Ryan Timbrook

Data Science 450, Spring 2017

Date: 05/27/2017

Assignment 5

Shared Azure Notebook URL:

https://notebooks.azure.com/n/uoSuVfd6HXI/notebooks/ds_450_ass5_v1.0.ipynb

Description: Support Vector Machine

Question 1

Using what you learnt in the lecture, search for additional resources related to the following two approaches.

- Two-class Support Vector Machine
- Two-class Locally-Deep Support Vector Machine (MSR)

Explain what's the difference between the two algorithms, and provide examples of when you will prefer to use one over the other.

Algorithm characteristic comparisons:
(*X denotes the algorithm has the characteristic specified*)

	TC-SVM	TC-LD-SVM
Creates a binary classification model	X	X
Trained on Positive and Negative examples	X	X
Useful for predicting between two possible outcomes	X	X
Supervised Learning model	X	X
Creates a two-class LINEAR SVM classifier	X	
Creates a two-class NON-LINEAR SVM classifier		X
Optimized for efficient prediction		X
Optimized training time (speed)		X
Used in information retrieval	X	X
Used in text classification	X	X
Used in image classification	X	X

When to use one over the other:

	TC-SVM	TC-LD-SVM
Linear classifiers are not performing well		X
Non-linear classifiers work well, but are slow to predict		X
Prediction speed off sets prediction accuracy		X
Linear classifier are yielding good results	X	

Question 2

For this exercise, we will use the veh-prime.arff file, and support vector machines for classification. Try various parameters, and explain what you observe.

Observations:

Modifying the cost parameter of the svm function effects the number of Support Vectors selected during the modeling process. While tuning the svm model the best parameter cost was 100. Comparing Linear, RBF and Polynomial SVM Kernels using the best parameter cost value of 100 yields:

- Linear
  - Number of support Vectors: 177
  - Training error: 0.045608
- RBF
  - Number of support Vectors: 325
  - Training error: 0
- Polynomial
  - 117
  - 0.045608

Exploring the principal components of this data set reflects PC1 as being the most important component. This is shown in Table 1 below.

Evaluation of the SVM Model classifier (ANOVA RBF) yielded an AUC score of 97% when using the default cost parameter. After modifying the model with the best parameter cost value of 100 the AUC score ran on the test data set yielded an AUC score of 98%. This is shown in Table 2 and 3 below.

Table 1: Principal Component Analysis

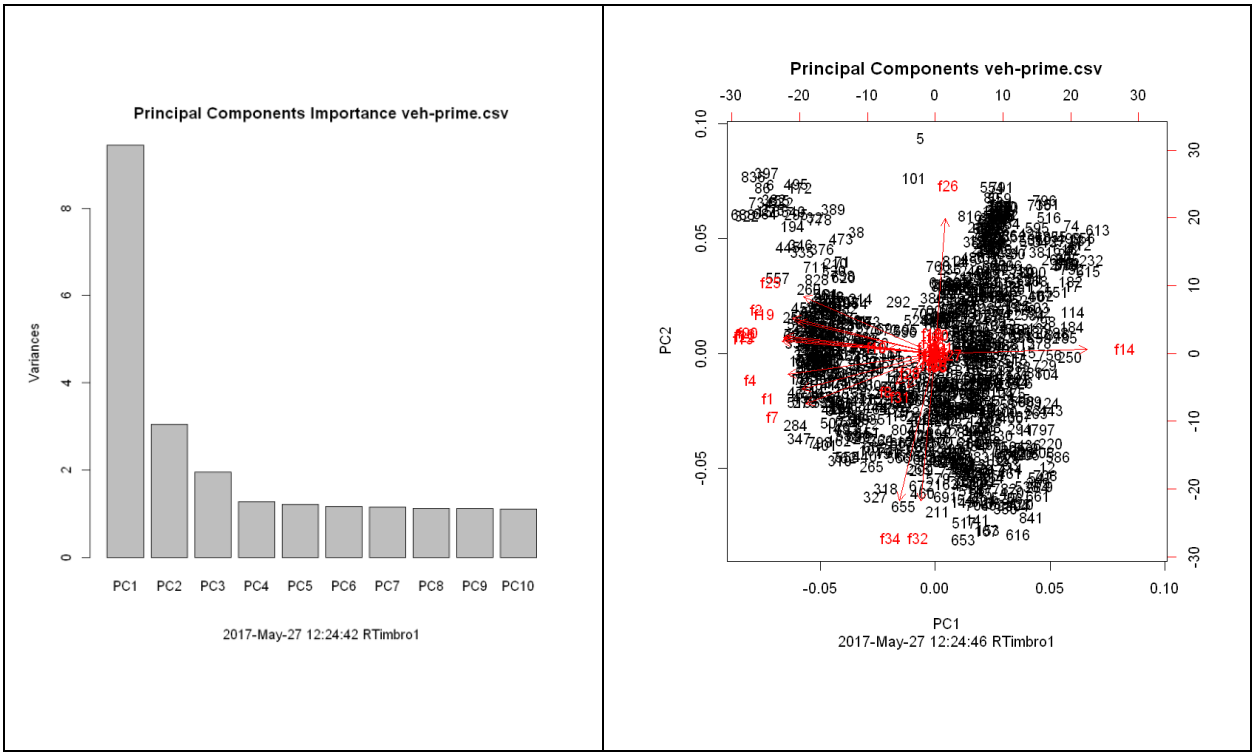

Table 2: Evaluation of SVM Model (ANOVA RBF – Default Cost)

Number of Support Vectors: 151

	Predicted		
Actual	-1	0.972	Error
-1	55	5	.08
.972	8	60	.12

	Predicted		
Actual	-1	0.972	Error
-1	.43	.04	.08
.972	.06	.47	.12

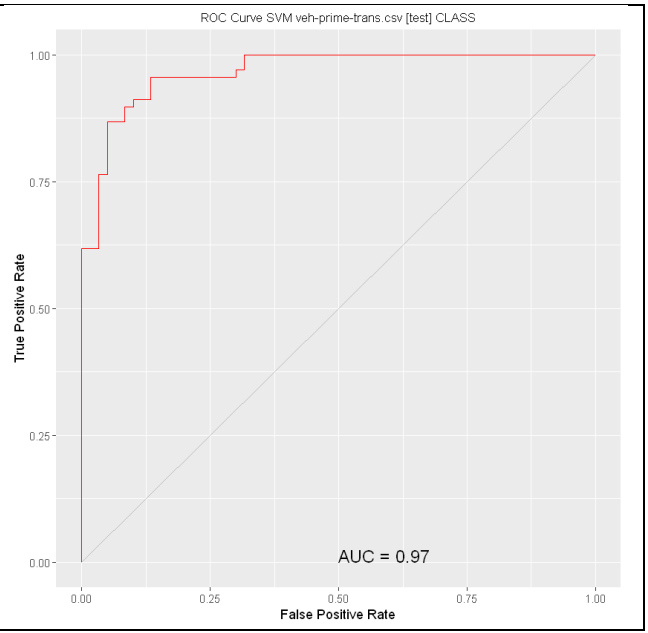

Table 3: Evaluation of SVM Model (ANOVA RBF – Best Fit, Cost 100)

Number of Support Vectors: 94

	Predicted		
Actual	-1	0.972	Error
-1	56	4	.07
.972	2	66	.03

	Predicted		
Actual	-1	0.972	Error
-1	.44	.03	.07
.972	.02	.52	.03