Joel Kerfoot
V00855134

Assignment 1 – SENG 474

Logistic Regression

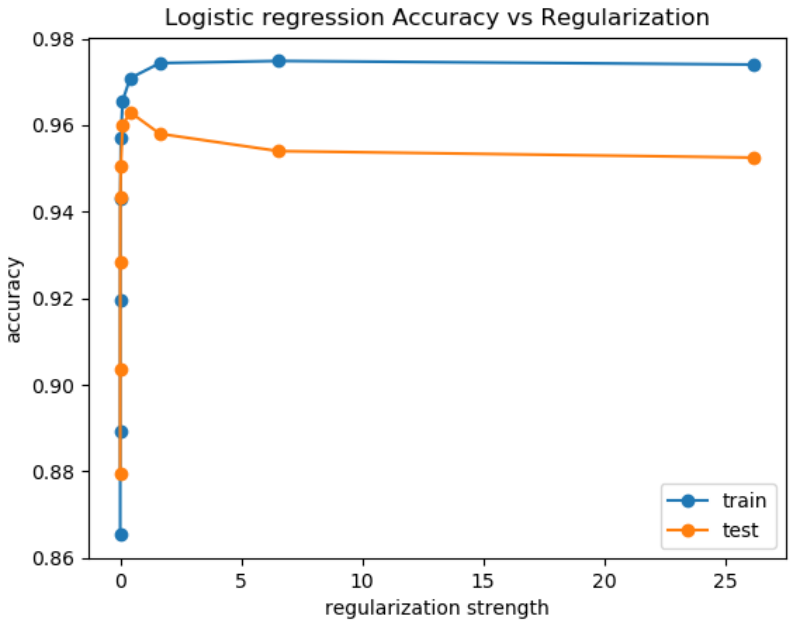
The task for logistic regression was to find a regularization strength *C* the has the best test accuracy/least test error. The data set consisted of 6000 28x28 greyscale images. The images are classified as 0 – 'sandal' and 1 – 'sneaker'. This was done using the *l2* penalty where training is penalized according to $||w^2||$. To do this the following formula was used:

$$C_0 * \alpha^i$$

Where C_0 = 0.0001; α = 4; and $0 \leq i < 10$

This produced values of *C* ranging from 0.0001 to 26.2. *C*=0.4096 (*I* = 6) was the most accurate with 96.3% test accuracy or 3.7% test error (97.5% training accuracy / 2.5% training error). See figure 1 below for a comparison of *C* and the respective accuracy.

FIGURE 1 – LOGISTIC REGRESSION TEST ACCURACY

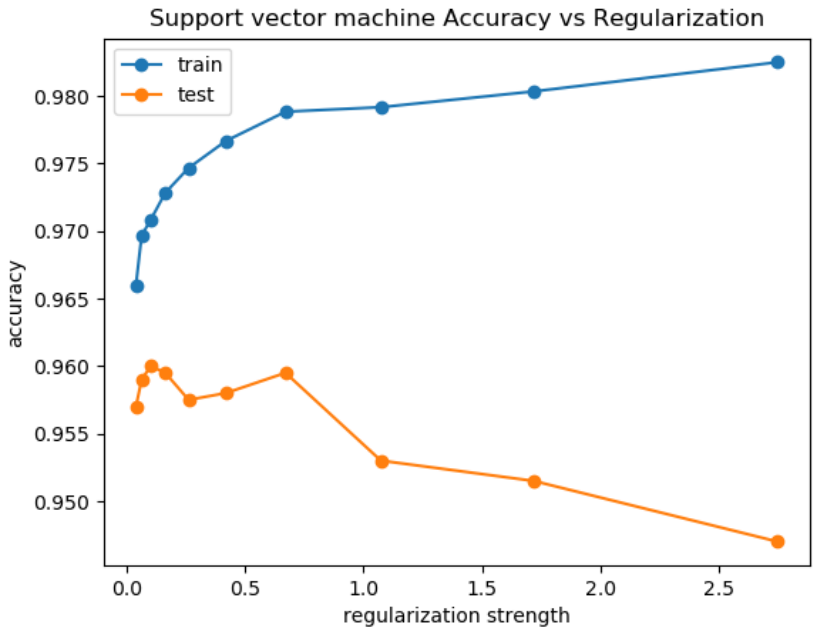


Joel Kerfoot
V00855134

Support Vector Machines

The task for training the support vector machine (SVM) was the same as logistic regression, determining a value of *C* that has the least amount of error on the test data. In order to improve training time only half of the 6000 data points were used. Reducing the training size will have an impact on the total accuracy of the algorithm but significantly improved the time required to train. This was done on an SVM with a linear kernel. The same formula was used to determine the best value of *C.* In this case $C_0 = 0.004$; $\alpha = 1.6$; and $0 \leq i < 10$. The least error obtained on the test set was 4.0% or 96.0% accuracy for both $C = 0.164$ and 0.671. Figure 2 below illustrates the training and test error for all values of *C* that were tested.

FIGURE 2 – LOGISTIC REGRESSION TEST ACCURACY



Joel Kerfoot
V00855134

K-Fold Cross-Validation – Logistic Regression

Using the value C=0. 4096 obtained earlier k-fold cross validation was done to further tune the parameter to import the test accuracy. k=5 was chosen because it results in 80% training 20% test split and has a shorter training time than picking a larger value of k. K-fold cross-validation was done on the training data set. Using the same formula as before C was chosen using the initial values $C_0 = 0.7$, and $\alpha = 1.1$. The best classifier was chosen and trained on the entire training data set. This classifier was 97.4% accurate (2.6% error) on the training data and 95.9% (4.1% error) accurate on the test data set which it had never seen before.

K-Fold Cross-Validation – Support Vector Machine

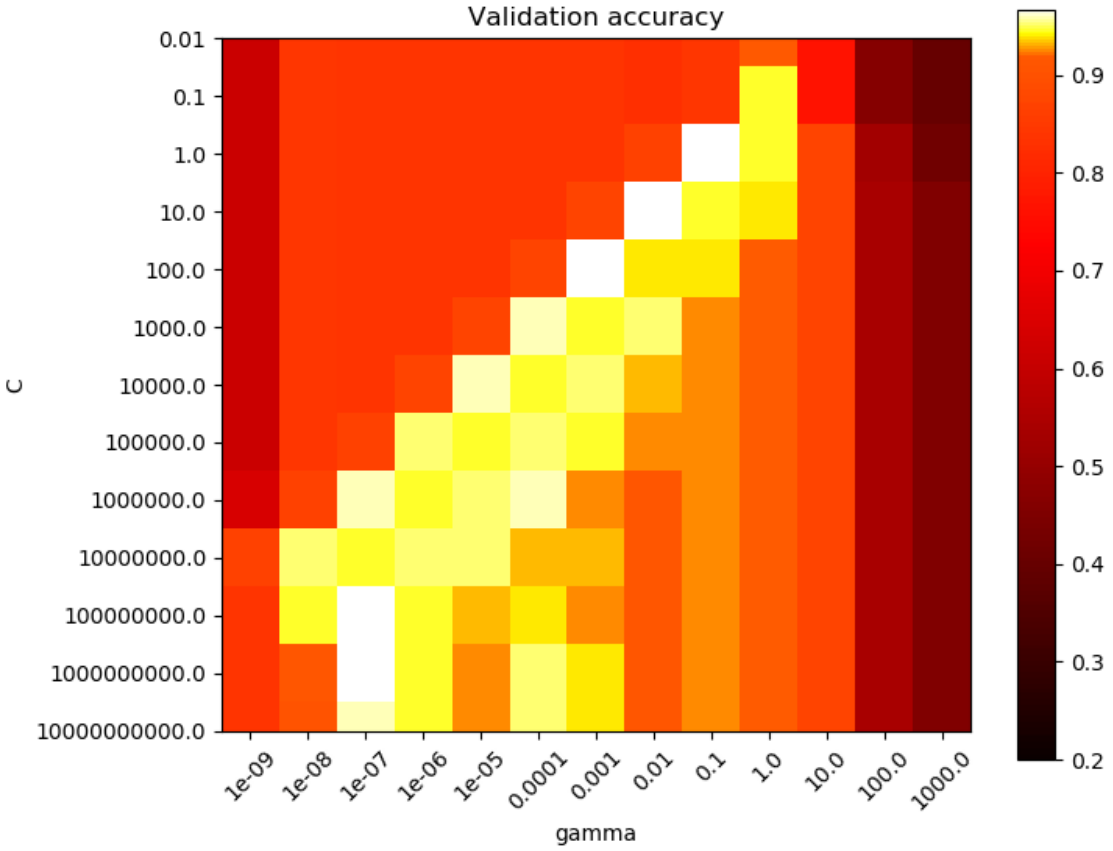
The same procedure as K-fold logistic regression was done for support vector machines but again only 3000 of the 6000 data points were used to reduce training time. The initial values used for support vector machines was much closer to the original values used because there were values of C that achieved the best accuracy. The values used were $C_0 = 0.04$, and $\alpha = 1.5$. The best performing classifier achieved 96.8% training accuracy (3.2% error) and 95.8% test accuracy (4.2% error).

Joel Kerfoot
V00855134

Gaussian Kernel

The final task was to use support vector machines with a non-linear kernel, in this case the gaussian or radial basis function (RBF) kernel was used. This kernel takes 2 parameters *C* and γ. The values shown in figure 3 was used to determine reasonable values to try for *C* and γ (gamma).

FIGURE 3 – C VS GAMMA ACCURACY



src: https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

To determine values for gamma the following formula was used:

$$\frac{1}{len(x_{train})/\ i}$$

Where i are logarithmically spaced values ranging from 15.8 to 158.5. This produces values of gamma ranging from 0.0053 to 0.053.

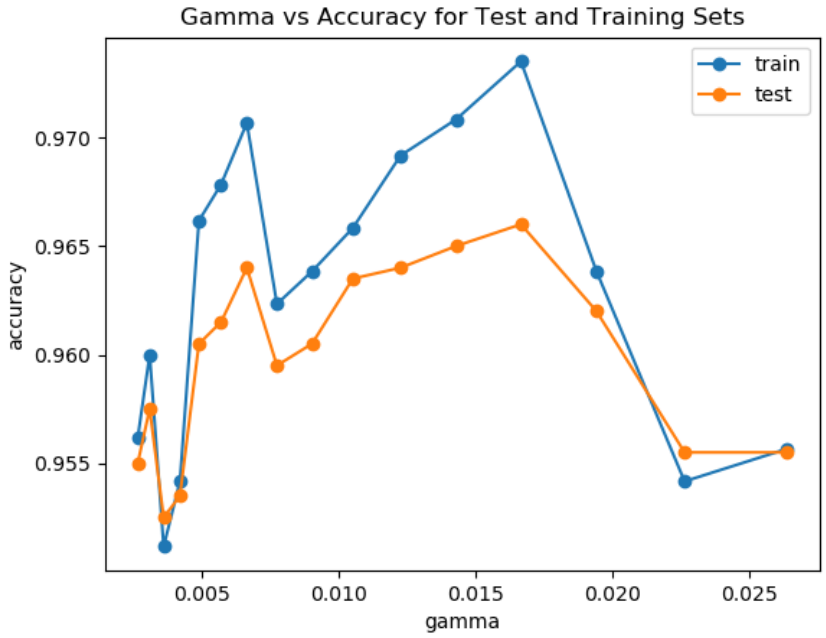
Joel Kerfoot
V00855134

To determine values of *C* the following formula was used:

$$c_0 * a^{2j}$$

Where c_0 = 0.04 and a = 1.5 and j ranged from 5 to 14 inclusively. This produced values of *C* ranging from 2.3 to 7670. The ranges of the values fall nicely into the expected higher accuracy percentile shown in figure 3. The best combinations of *C* and gamma were then trained on the full dataset and the test and train error was recorded. The results of the training and test accuracy is show in figure 4 below.

FIGURE 4 – GAMMA VS ACCURACY

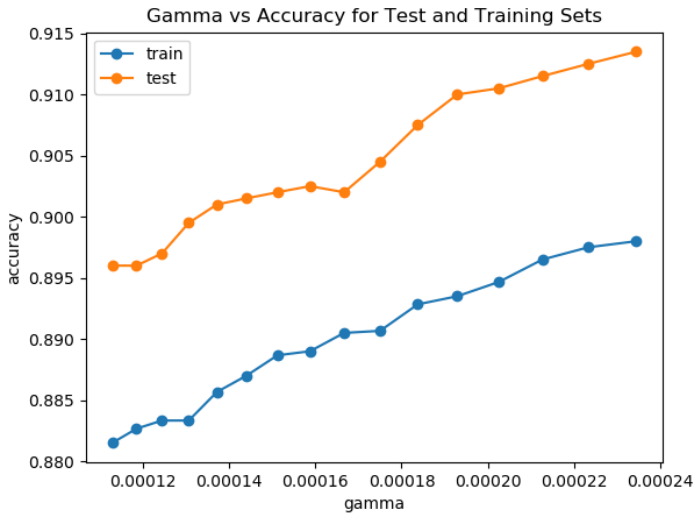


The maximum accuracy of ~96.6% (3.4% error) was achieved with a gamma value of 0.0166. Appendix A shows some less than successful attempts to determine reasonable values for *C* and gamma.

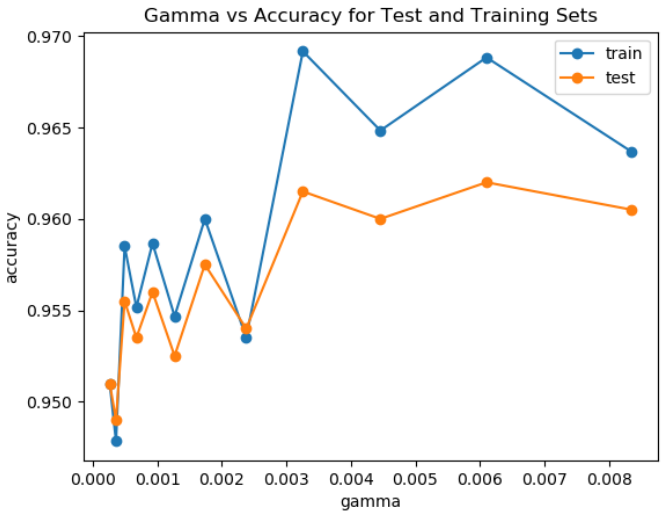
Joel Kerfoot
V00855134

Appendix A

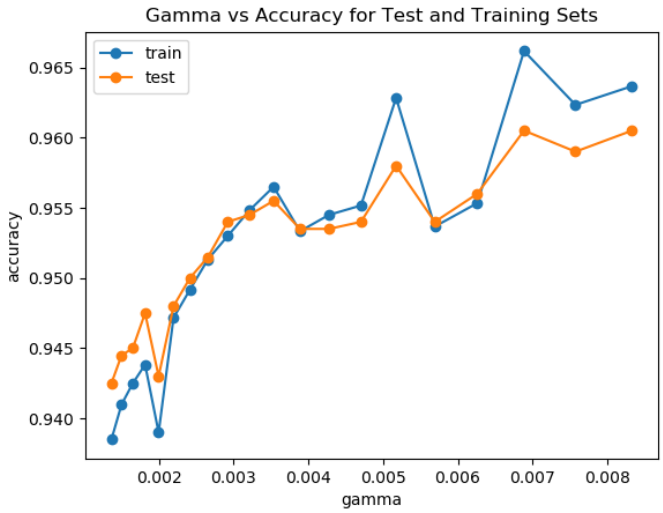
Attempt 1



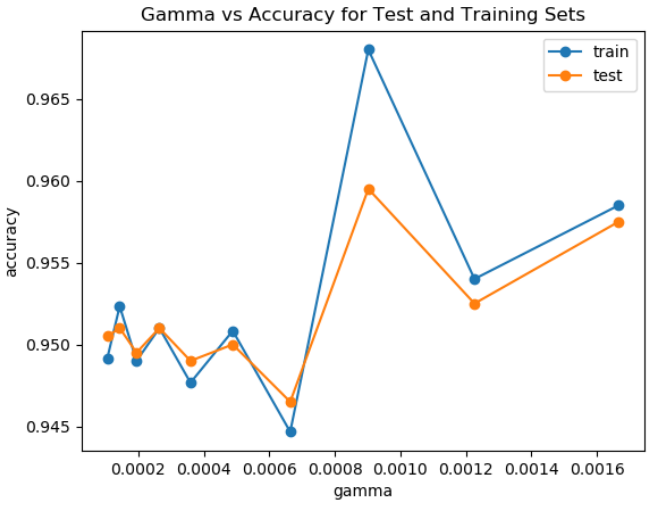
Attempt 2



Attempt 3

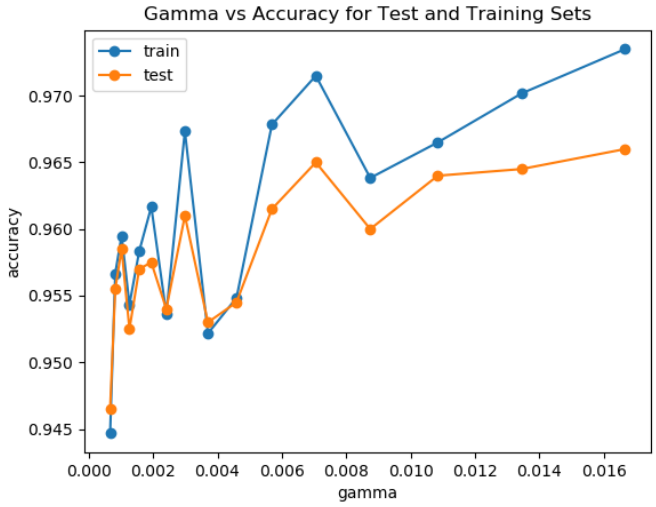


Attempt 4



Joel Kerfoot
V00855134

Attempt 5



Attempt 6