1.Introduction

1.Machine learning is the process of learning examples by machine itself, so that, it can generalize the knowledge on new data.

2.Given set of data of which each item has attributes and a certain class, the classifier can learn from this data. After the learning step, given input of same kind of data item which has attributes, the classifier can predict the class of such item.

2.Learning

1.The three components are representation, evaluation and optimization.

Representation is implicit rules of data attributes, and a proper representation can fit the data well.

Evaluation weights the performance of learners, so people can tune the learners based on evaluated outcome.

Optimization is the method of fitting learners to the data.

2.In the process of constructing decision tree, we need to determine what attributes to choose to classify data from root to leaves. And decision tree takes the attribute which maximizes the information gain to be the next attribute.

Information gain is difference between two information entropy. One is calculated based on n attributes and the other on n+1. If the information gain is large, that means the added attribute has useful information in classifying data. The following is the information gain equation.

3.Generalization

1.Our goal in machine learning is learning and then acquiring a high performance on certain tasks. High accuracy came from the learner does not mean it can get a high score in new data which is not labeled, instead there might be problems like overfitting in the learner. Generalization is the capability to apply knowledge to new data. Conversely, incapability to generalization means the leaner does not learn much meaningful knowledge.

2.In cross-validation, data set is split in n subsets. Then the data will be trained n times. In each time, only n-1 subsets are trained, and the one left behind will be used as test set.

In cases where data is not enough, this technique can make full use of all the data. In some time, we don’t test the learner on test set until the last moment. So, cross-validation can be applied to training data to measure the capability of generalization of the learner. Particularly, if only one fixed validation set is selected in training, and error rate of this validation set is used to tune the learner, this might add bias to the learner. Because data is not homogeneous all the time. The average accuracy from cross-validation could be a better measure.

3.In other optimizations, there exist explicit functions to be optimized while in generalization, there doesn’t exist. For generalization, there is no way to optimize to the true goal, and training error is the only measure we can rely on.

4.Data alone is not enough

1.If the function includes 10 boolean variables, there would be 210 possible examples. 100 examples accounts for 9.77% of these instance space.

2.In some cases, the data people have usually not enough. Data or example they have might only accounts for part of the instance space. It is clearly that, without other knowledge, it is very difficult to learn from this limited data. “No free lunch” means, only data alone is not enough to train a learner pretty well, but other prior knowledge like assumptions can be added to make it a good learner.

3.Those general assumptions are smoothness, similar examples having similar classes, limited dependences or limited complexity.

Induction is turning a small amount of knowledge in to large amount of knowledge.

4.Farming is collecting seeds and let them grow with nature. Whether they would grow well depends on the seeds. Data like seeds, programmers use their knowledge to choose data, preprocess data and construct programs for the data, then the programs just learn by itself.

5. Overfitting

1.In case of overfitting, accuracy from training data is almost near to 100%, but accuracy from test data is far below that. Overfitting has low capability of generalization but only fits the training set. It leads people to the wrong idea that the learner does a good job on training set with its high accuracy rate.

2.Variance like noisy signal added to the true signal. Despite variance of noise in output, bias is overall deviation to the true output.

3.Methods to combat overfitting are cross-validation, regularization, statistical significance test, etc.

6. Intuition falls in high dimensions

1.In high dimension, the instance space becomes exponentially large, however, the training set is fixed size. And training set would account for a lower portion to the instance space. And higher dimension adds complexity to the task as well as the computing time.

2.If data is distributed uniformly in the search space, each degree of dimensions plays equal roles in determining the output. However, data is often non-uniformly distributed, which means some degrees have little effect on the output and we can actually ignore them.

7. Theoretical guarantees

1.One of the developments is we can have probabilistic guarantees on the results of induction.

8. Feature engineering

1.Feature is the key factor for machine learning’s success. If features are of simple function to the output or it is easy to learn from the features, machine learning is easy to succeed.

2.Feature engineering is more time-consuming.

Machine learning process is iterative because the model needs to be tuned after each training process. To improve the performance of model, modifications on assumptions of features and even the representations are sometimes necessary based on previous measure and output.

3.One of the holy grails in machine learning is automating more and more of the feature engineering process.

9. More data beats a cleverer algorithm

1.The two things are, we can choose a cleverer algorithm or use more data. And the latter one is better.

2.3 limited resources are time, memory and data. Time is the bottleneck today. One of the solutions to this is coming up with faster algorithm.

3.Even different representations draw different boundaries or frontiers, training data is not drawn uniformly in the instance space. Various types of frontiers can be applied to the instance space resulting similar classifications. Simple learners can be used at first.

4.There are two major types of learners: fixed-size learners and variable-size learners. The fixed-size learners have fixed size of representation no matter the size of data. Representation size of variable-size learners variates depends on the data. When data is large, its representation include more parameters resulting much more computational cost and time. But variable-size learners can usually deal with problems much more complex than fixed-size learners.

10. Learn many models, not just one

1.Yes, it is better to have more variation of a single model of have a combination of different models. Ensemble techniques are bagging, boosting, stacking, etc.

Combining more than one model enables the model to learn much more complex hypothesis and can reduce bias introduced by single model.

11. Simplicity does not imply accuracy

1.Simplicity does not connect to theoretical accuracy, but it by virtue of itself right according to Domingos.

13. Correlation does not imply causation

1.Even observational data is usually predictive, it might have potential predictive effect on the result. Correlation can be considered as effect to some action, indirectly effecting causal variables.