Noteworthy - The Journal Blog
The Official Journal Blog
Follow

A Comparison of Grid Search and Randomized Search Using Scikit Learn

Peter Worcester | Follow
Jun 5, 2019 · 4 min read

Introduction:

If you're planning on doing Machine Learning in Python, Scikit Learn sets the bar.

Scikit Learn offers various important features for Machine Learning including classification, regression, and clustering algorithms. Scikit Learn is designed to inter-operate with python libraries such as NumPy and SciPy and is by far the cleanest Machine Learning library.

Scikit Learn supports a host of models for both supervised and unsupervised algorithms including:

- **Regression:** Fitting both linear and non-linear models
- **Support Vector Machines (SVMs):** Supervised method for classification
- **Neural Networks:** Supervised learning algorithm that learns a function by training a data set
- **Decision Trees:** Tree induction and pruning for classification and regression
- **Clustering:** Unsupervised classification
- Many others…

In addition, Scikit Learn offers some advanced functions not typically offered by other libraries:

- **Outlier detection:** Fits a regression model in the presence of corrupt data
- **Ensemble methods:** Bagging, Forests, Gradient Tree Boosting, and Model Voting
- **Feature selection and analysis**
- **Model selection:** Cross-validation, hyperparameter tuning, and metrics

Objectives:

In this article, you'll learn an introduction to model selection. Specifically, you'll learn the process of tuning hyperparameters and the difference between Grid Search and Randomized Search

Grid vs. Random Search:

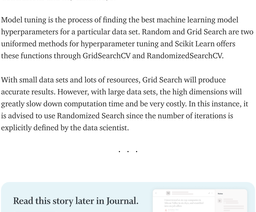
In contrast to model parameters which are learned during training, model hyperparameters are set by the data scientist ahead of training and control implementation aspects of the model. Hyperparameters can be thought of as model settings. These settings need to be tuned because the ideal settings for one data set will not be the same across all data sets. When tuning the hyperparameters of an estimator, Grid Search and Random Search are both popular methods.

Grid Search can be thought of as an exhaustive search for selecting a model. In Grid Search, the data scientist sets up a grid of hyperparameter values and for each combination, trains a model and scores on the testing data. In this approach, every combination of hyperparameter values is tried which can be very inefficient. For example, searching 20 different parameter values for each of 4 parameters will require 160,000 trials of cross-validation. This equates to 1,600,000 model fits and 1,600,000 predictions if 10-fold cross validation is used. While Scikit Learn offers the GridSearchCV function to simplify the process, it would be an extremely costly research point in computing power and time.

Top highlight

By contrast, Random Search sets up a grid of hyperparameter values and selects random combinations to train the model and score. This allows you to explicitly control the number of parameter combinations that are attempted. The number of search iterations is set based on time or resources. Scikit Learn offers the RandomizedSearchCV function for this process.

While it's possible that RandomizedSearchCV will not find as accurate of a result as GridSearchCV, it surprisingly picks the best result more often than not and in a *fraction* of the time it takes GridSearchCV would have taken. Given the same resources, Randomized Search can even outperform Grid Search. This can be visualized in the graphic below when continuous parameters are used.

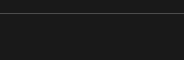


With grid search, nine trials only test three distinct places. With random search, all nine trails explore distinct values.

Application:

In order to compare the efficiencies of the two methods, I analyzed mobile phone data. Specifically, I analyzed the relation between features of a mobile phone (eg: RAM, Internal Memory, etc) and its selling price. The database has 2000 data points and was built from collecting sales data of mobile phones of various companies.

I was surprised to see that Grid Search performed better so I performed the same operation on a dummy data set with the same number of samples and features using Scikit Learn's "make_classification" function and compared the two again.



Once again, the Grid Search outperformed the Random Search. This is most likely due to the small dimensions of the data set (only 2000 samples). With larger data sets, it's advisable to instead perform a Randomized Search.

Conclusions and key takeaways:

Model tuning is the process of finding the best machine learning model hyperparameters for a particular data set. Random and Grid Search are two uniformed methods for hyperparameter tuning and Scikit Learn offers these functions through GridSearchCV and RandomizedSearchCV.

With small data sets and lots of resources, Grid Search will produce accurate results. However, with large data sets, the high dimensions will greatly slow down computation time and be very costly. In this instance, it is advised to use Randomized Search since the number of iterations is explicitly defined by the data scientist.

. . .

Machine Learning | Scikit Learn | Hyperparameter Tuning

WRITTEN BY
Peter Worcester
Follow

Noteworthy - The Journal Blog
The Official Journal Blog
Follow