Problem Solving in Dataset Using Support Vector Machine with Python

Faza Thirafi - 13514033
Informatics Engineering Department
School of Electrical Engineering and Informatics
Institut Teknologi Bandung, Bandung, West Java, Indonesia
faza.thirafi@gmail.com

***Abstract*—This paper contains the author research about a tool for Machine Learning topics that used in Python packages. Main packages in Python that provided to help us to solve Machine Learning problems are SciPy, NumPy, and Scikit Learn. This research mainly talking about solving a problem utilizing those tools. Here, the author solved a simple problem in Machine Learning which specifically called "Support Vector Machine" or popularly abbraviated as SVM. In SVM, the goal that should be reached is finding "vector" to make barrier between 2 classes in classification. So this concept would help us to classify new data easier.**

Keywords—machine learning; support vector machine; python; classification; regression

I. Introduction

Machine Learning is considerable to be a solution for many problems either in Artificial Intelligence itself, or in supporting business in general. Many companies (even for non IT companies) consider to use Machine Learning to enhance their business process, such as understanding customer's needs, predicting future trends, and many more. So, as a student of Informatics Engineering or Computer Science, learning this topic is a must nowadays.

Many algorithms, theories, and concepts included in Machine Learning. One of the most important algorithm in Machine Learning is Support Vector Machine (SVM). As stated by Lamp (2012), SVM is used to help engineers in classification or regression problems. In a dataset, with SVM and Kernel trick we could find the optimal boundary that separates 2 classes in a diagram called hyperplane. So, it would optimize the classification ability to our algorithm in machine learning.

As an important theory, SVM was developed by Machine Learning researchers since invented in 1992. Beside of theories, engineers also build up some tools to simplify solving problems with computer programs. One of them, Python provides programmers in analyzing and visualizing a dataset into graphical diagram which easy to understand. Let's say Scikit Learn and SciPy for data analysis, then pandas and matplotlib for data visualization. Those tools are important to solve this problem.

So, this paper would talking about how to solve a simple Machine Learning problem with SVM and using tools in Python. Both theory and practices are included, but this paper is concerned in using the tools.

II. Main Theory

A. Machine Learning

Machine Learning is one of Artificial Intelligence field which mainly discusses about how computer could "learn" to a new data input without explicitly programmed. It is also talking about data classification, classifier building, and data regression for data analysis.

As Marr (2016) points out, Machine Learning research started in 1950 when Alan Turing created "Turing Machine" which tested computer to "fool" human with being undifferentiable from human in the way of communication. After that, many researches and invention contributed to build up a new field of knowledge in Computer Science named Machine Learning as a branch of Artificial Intelligence.

B. Support Vector Machine

By definition, Support Vector Machine is a "machine" that support problem solving with vector. SVM was invented by Vapnik, Boser, and Guyon in 1992 with a paper for COLT (New York). Well, if we compare this algorithm to the others in Computer Science, it's relatively new. But, SVM will perform better in many application than other, such as bioinformatics and text classification.

In Machine Learning, Support Vector Machine is included as supervised learning that is learned with defined and fixed classes. As we know, in classifying data if we use machine learning, we want to know the pattern that occurs.

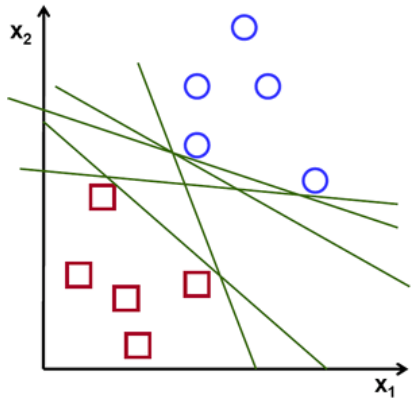

Fig. 1. Boundary possibilities (source: http://opencv-python-tutroals.readthedocs.io/en/latest/_images/svm_basics1.png)

In figure above, we can see that we have many possibilities to define the boundary that separate each classes. Those boundaries are valid to separate the area of "red box" class and "blue circle" class. So, Support Vector Machine helps us to find an optimal solution among them. After finding the optimum boundary, in Support Vector Machine there is also algorithm to find the maximum margin of separation.

C. *Python*

Python is a "powerful" programming language which provides programmer with many packages and libraries. Beside of pure programming, with Python we could do advanced tasks such as analyzing datasets, making Graphical User Interface (GUI), building remote networks, etc.

Python was originally founded by Guido Van Rossum in 1991. This programming language is relatively new if we compare to C, Java, etc. But, the packages available for Python are very helpful for programming development. Futhemore, Python packages are available on internet, or we could install it through command line with pip commands (package manager for Python).

III. Experiment

First of all, we have to ensure that our CPU machine or Laptop has Python installed into it. There are two versions of Python that is Python 2 and Python 3. Both version have few differences. Each version has specific packages for the same package (ex. NumPy). So, we must ensure that the packages is suitable to our Python version. In this case, the author uses Python ver. 2.7.

Then, we need the packages available in Python to support the experiment. Here is the packages which the author uses.

- NumPy

  This package is used to handle dataset

- Matplotlib

  In this experiment, this package is used to sketch the diagram and show the GUI

- Scikit Learn (Sklearn)

  This package is the main tool here. Sklearn provide many datasets as a "toy" (stated in Python Documentation) for experiment purpose. Then, Support Vector Machine also available as a class in Python. Therefore, as the author stated before Python is a "powerful" programming language.

After those Python configuration, we could start the experiment.

In this paper, the author uses dataset which is provided in Sklearn, that is **Breast Cancer data** which has two classes as tumor kinds in breast cancer. They are "malignant" and "benign". Actually, there are 32 features in complete. But, In this case the author uses first 2 of them to simplify the diagram drawed. They are "Clump Thickness" and "Uniformity of cell size".

Now, let's discuss about how to use SVM in e. First, we should import the data and define the mesh (h)

```
cancer_data = datasets.load_breast_cancer()
X = cancer_data.data[:, :2]
y = cancer_data.target
h = 0.02
```

It will automatically results the training data (X) and its classes (y). With the defined data, we can call the `svm` classifier from the imported svm class in sklearn. Here, the author use SVC with linear kernel. Then, with fit method we could "fit" the data that we have.

```
C = 1.0  # SVM regularization parameter
svc = svm.SVC(kernel='linear', C=C).fit(X, y)
```

As SVM defined here, we can use svc variable to find the optimal hyperplane from data. After that, for the plot which we draw, we have to define the mesh

```
x_min = X[:, 0].min() - 1
x_max = X[:, 0].max() + 1
y_min = X[:, 1].min() - 1
y_max = X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max,
h),np.arange(y_min, y_max, h))
```

Then, after naming every part of the plot, we can run the program and get the graph as figured below

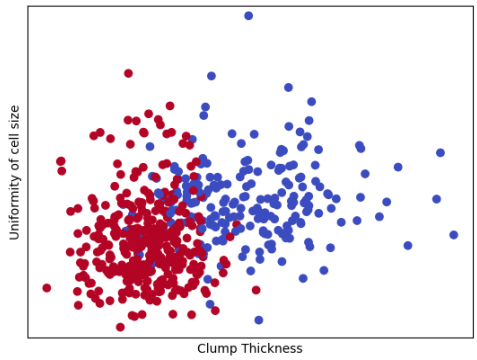

Fig. 2. Graph of breast cancer dataset

From the first sight, we would know that the red class and blue class are mostly located in each area, although the noise is exist. For the simple way, those noise is neglectable considering that it is not dominant here. So, from this graph we can find the regression using SVM with this code

```
plt.contourf(xx, yy, ex, cmap=plt.cm.coolwarm, alpha=0.5)
```

Then, the running program would generate figure below

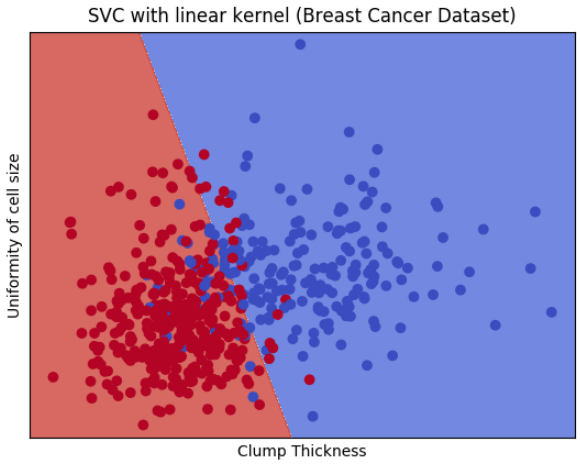

Fig. 3. Graph regression with SVM

So, for the prediction case, a data located in red area will be classified as red class, but if the data located in blue area will be classified as blue class. it shows that SVM will help us to ease data classification also.

IV. Summary

Support Vector Machine is an important concept in understanding Machine Learning. This algorithm help us to find classification area of a dataset. Without implementing the algorithm, we also can use SVM with Python. In Python, there is a package provided to do data analysis. So, we can easily use it to support our Machine Learning program.

References

[1] B. Marr. (2012, Dec. 3). *A Short History of Machine Learning – Every Manager Shoul Read* [Online]. Available: https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/

[2] G. Lamp. (2012, Dec. 3). *Why Use SVM?* [Online]. Available: http://www.yaksis.com/posts/why-use-svm.html

[3] R. Berwick, 'An Idiot's guide to Support Vector Machine (SVM)', MIT, 2003.

[4] Unknown. (Unknown). *Plot Different SVM Classifiers in the Iris dataset* [Online]. Available: http://scikit-learn.org/stable/auto_examples/svm/plot_iris.html

Acknowledgment

This paper is written for the final assignment of a course with title "Socio-Informatics and Profesionalism" in Institut Teknologi Bandung. So, first of all the author want to give thanks to The Almighty God for His blessings. Then, for the author's lecturer; Mr. Rinaldi Munir, Mrs. Ayu Purwarianti, and Mrs. Dessy Puji Lestari for their dedication to the students. Also for all people through their papers, resources, and the other that indirectly support this paper completion.

For the last part, I hereby as author acknowledge that this paper is originally written and there is no plagiarism behind it.

Bandung, 5th of May 2017

Faza Thirafi .