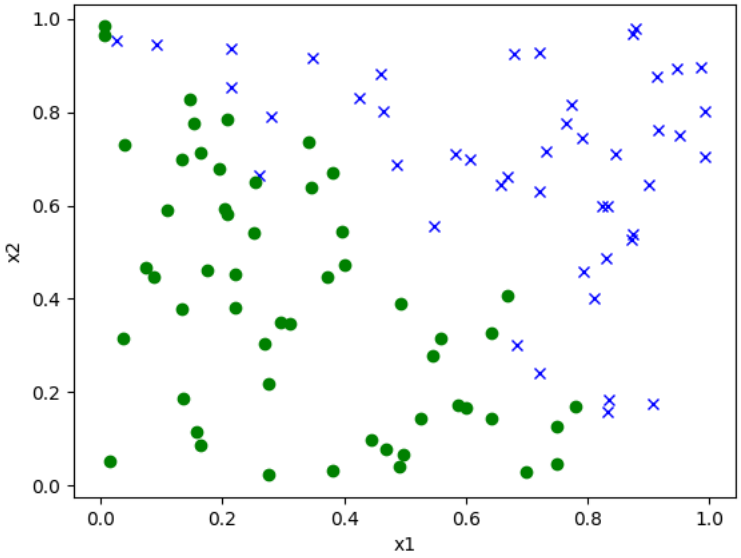
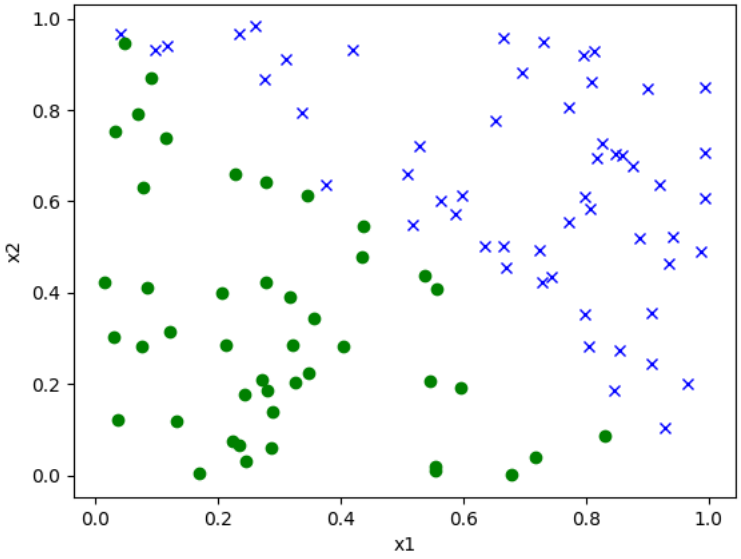
Stanford CS229 Machine Learning Problem Set #2 Coding Problems

1. Logistic Regression: Training Stability

1-b. Plot dataset a and b on two diagrams.
Dataset a: Cannot linearly separate positive and negative samples.



Dataset b: Positive and negative samples can be linearly separated.



2. Spam Classification

2-a. Report the vocabulary size after the pre-processing step.
Answer: Size of dictionary: 1722

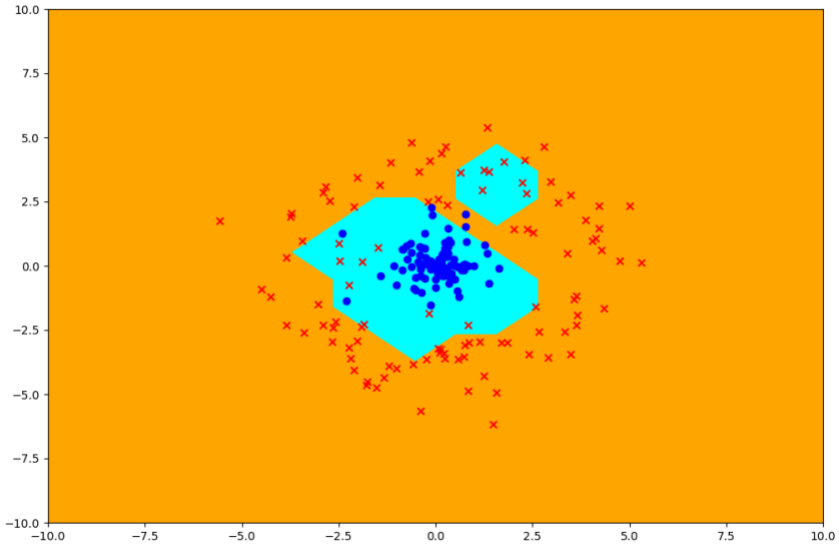
2-b. Report the accuracy of the trained model on the **test set**.
Answer: Naive Bayes had an accuracy of 0.978494623655914 on the testing set.

2-c. Report the top five words.
Answer: The top 5 indicative words for Naive Bayes are: ['claim', 'won', 'prize', 'tone', 'urgent!'].

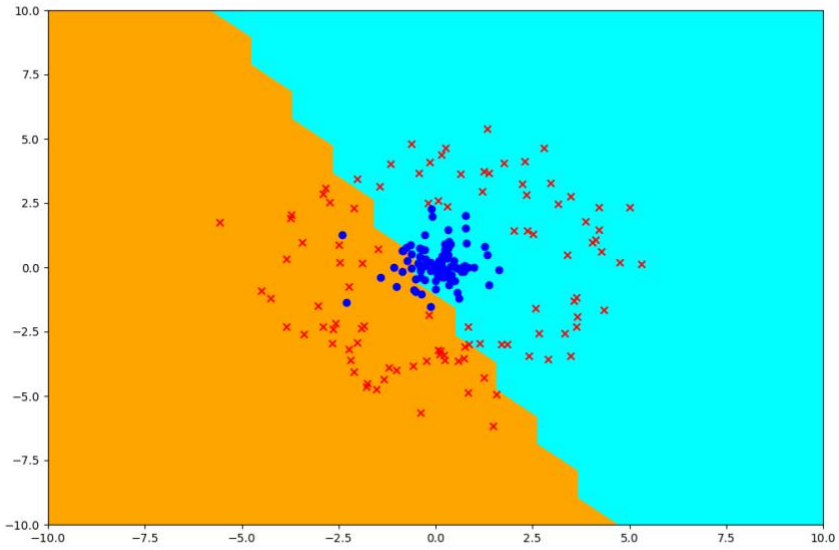
2-d. Report the best kernel radius.
Answer: The optimal SVM radius was 0.1. The SVM model had an accuracy of 0.9695340501792115 on the testing set.

4. Kernelizing the Perceptron

4-b. Include the two plots and indicate which plot belongs to which kernel
Perceptron output using a dot-product kernel:



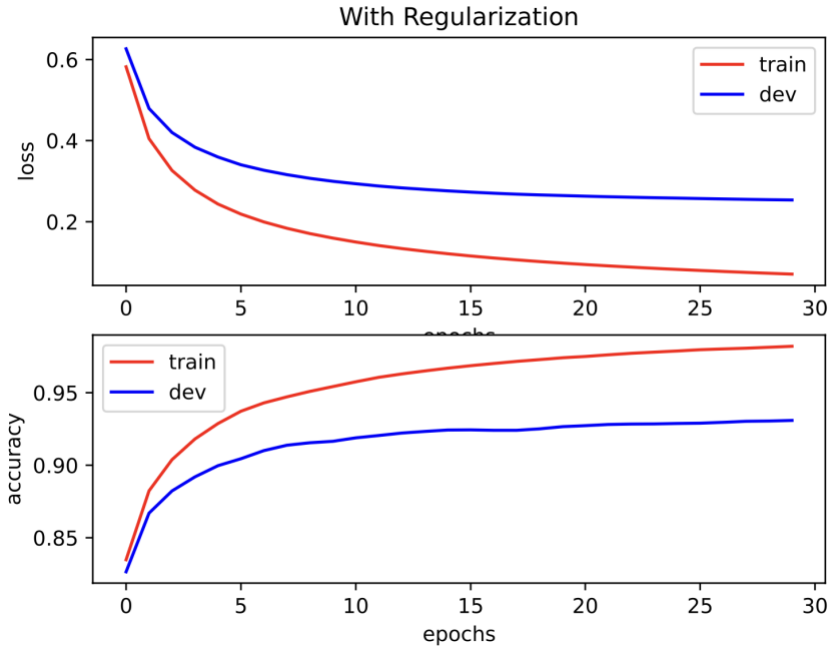
Perceptron output using a radial basis function (RBF) kernel:



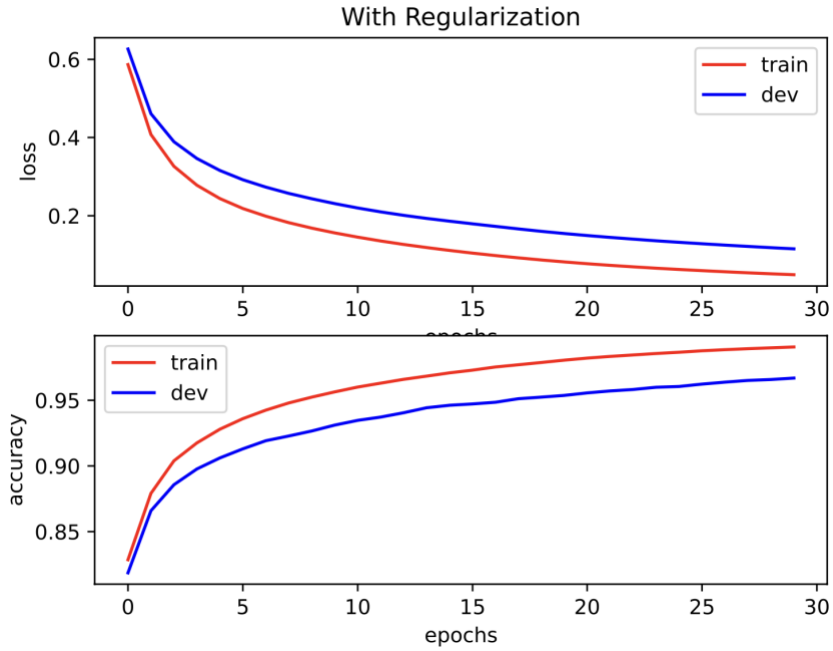
4-c. Which kernel performs badly and why does it fail?
Answer: The dot-product kernel performs badly. The reason is that the data is not linearly separated.

5. Neural Networks: MNIST image classification

5-a. Non-regularization Batch Gradient Descent (Loss vs epoch, accuracy vs epoch)



5-b. Batch Gradient Descent with Regularization (Loss vs epoch, accuracy vs epoch)



5-c. Report test accuracy for both regularized model and non-regularized model.

Answer: For model baseline (non-regularization), got accuracy: 0.931400. For model regularized, got accuracy: 0.964900.