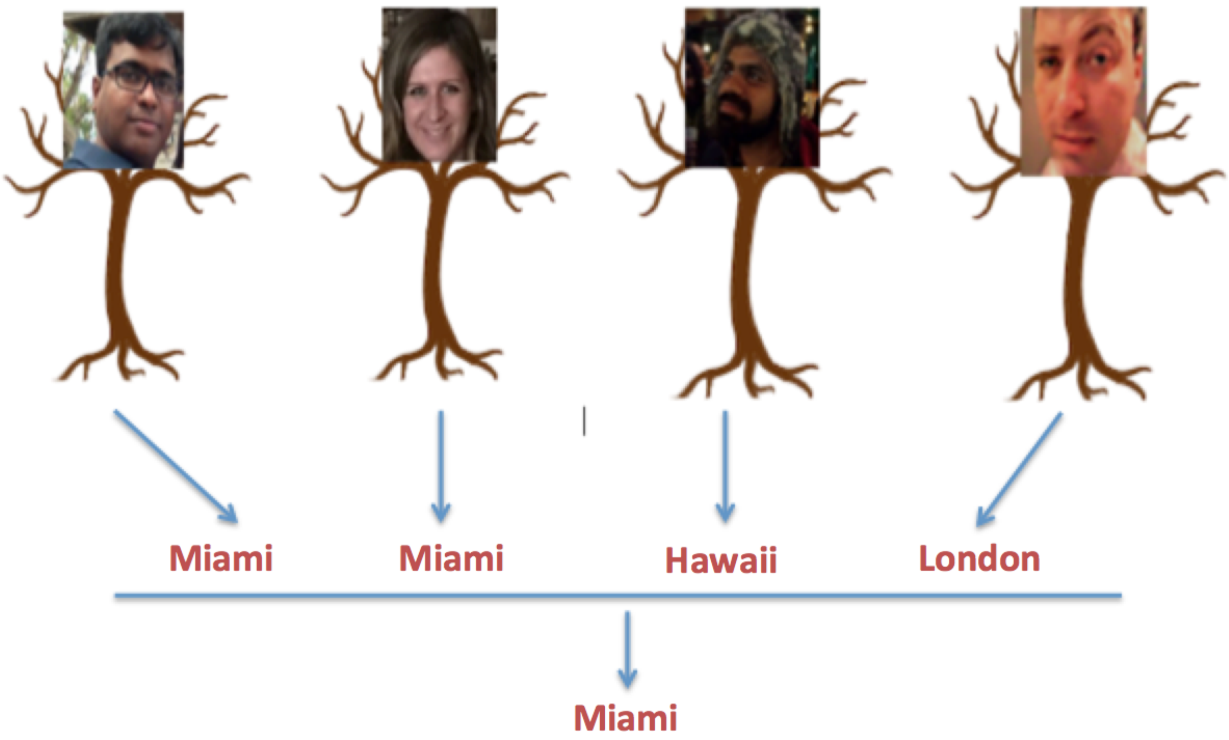
But Vatsan is usually drunk and would make mistakes and not generalize your choices well (**i.e he will tend to overfit**). So to get more accurate recommendations you decide to ask a bunch of your friends and go to a place X if *most of them* say you will like that place.

So now instead of only asking Vatsan, you ask Sarah, Ambarish and Michael and they vote on whether you will like a place or not (i**.e you build an ensemble of trees aka forests**).

But you don't want all of them to give you the same suggestions by having the same information. So you change the data which you give them by providing them slightly different information. You are perturbing the data.

Maybe you are not sure yourself of the preferences. So while you tell Vatsan that you ***liked*** Miami, you tell Sarah that you ***really loved*** Miami and that maybe she should give more weight to it. Your love/hate decision remains the same, you just say you love or hate a place *little more or less* (**i.e you are giving a bootstrap version of your original training data to your friends**). For example while you tell Vatsan, you ***liked*** Miami, ***loved*** Hawaii, and ***liked*** NYC, you tell Ambarish you ***loved*** Hawaii so much that you went twice, ***liked*** Miami and no mention of NYC.

Your friends will now ask you the *same* questions and give their preferences which they think you might like. You take all of their recommendations and aggregate them and take a choice which is most suggested (i.e **your friends now form a bagged (b**ootstrap **agg**regated**) forest of your vacation place preferences).**



But you are not happy as there is still a problem with the data. You don't want all of your friends to ask the same questions and make a suggestion based on that.

While You liked Miami and you also liked Hawaii, but maybe it's not because of the beach and sand but for some other reasons. You *don't want* all your friends to base their decision on the *same* question -"Do you like beach and sand?". So you don't allow all of them to ask the same question-"*Do you like beach and sand"* whenever they want. Now when each friend asks a question, only a random subset of possible questions are allowed (i.e **each decision tree node can have only a random subset of attributes to split on).** So earlier while you introduced randomness at data level now you ***also*** introduced randomness at attribute level by *not allowing* all your friends to ask the same questions.

Vatsan, Sarah, Ambarish and Michael now form a Random Forests of your vacation places preferences.

Cool. Understood. Hope you could do more :-(. Provided you talk less 😠
Yeah, but it ends here. Thanks!