CSCI 181 / E-181 Spring 2014 Practical 3

Kaggle Team ”Capt. Jinglehiemer”

David Wihl
davidwihl@gmail.com

Zack Hendlin
zgh@mit.edu

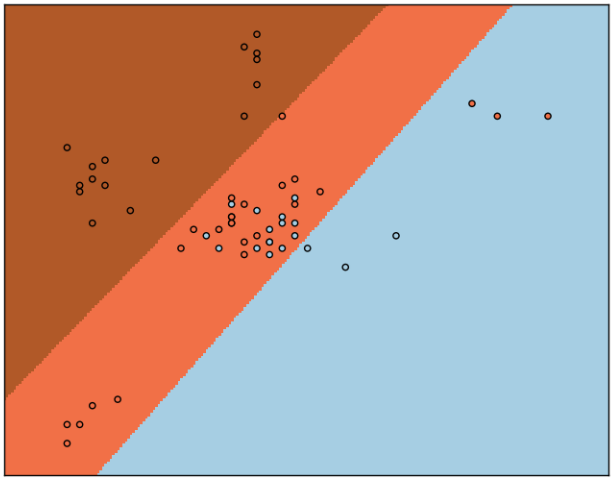
March 12, 2014

Warm-Up

We consider two approaches for classifying fruits (with length and width measurements provided) into one of three categories.

[Plot]

Logistic Regression

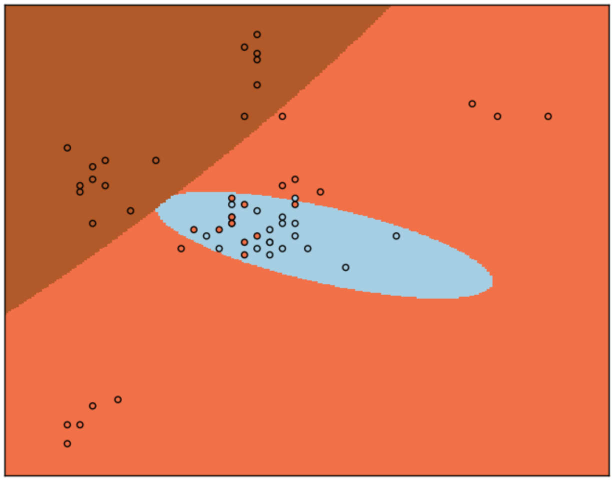


Generative Models

The multivariate normal is given by:

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\right.$$

where k=2 in our case because we have (1) height and (2) width.



Warmup Summary

Classifying Malicious Software

Preliminary Data Analysis

NOTES: 4GB of XML to parse and process, first step was to split the training and the testing. broke into vectorize, train and test steps, persisting appropriate intermediate data at each step. This also enabled parallelization of test runs over a cluster of machines.

Using Cross-validation

Ran 5 CV sets of train / CV data 70/30, 80/20 and 90/10 for each classifier.

Approaches considered

Feature Engineering

Aggregate Features per training file: selected all process features (e.g. 'startreason', 'terminationreason', 'username', 'executionstatus', 'applicationtype') and summary thread features (num of each type of system call).

used CV to generate Logistic Regression weights. Took mean and std of resulting matrix, then eliminated any features where abs(mean) < 0.001 and std <0.01.

Selection of fitting technique

Tried LogisticRegression and SVM with a number of different C values, none of which made a significant difference.

Exploratory Data Analysis

Conclusion