Rayanna Harduarsingh

October 28th, 2020

IST 687

Module 10 Notes

Chapter 18: What Your Vector, Victor?/ Module 10: SVM (Support Vector Machine)

- Supervised Learning Data Mining
    - Train algorithm on an initial set of data
    - Test algorithm on a new set of data
    - Validate trained algorithm predicted the right outcome
    - Ex: Email is spam/not spam, Weather is cloudy/not cloudy
- Predicting the Weather Example
    - Collect weather data over period of time
        - Sunny, cloudy
        - Temp
        - Barometer
        - Wind Speed and direction
    - We would train this algorithm with the above variables
    - Collect more weather data and predict the weather via our trainees algorithm and see if it is valid
- Supervised Learning Strategy
    - Substrainial number of truing cases that the algorithm can use to discover and mimic the underlying pattern
    - Use the results of this process to determine how well the algorithm performed
    - Cross validate the process of verifying the algorithm can carry its prediction
- Kernel: Mapping Algorithm
    - Must install in R (“kernlab”)
    - Input Data
        - Independent variables from a given case

    - Kernel:
        - Formula run on each case's input data
    - Output Data:
        - Position of that case in a multidimensional space
- What are some examples where something like SVM could be used?
    - An example would be of course in shopping. Stores can predict who are their most frequent and popular customers and who is not. They can also predict which of their items will sell fast or not, perhaps during a certain season like Christmas. Valentines, or Halloween. It can also be used to predict the health of a human or predict signs of cancer.

SVM in R:

- Create mini table from data set form certain column:

  > table(spam$type)
- Calculating the cut point that would divide the spam data set into a two-thirds training set and a one-third test set:

  > "cutPoint2_3 <- floor(2 * dim(spam)[1]/3)"
    - the expression dim(spam)[1] gives the number of rows in the spam data set
    - the floor() function chops off any decimal part of the calculation
- Build a training set:

  > trainData <- spam[randIndex[1:cutPoint2_3],]
- Generating a model based on training data set:

```
> svmOutput <- ksvm(type ~., data=trainData,  kernel =
+      "rbfdot",kpar="automatic",C=5,cross=3,
+       prob.model=TRUE)

Using automatic sigma estimation (sigest) for RBF or
laplace kernel
```

•The kpar argument refers to a variety of parameters that can be used to control the operation of the radial basis function kernel.

- The C argument refers to the so-called cost of constraints.
- prob.model=TRUE, dictates that we use a so-called threefold cross-validation in order to generate the probabilities associated with whether a message is or isn’t a spam message.