"Science Stays True Here"
Biological and Chemical Research, Volume 2015, 306-311 | Science Signpost Publishing

Estimating The Religion of Countries According to Shapes of The Flags Using Support Vector Machines and Kernel Methods

Taha Eren Sarnıç
Department of Statistics, Hacettepe University, Turkey

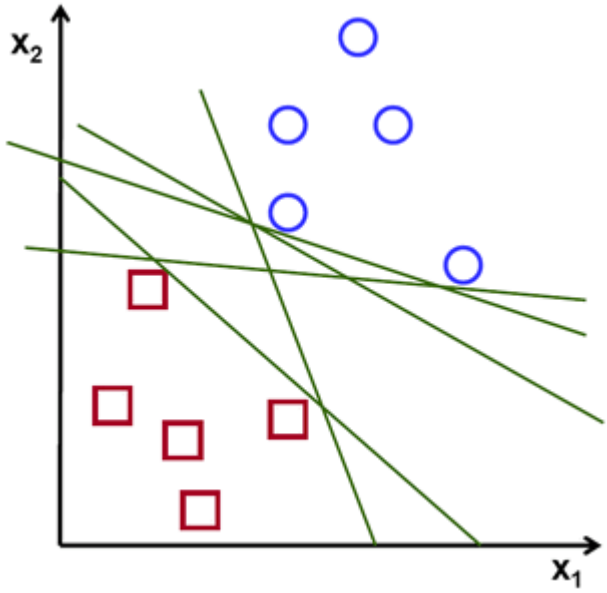
Received: July 01, 2015 / Accepted: July 30, 2015 / Published: October 25, 2015

1. Introduction

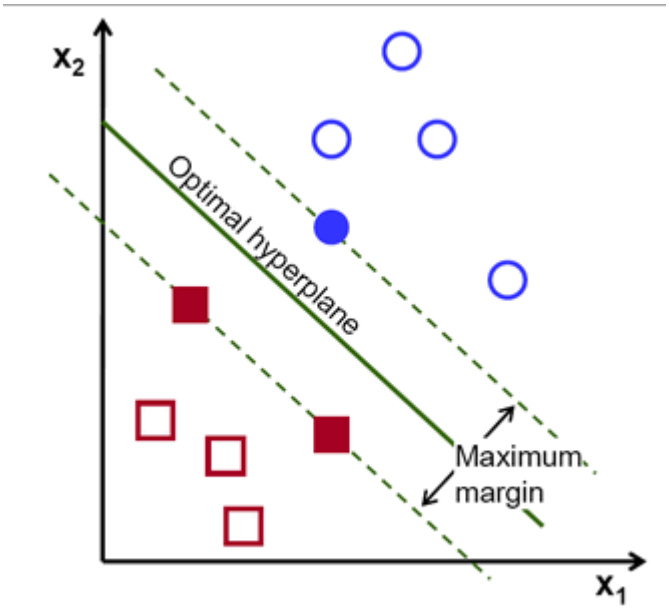
Support vector machines are a set of related supervised learning method used for classification and regression. In simple words, given a set of training examples, each marked as belonging to one of two categories, a SVM (Support Vector Machines) algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

The categorization of the new examples are made with hyperplane. As it is seen in below, two classes are seperated with multiple straight lines.

Corresponding author:
Taha Eren Sarnıç, Department of Statistics, Hacettepe University. E-mail: tahasarnic@gmail.com.



Then, the operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Therefore, the optimal separating hyperplane maximizes the margin of the training data.



2. Methods and Implementation

In the data set there are 30 attributes. In order to deal with many features, kernel methods are the best way to handle with it. Kernel methods map the data into higher dimensional spaces in the hope that in this higher-dimensional space the data could become more easily separated or better structured. There are also no constraints on the form of this mapping, which could even lead to infinite-dimensional spaces.

In the implementation we are going to use 4 kernel functions, which are Radial Based Function(RBF), Sigmoid, Linear, and Polynomial functions. Input datas are, name of the countries, landmass, zone, area, popuation, language, bars, stripes, colours, red, green, blue, gold, white, black,and orange colors, mainhue, circles, crosses, saltires, quarters, sunstars, crescent, triangle, icon, animate, text, topleft color, botright color and our output data is religion. In the table below we can see the important variables that are considered in estimation and their indexes.

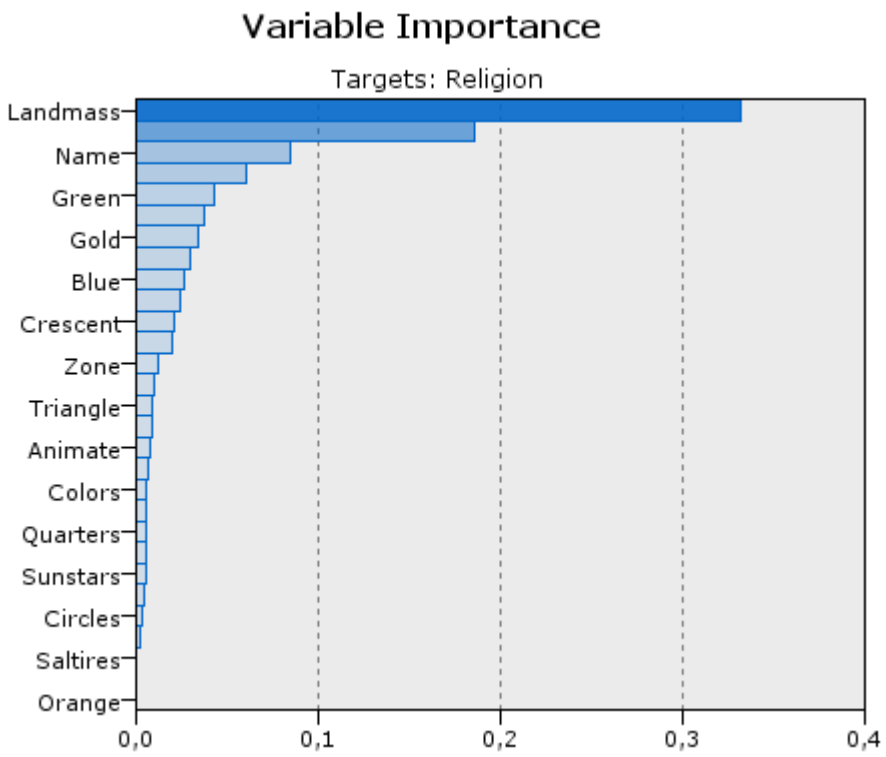


Table 1

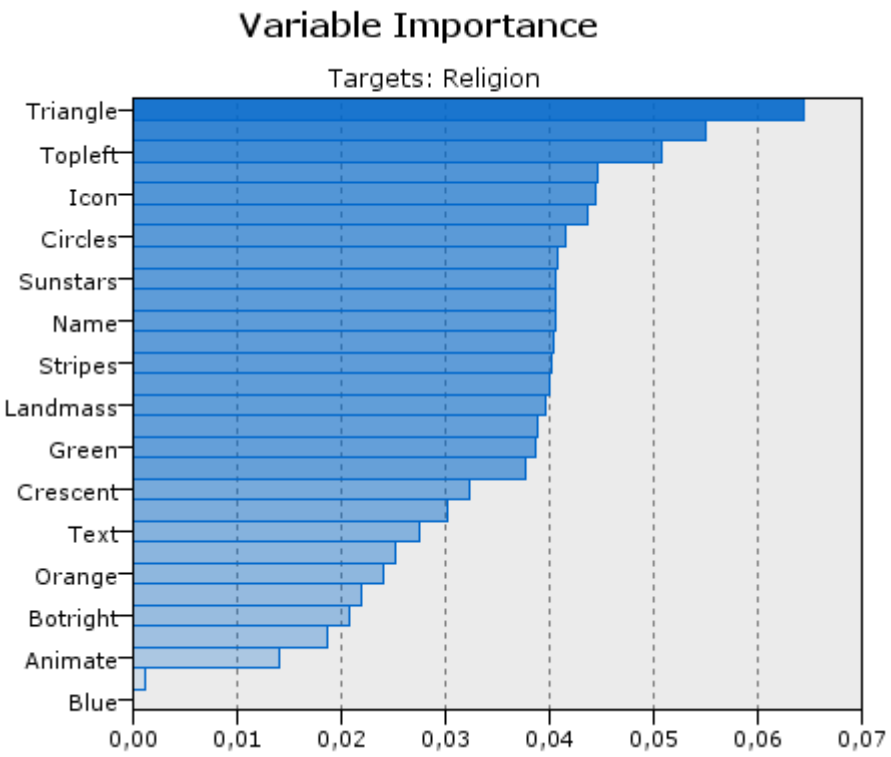


Table 2

According to Linear Function (Table 1) the most important variable is Landmass, and then the name of the country. In spite of that the Sigmoid function considers the important variables as features of the flag such as, if the flag has a triangle or not, or the topleft color of the flag etc. The indexes in Table 2 are higher than Table1.

However sigmoid function has higher indexes and it only considers features of the flag, when it comes to estimation Sigmoid function will have the lowest estimation accuracy as it is seen in Table 3 below.

Results for output field Religion

Individual Models

Comparing $S-Religion with Religion

Correct	194	100%
Wrong	0	0%
Total	194	

Comparing $S1-Religion with Religion

Correct	194	100%
Wrong	0	0%
Total	194	

Comparing $S2-Religion with Religion

Correct	59	30,41%
Wrong	135	69,59%
Total	194	

Comparing $S3-Religion with Religion

Correct	194	100%
Wrong	0	0%
Total	194	

Agreement between $S-Religion $S1-Religion $S2-Religion $S3-Religion

Agree	59	30,41%
Disagree	135	69,59%
Total	194	

Comparing Agreement with Religion

Correct	59	100%
Wrong	0	0%
Total	59	

Table 3

$S-Religion represents RBF, $S1-Religion represents Polynomial function, $S2-Religion represents Sigmoid funciton and $S3-Religion represents Linear Function. Polynomial, Linear and RBF functions have a %100 correct estimation. Where Sigmoid function has a %30.41 correct estimation which can be considered as a failure in estimating the religions. However, many kernel functions which aren't strictly positive definite also have been shown to perform very well in practice. An example is the Sigmoid kernel, which, despite its wide use, it is not positive semi-definite for certain values of its parameters. This is the reason of the bad estimation accuracy.

3. Conclusion

In this implementation we saw that Linear, RBF, and Polynomial functions have the same estimation accuracy and they have an perfect classification and prediction feature. With other types of data Sigmoid function will also do good things. Choosing the most appropriate kernel highly depends on the problem at hand - and fine tuning its parameters can easily become a tedious and cumbersome task.

Also there are other applications of SVMs such as: Bioinformatics, Machine Vision, Text Categorization, Handwritten Character Recognition, and Time Series Analysis.

References

[1]. Pang-Nıng TAN,Micheal Steinbach, Vipin Kumar, Introduction to Data Mining Ian H. Witten&Eibe Frank, Data Mining Practical Machine Learning Tools and Techniques.

[2]. George H. John and Ron Kohavi and Karl Pfleger. Irrelevant Features and the Subset Selection Problem. ICML. 1994.

[3]. Ron Kohavi and George H. John and Richard Long and David Manley and Karl Pfleger. MLC++: A Machine Learning Library in C. ICTAI. 1994.