title: "INFO523\_Project\_SVM"

author: "Rice"

date: "2023-08-12"

output: html\_document

**Abstract**

The project is an application of a support vector machine algorithm (SVM) to classify letters utilizing the letter recognition dataset located at url <- <https://archive.ics.uci.edu/ml/machine-learning-databases/letter-recognition/letter-recognition.data>. The project introduces the concepts of support vector machines and leverages the Cats dataset in the Mass package to explain linearly separable data, hyperplanes, margin, kernels, and non-linearly separable cases.

The Cats dataset comprises one hundred and forty-four instances of data. The response Yi = sex with values male vs. female. The dataset includes two variables X1 cat weight in kilograms and X2 heart weigh in grams. The advantage to the Cats dataset is that the two variables can be plotted, and the effect of model parameters displayed visually. The SVM approach to separate nonlinear data using kernels is explained. The Cats dataset is re-analyzed using both a radial and a polynomial kernel. For each kernel, linear, radial, and polynomial, the SVM model is evaluated using a test dataset comprising forty-four instances and the fraction correctly classified is reported.

The letter recognition dataset displays the twenty-six capital letters of the English alphabet. The character images were converted into sixteen primitive numerical attributes which were then scaled to fit into a range of integer values from 0 to 15. The letter recognition dataset comprised 20,000 instances. The dataset was separated into training and test datasets with 16,000 and 4,000 instances respectively.

Initially, to access the letters training dataset an SVM model was created using the linear “vanilladot” kernel in the kernlab package. That initial analysis of the dataset resulted in correctly classifying 84% of the instances in the test dataset. To test the effect of using a non-linear kernel the training data was reassessed using the radial “rbfdot” kernel. The radial kernel also resulted in a correct classification of 84% of the test dataset.

To improve model accuracy various values for the cost function were trialed using an iterative loop. Cost values “C” evaluated ranged from 4 to 20 in steps of 2. The results were plotted with cost values on the x-axis and % accuracy predicting the class of the test data on the y-axis. That analysis indicated model accuracy could be maximized with values of the cost function equal to twelve or greater (C>12). With the cost value (C=12) the SVM model correctly predicted 97% of the instances in the test dataset.

The analysis of the letter dataset using SVM indicates that the algorithm is very effecting as a classification tool. The algorithm can assess multiple classes provided the correct kernel is selected and trail and error used to define key parameters. When analyzing the letter recognition dataset, the SVM algorithm achieved 97% accuracy when classifying the 4,000 instances within the test dataset.