Datasets Complexity

Dieudonne

5/2/2018

Motivation:

What we are currently doing now in ML we look at the data, we try several algorithms, we see the accuracies and compare and select the best model.With another dataset, we will have to retrain and pick other algorithms and see again .This is called "no free lunch" theorem: no one algorithm performs well for all datasets, this leads to a question: can we determine or estimate the dataset complexity and assign the right algorithm(s) before training models?.For models that run for days before returning results, I think running many models and waiting to see results is not efficient.

Supervised Learning

I will talk about the supervised learning context, but I believe we can generalize on the unsupervised side also.

Dimensions

n: number of rows

p: number of columns

In trying to predict the outcome of an event(s), we have a dataset let say D with n rows and p columns. One column is the output or target we are trying to predict based on a combination of other columns (p-1 or sometimes less). We know that if we don't have enough data let say n<p-1 it is hard to even in some case impossible to find the right predictions on a new test dataset. So criteria to test on a dataset is the relationship between n(number of rows or observations) and p-1(the number of possible inputs),n must be large enough to capture the entire space of possible solutions.So I believe this must be quantifiable in some ways! Also even though n must be large enough, the observations must be diverse(repeated instances should not count much cause we don't learn anything by having the same thing repeated) for better predictions.

Output type(Target Type)

If we take a target variable which is discrete we know that using a decision tree is better, we know that if the target value continues and the inputs variable are continuous most likely a linear regression is better than the decision tree. So there must be a way to assign a value(or an indicator) to output variable type(maybe one if it continues or 0 if it is discrete, based on that we can choose the correct algorithm to tackle the problem).

Inputs types

If an input is discrete, we can measure the entropy, that measure could be used in the complexity of the dataset D. If the input is continuous,the variance, mean and the type of distribution of the input can play a role in determining the complexity of the dataset.

Relationships between inputs

The relationship between the inputs (correlations I mean here) could be indicators of the complexity(computing the correlations we can find the number of uncorrelated(defining a threshold) inputs and this could be a good indication of the complexity of the dataset).

Relationships between inputs and output

a two dimensions plot(scatterplot) can give indications

Missing values

Missing values either in input variables or the target(the output) affect the predictions, so this should be incorporated into the complexity of the dataset.

Presence of outliers

The presence of outliers(values that are too large or small) aways of the typical values in the inputs should be part of the complexity as well because most machine learning algorithms perform poorly with outliers in datasets.