|  |  |
| --- | --- |
| Support Vector Machine (SVM) | Jim Moraga |

Contents

[1. Introduction 2](#_Toc54594499)

[2. Data 2](#_Toc54594500)

[3. Methodology 2](#_Toc54594501)

[3.1. Data splitting 3](#_Toc54594502)

[3.2. Training and testing 3](#_Toc54594503)

[3.3. Calculating summary and stats 4](#_Toc54594504)

[4. RESULTS 4](#_Toc54594505)

[References 5](#_Toc54594506)

# Introduction

Machine learning is a subset of the artificial intelligence family of algorithms that deals with computer algorithms that improve through experience.

Among these, supervised algorithms are those that require a labeled set of samples as training and testing set to achieve automatic learning. Support Vector Machines are one Supervised Machine Learning group of algorithms, that can achieve very precise two-class classification for problems that can be solved through linear separation in a multidimensional space.

In particular, we used the e1071 package from CRAN to implement a Linear SVM to classify pixels with consolidated data from each geothermal site (Brady Hot Springs and Desert Peak), based on the labels created through our labeling algorithm.

The scripts are available in:

* https://github.com/jmoraga-mines/doe\_svm

# Data

The data is specified in the Labeling Methodology, and it consists of georeferenced layers of data:

* Geothermal
* Land Surface Temperature
* Mineral Markers
* Faults

The “Geothermal” layer is the label for each data points, and its values consist of two categories:

* Zero: Non-geothermal
* One: Geothermal

# Methodology

The classification requires a two-step process:

* Training: A subset of the data is selected randomly, and is used to train the SVM
* Testing: The remaining data is used to test the accuracy of the SVM in the dataset

Although the process is simple, care must be taken to ensure accuracy, and several cycles of training and testing were performed to achieve the accuracy estimates.

## Data splitting

The data is read by using the “raster” R package, and the data converted to a data frame, which will allow further processing and is the standard data type used by most R packages as an input.

An index is created and a random selection of the data is performed. Part of the data will be used to train the data, and the rest to test.

Monte Carlo experiment will be run by using a randomizer, and running the test several times with different sets of training and testing samples.

An important consideration is that any spatial data is removed (i.e. latitude/longitude or x/y coordinates).

## Training and testing

The cycles of training and testing are handled by the “multiple\_svm” function:

|  |
| --- |
| multiple\_svm <- function(x, y\_colname, data\_model, number = 10, p=0.1, tuneLength=3, plot\_roc=FALSE){  ### Creates "number" of iterations of svm for the dataset "x"  ### In each iteration, a portion equal to "p" is used for training  y\_index <- as.integer(which(sapply(colnames(x), FUN=function(x) x==y\_colname)))  v <- vector('list', number) # we will store the values in this list  for (i in 1:number){  training\_indexes <- createDataPartition(y = x[[y\_colname]], p = p, list = FALSE)  training\_set <- x[training\_indexes,]  testing\_set <- x[-training\_indexes,]  print(paste("Calculating SVM Linear Model #", i))  svm\_model <- svm(data\_model,  data=training\_set, probability=TRUE,  cost=1, gamma=1)    print(paste('Result #', i))  print(svm\_model)  print("########################################################")  print ("Now, to predict...")  svm\_predict <- predict(svm\_model, testing\_set, probability=TRUE, decision.values=TRUE)  if (plot\_roc){ # Do we want to plot the ROC curve?  svm\_prob\_geo <- attr(svm\_predict, "probabilities")[,"1"]  svm\_pred <- prediction(svm\_prob\_geo, testing\_set$Geothermal=="1")  svm\_perf <- performance(svm\_pred, "tpr", "fpr")  plot(svm\_perf, col="red",  main=paste("ROC curve for SVM, result #", i),  width = 1000, height = 600)  }  # calculate confusion matrix  svm\_table <- table(pred = svm\_predict,  true = testing\_set[,y\_colname])  svm\_cmatrix <- confusionMatrix(table(svm\_predict,  testing\_set[[y\_colname]]))  print(svm\_table)  print(svm\_cmatrix$overall)  print("########################################################")  v[[i]] <- list(svm\_model, svm\_predict, svm\_table, svm\_cmatrix, svm\_pred)    }  return (v)  } |

In the function call, the key parameters are:

* x: the data frame that contains all the layers
* y\_colname: the name of the column that contains the labels (this should be “Geothermal”)
* number: the number of times the SVM algorithm will be run with different randomized test sets
* p: the percentage of the data that should be used to train (this should be 5% or 10%)
* TuneLength: number of levels generated for each SVM algorithm parameter
* plot\_roc: Boolean, selects whether to plot the ROC curve or not

## Calculating summary and stats

The results are saved, and then a summarization script is run to show the results of each iteration.

The script is: summarize\_svm.R

# RESULTS

The results are very good for Brady and Desert Peak, but poor for Salton Sea:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Site** | **% data** | **Accuracy** | **Kappa** | **Accuracy Evaluation** |
| Brady | 5% | 95.58% | 0.9083 | Very High |
| Brady | 10% | 95.59% | 0.9086 | Very High |
| Desert Peak | 5% | 91.93% | 0.8383 | Very High |
| Desert Peak | 10% | 91.93% | 0.8382 | Very High |
| Salton Sea | 5% | 52.71% | 0.2474 | Poor |
| Salton Sea | 10% | 53.08% | 0.2556 | Poor |

The main causes of the underwhelming results are location-specific issues and data problems:

* Location issues
  + The area of analysis ended up being relatively small and very heterogenous
  + The coverage area was not fit for proper analysis due to land cover (i.e. water and agricultural areas skew temperature and deformation measurements)
* Data Problems
  + Learning models work better with large sets of data, but the area analyzed was small due to the small coverage of the HyMap/MAKO images
  + Learning models require a good mix of samples of each class. In this case, the whole area of analysis is geothermal
  + There was no usable Fault data for the area. There are studies that propose a fault that covers the whole area of analysis, and also the whole area is seismically active, once again preventing any type of discrimination
  + Given the land cover, both temperature and deformation analyses turn out poor outcomes (e.g. spotty areas, known geothermal zones marked as low temperature, distorted InSAR results)

# References

**There are no sources in the current document.**