Algorithm primitives

Master Thesis

L.D. Stooker, 0819041

April 6, 2018

Abstract

To improve existing automated picking of a machine learning algorithm

Contents

1 Introduction

This report is the result of my graduation project which completes my Business Information Systems study at Eindhoven University of Technology. The project was performed internally at the University in the Data mining department. In this project we investigated annotations of primitives, more specifically primitives in the scikit-learn library. In the section 1.1 we will briefly explain more about primitives and the annotations. To elaborate on this we will outline the research questions and thesis structure

1.1 problem description

Machine learning is a growing field that can help process the increase of available data[4][3]. Python is a language which holds premade machine learning algorithms in libraries like scikit-learn[**?**]. In recent years python is also increasing in so called market share for machine learning[2]. To help machine learners using the scikit-learn library made a model to indicate what algorithm to use for what problem. In figure **??** you can see that depending on size of the data and early results different algorithms are recommend.

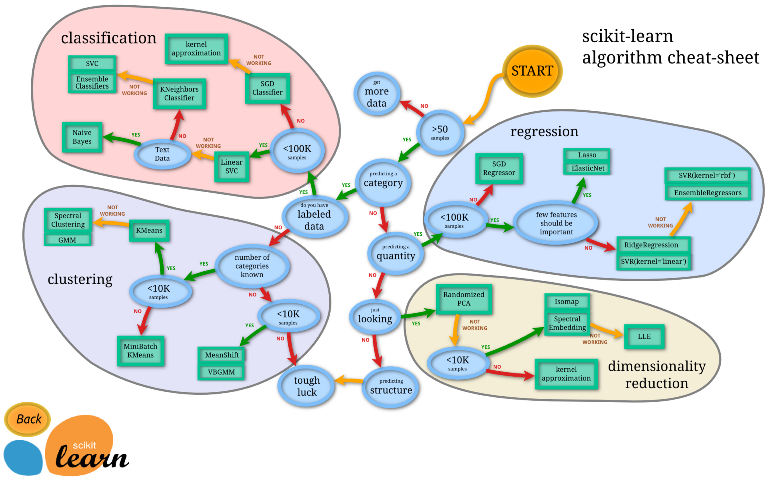

Figure 1: FlowChartML

1.2 research question

We base our research question on the work of Joaquin to give properties to algorithms[5]. More specifically we look more closely to the resilience properties. Earlier research has been done on scalability and resilience to irrelevant variables[6]

1.3 thesis structure

1.4 Outline

2 Preliminaries

Before we discuss in detail the solutions for the steps of our approach, this chapter provides some background knowledge and definitions which are required for a good understanding of the remainder of this thesis.

2.1 Sklearn/scikit-learn library

The scikit-learn library is based in Python and is made to make machine learning in python accessible and organized. All resources are open source and hosted on Github.

2.1.1 KNeighborsClassifier

In the scikit-learn library KNeighborsClassifier is an implementation of the k-nearest neighbors algorithm(k-NN). K-NN uses instance-based learning or non-generalizing learning. This means that during fitting no complete model is made but only the given feature set is stored in order of appearing in a tree for example. The k-NN uses as it names tells the nearest neighbors for calculation. The inputted feature set is traversed to find the nearest points and depending on the parameter k, an amount of k points is searched for. The default metric for distance measuring is Euclidean distance another option is the Manhattan distance which is less accurate but needs less computing. To find the nearest points an option can be made between a ball tree, a kd-tree or a brute search. This can heavily influence the search time, depending on the amount and size of the input (features and instances) this can influence the prediction time heavily but will not the influence predictive accuracy. The previously mentioned k parameter is an influencer for prediction quality—find source—

2.1.2 GaussianNB

GaussianNB is a naïve Bayes classifier implementation assuming the feature set is gaussian distributed [11]. For fitting the data, a partial fit function is used based on the work of Chan, Golub and LeVeque [7]. This calculates the assumed means and variances of a gaussian distribution of the inputted feature set. Based on this distribution the prediction is made by filling in the maximum likelihood. The limited calculation needed for classification and prediction makes this one of the fastest algorithms. The only parameter of this classifier specifies the prior probabilities of the classes, which will when specified not be adjusted to the given input.

2.1.3 BernoulliNB

BernoulliNB is a naïve Bayes classifier implementation assuming a Bernoulli distribution with Boolean like values [12]. The first step of this implementation is checking if the features are binary-valued, if any other data is found this input will be binarized. This setting can be disabled or reduced by a threshold for the input. Based on this Boolean model a smoothed version of the maximum likelihood is used for prediction. This classifier is mostly used in document classification as it can binary store occurrence useful for prediction class probability.

2.1.4 SVC-rbf

SVC-rbf is a support vector classifier(SVC) implementation with a radial basis function. The radial basis function(rbf) is used to handle a large feature dimension, since the standard support vector machine splits the spaces with linearly lines computation grows too large for a large feature space [10]. The fit time is already quadratic with the number of samples based on the implementation of libsvm[8]. The fitting of a SVC will assign each example to one of two categories and will represent them in a dimension space mapped so there is a clear separation between the two categories. With the radial basis function this is with the distant from the points indicated by a separation area. Classifying a point is finding in which class area this point falls. For a multiclass problem this is done in pairs of two for all categories and then the most voted class is picked [9].

2.1.5 RandomForestClassifier

RandomForestClassifier is an ensemble method of directive trees [14]. During fitting a random forest classifier constructs directive trees on subsamples of the input data and averages the results for the result. These sub-samples are chosen randomly so the results can vary between runs on the same input. A directive tree is a decision tree classifier which splits the features on certain thresholds to decide on the type of class. This splitting of the data is either randomly or choosing the best split, to measure this split a criterion is used like Gini or entropy. The amount of splits, features and samples are also considered and can be inputted.

2.1.6 AdaBoost

AdaBoost is an ensemble classifier that fits other classifiers and outputs the weighted results of those classifiers [15]. AdaBoost trains these other classifiers on previously misclassified results by increasing their influence this makes it heavily subjected to noisy data and outliers. The scikit-learn library uses the multi class AdaBoost-SAMME implementation from J. Zhu et al [16]. The solution of J. Zhu also solves the lack of multi-class solution of the weak learners (other classifiers) by extending the initial AdaBoost algorithm with a forward stage wise additive step. In this step a continual calculation of a loss function will output the prediction and in a two class case it reduces to the initial solution.

2.1.7 SDGClassifier

SDGClassifier is an incremental function to stochastic approximate the gradient descent of the input [17]. It will iteratively minimize or maximize a set of differentiable functions, the input must fit these differentiable functions and this makes than an optimal input is with a mean of zero. This makes the classifier sensitive to raw data as it performs optimally with sparse features. The iterative steps needed are bound by the inverse of the learning rate and a threshold value. The threshold value indicates what degree of slope indicates a near minima or maxima. The learning rate is used to update the model in each iteration.

2.1.8 GradientBoostingClassifier

GradientBoost is an ensemble classifier that builds from weaker classifiers [18]. Like AdaboostClassifier it builds an additive model in a forward stage-wise fashion. The weak classifiers used in the scikit-learn implementation are decision trees.

2.2 Terminology

3 Experimental setup

3.1 Motivation

3.2 Description

4 Experimental Results

5 Discussion

6 Conclusion

7 References

References

[1] The Popularity of Data Science Software, Robert A. Muenchen, r4stats.com, (2017)

[2] Most Popular Programming Languages For Machine Learning And Data Science,Adarsh Verma, fossbytes.com, (2016)

[3] Machine learning: Trends, perspectives, and prospects, M. I. Jordan, T. M. Mitchell, Science Volume 349 issue 6245 pages 255-260, (2015)

[4] Storage predictions: Will the explosion of data in 2017 be repeated in 2018?, Nick Ismail, www.information-age.com/, (2017)

[5] Understanding Machine Learning Performance with experiment databases, Joaquin Vanschoren, KU Leuven, (2010)

[6] Quantifying the resilience of inductive classification algorithms, M. Hilario, A. Kalousis, Proceedings of the 4th European Conference on Principles of data mining and knowledge discovery, pages 106-115, (2000)

[7] Updating Formulae and a Pairwise Algorithm for Computing Sample Variances Tony F. Chan* Gene H. Golub'* Randall J. LeVeque, Stanford CS tech report STAN-CS-79-773(1979)

[8] LIBSVM: A library for support vector machines, Chang, Chih-Chung, Lin, Chih-Jen, ACM Transactions on Intelligent Systems and Technology (2011)

[9] Probability estimates for multi-class classification by pairwise coupling, Wu, Lin, Weng, Journal of Machine Learning Research 5 (2004)

[10] Support-Vector networks, Cortes, Corinna, Vapnik, Vladimir, Machine Learning Volume 20 issue 3 (1995)

[11] Idiot's Bayes—Not So Stupid After All?, David J. Hand, Keming Yu, International Statistical Review Volume 69 Number 3(2001)

[12] A Comparison of event models for naïve Bayes text classification, Andrew McCallum, Kamal Nigam, AAAI-98 workshops on learning for text categorization (1998)

[13] An introduction to kernel and nearest-neighbor nonparametric regression, N. S. Altman, American Statistician 46(3): 175-185 (1992)

[14] Random Decision Forests, Tin Kam Ho, Proceedings of the 3rd Iternational Conference on Document analysis and recognition (1995)

[15] A short introduction to Boosting, Yoav Freund, Robert E. Shapire, Journal of Japanse Society for artificial Intelligence 14(5):771-780 (1999)

[16] Multi-class AdaBoost, J. Zhu, S. Rosset, H. Zou, T. Hastie, Statistics and its Interface volume 2, pages 349-360 (2009)

[17] Solving Large Scale Linear Prediction problems using stochastic gradient descent algorithms, T. Zhang, ICML Proceedings of the 21 International conference on machine learning (2004)

[18] Stochastic Gradient Boosting, J. H. Friedman, Computational Statistics & Data analysis – Nonlinear methods and data mining volume 38 issue 4 pages 367-378,(2002)

[19] Greedy function approximation: A gradient boosting machine, J. H. Friedman, The annals of statistics volume 29 issue 5, pages 1189-1232,(2001) Previous bias-variance research has shown a trend of larger dataset increasing the bias component but still fluctuating with less than 10000 instances.

[20] Experiment databases, J. Vanschoren, H. Blockeel, B. Pfahringer, G. Holmes, Machine Learning Volume 82 issue 2 pages 127-158, (2012)