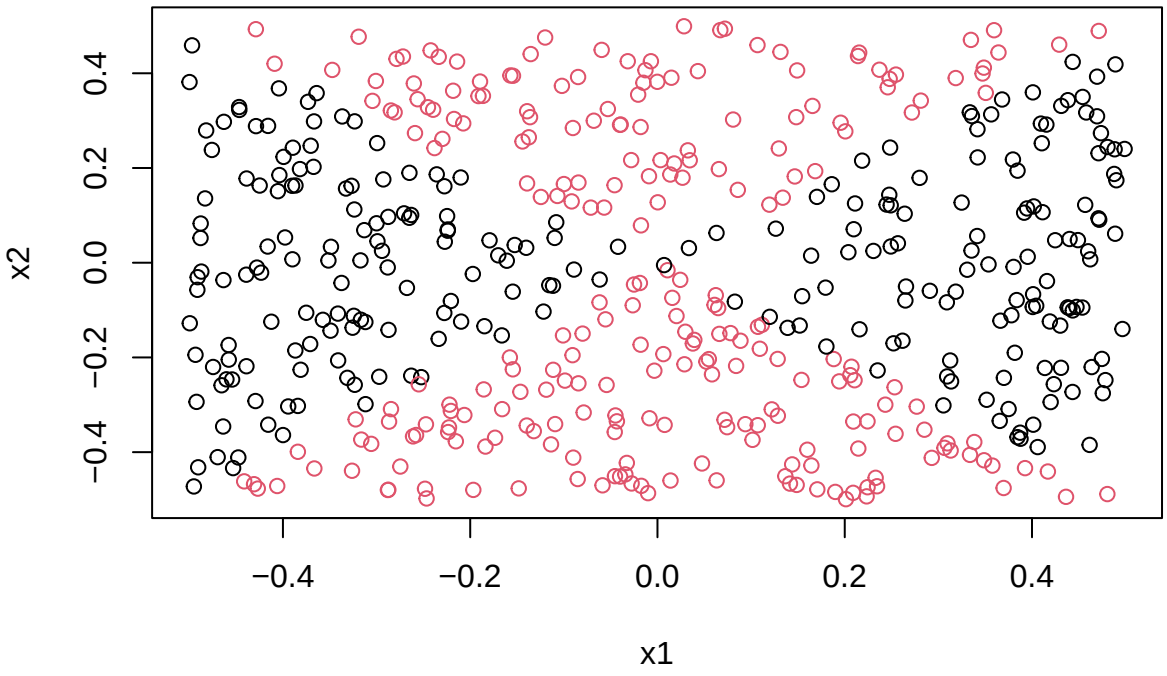
Homework Chapter 9

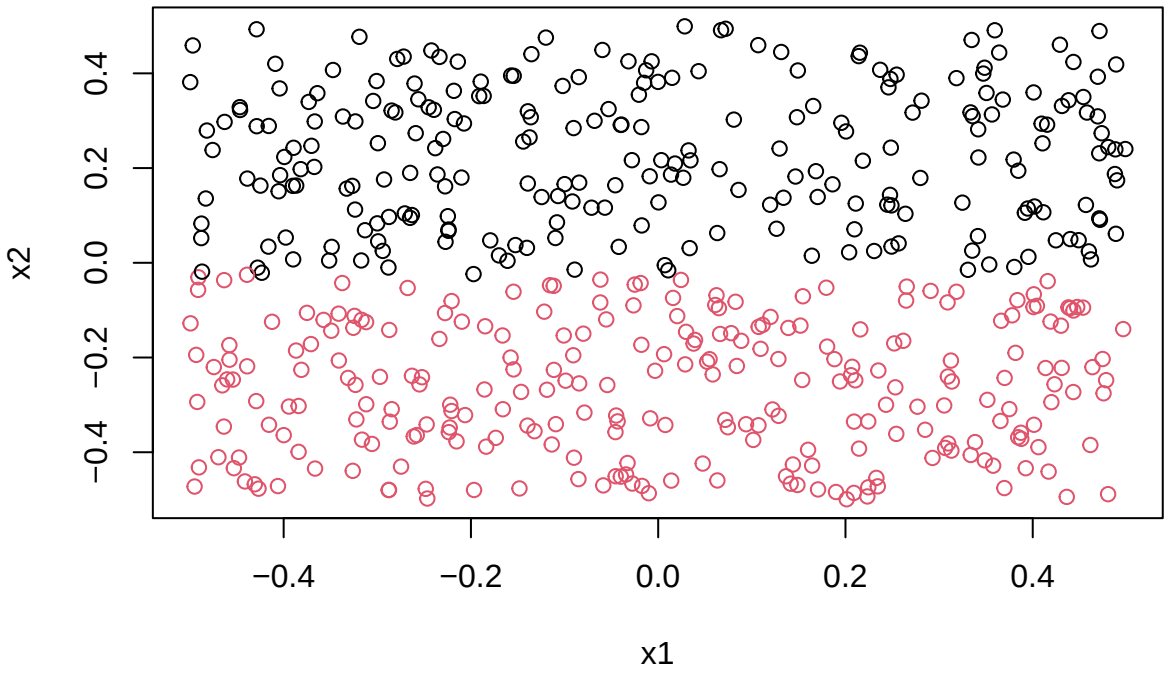
Jacob Thielemier

17 April 2024

Question 1

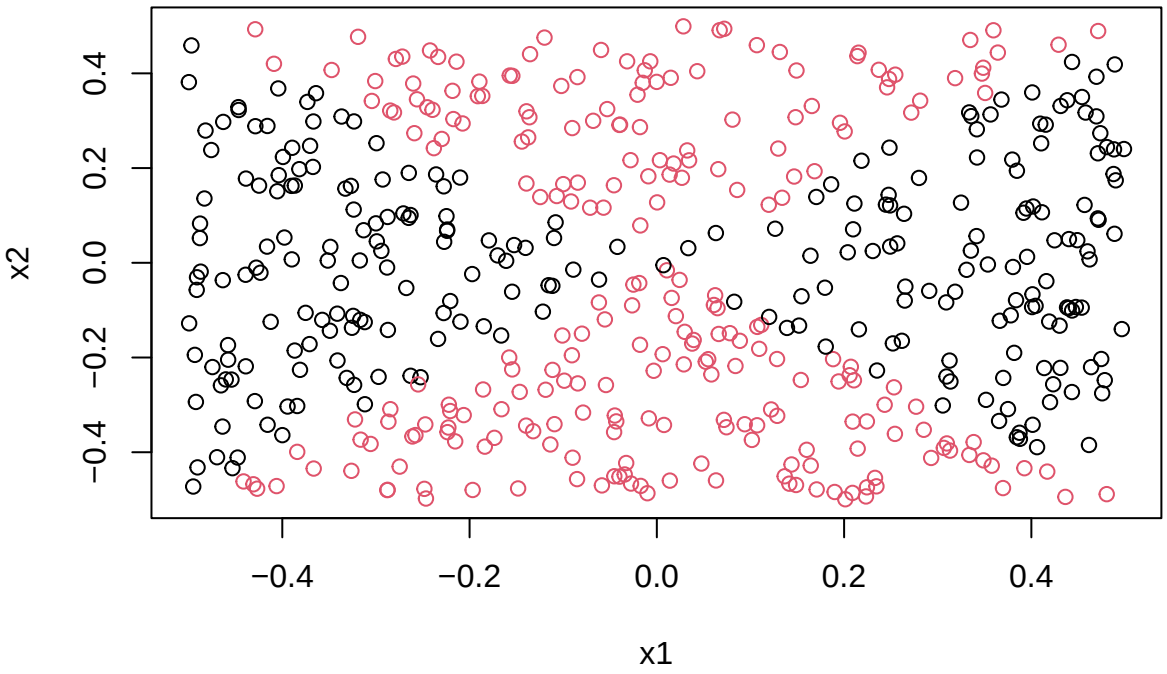


Part (a and b)



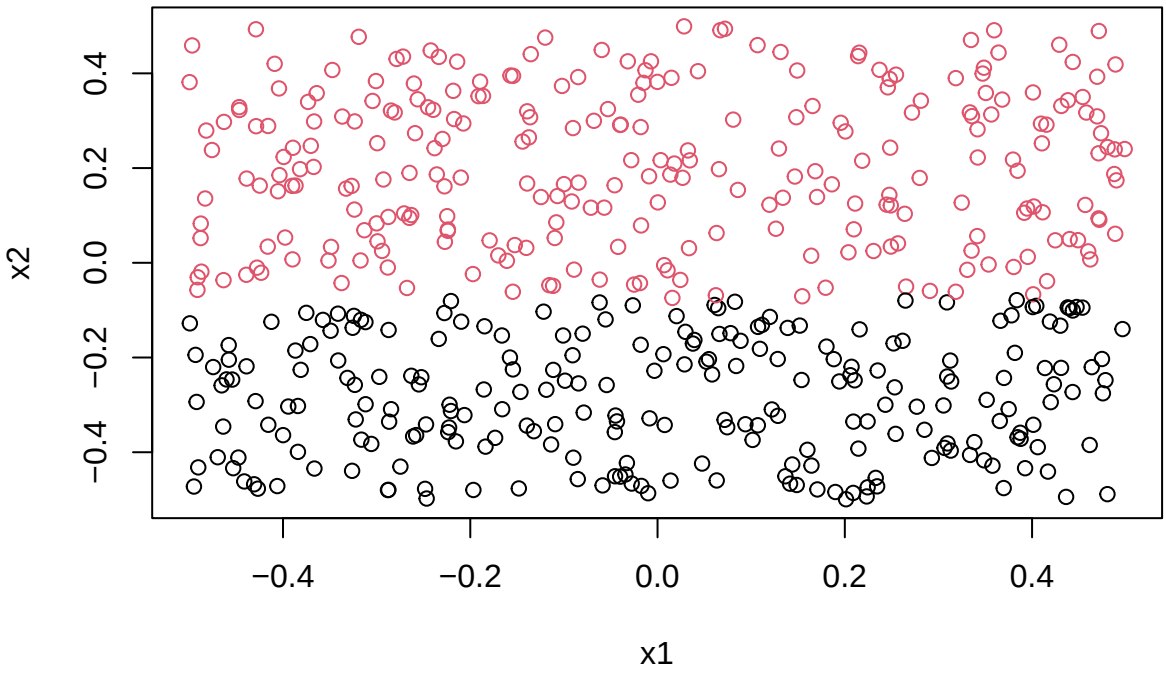
Part (c and d)

- As expected, the decision boundary is linear. Error rate: 43.8%. The predicted probabilities, and the error rates are impacted by the set.seed value. In some cases all the observations are predicted to be the same class.



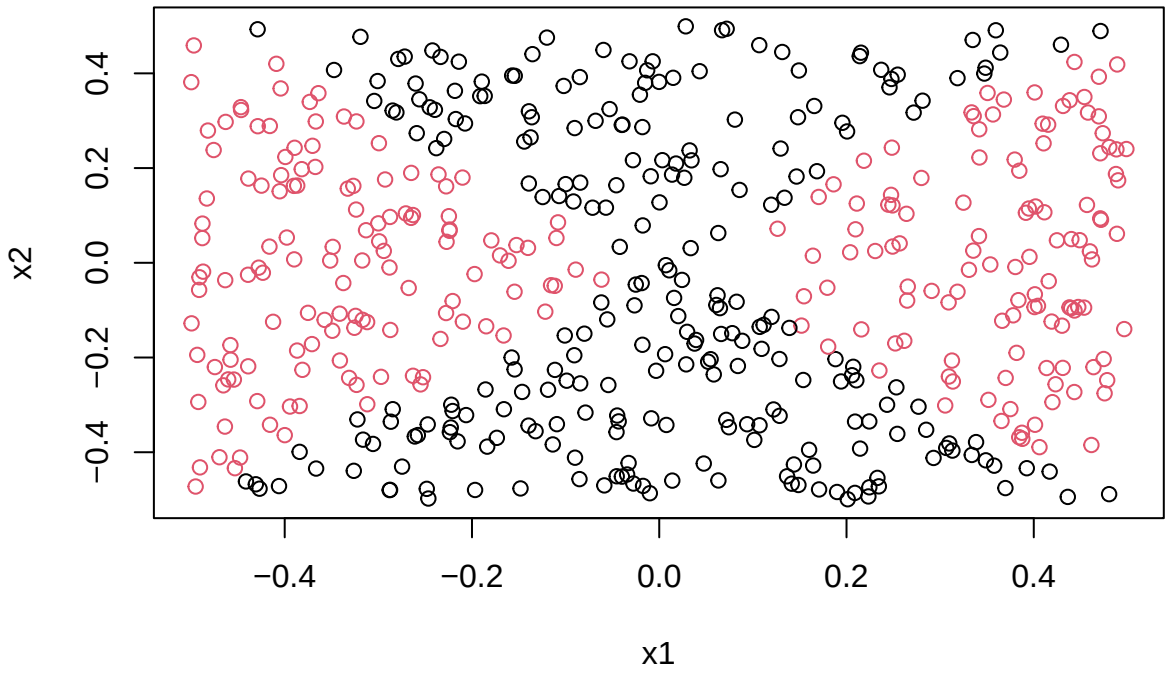
Part (e and f)

- Quadratic transformations of X1 and X2 result in perfect separation. All observations are correctly classified. Error rate : 0%.



Part (g)

- Error rate: 41.4%.



Part (h)

- Error rate: 3.6%.

Part (i)

- The logistic regression models using linear terms and non-linear terms closely match the SVM with linear and radial kernels respectively. As expected, logistic regression with non-linear terms and SVM with a radial kernel outperform the models using linear terms. These models achieve near perfect accuracy in predicting the class of the training observations. The results confirm that Logistic Regression and SVM are similar methods.

Question 2

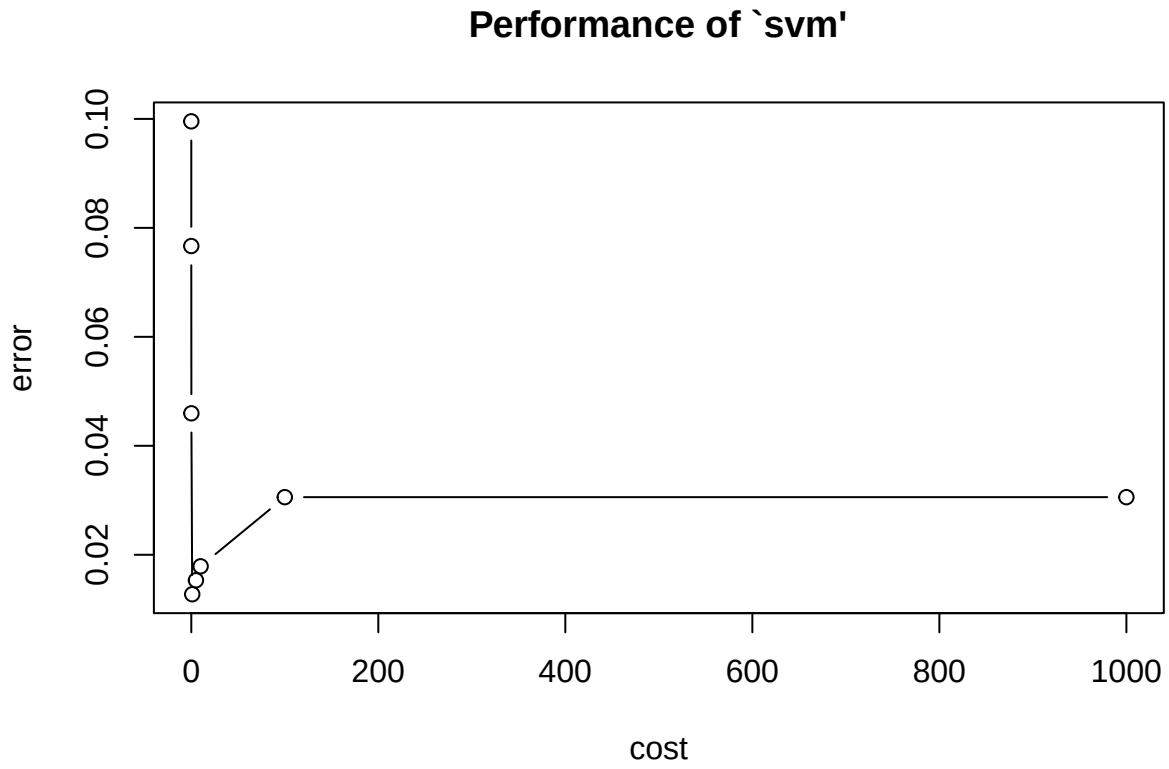
Part (a and b)

- The training CV error decreases as the cost increases, with a minimum at cost=1, thereafter it starts increasing.

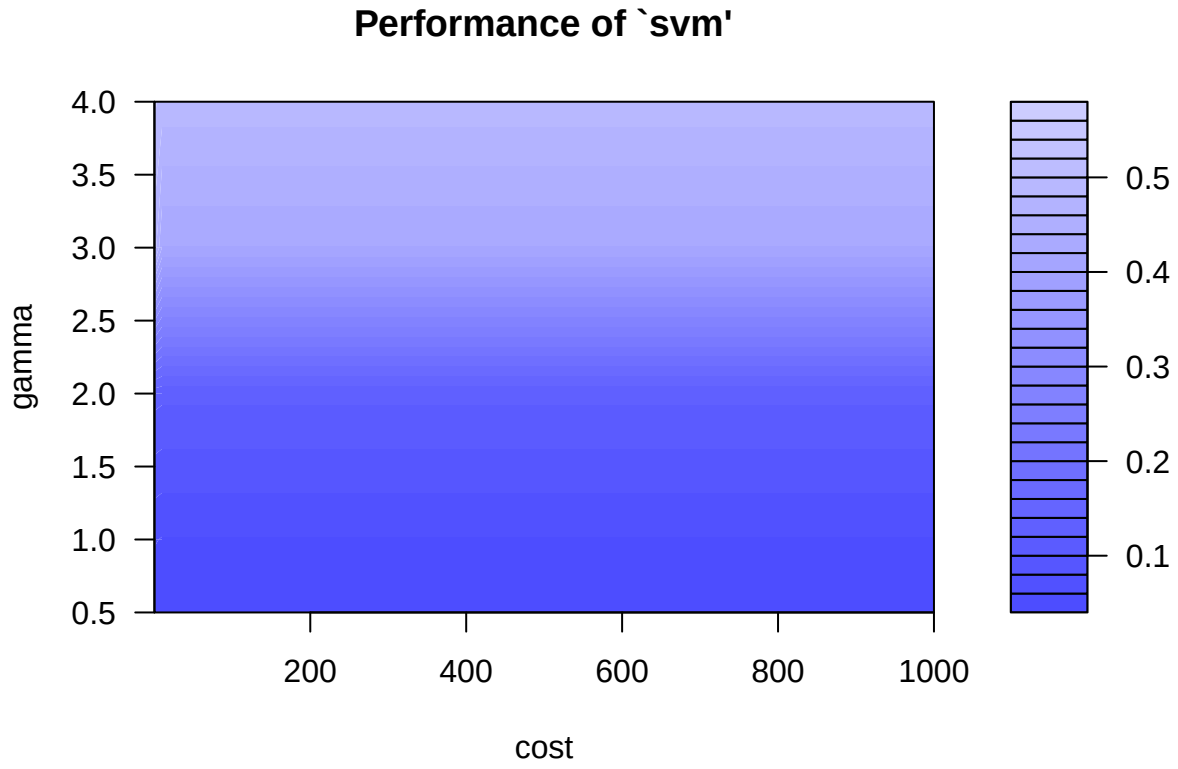
Part (c)

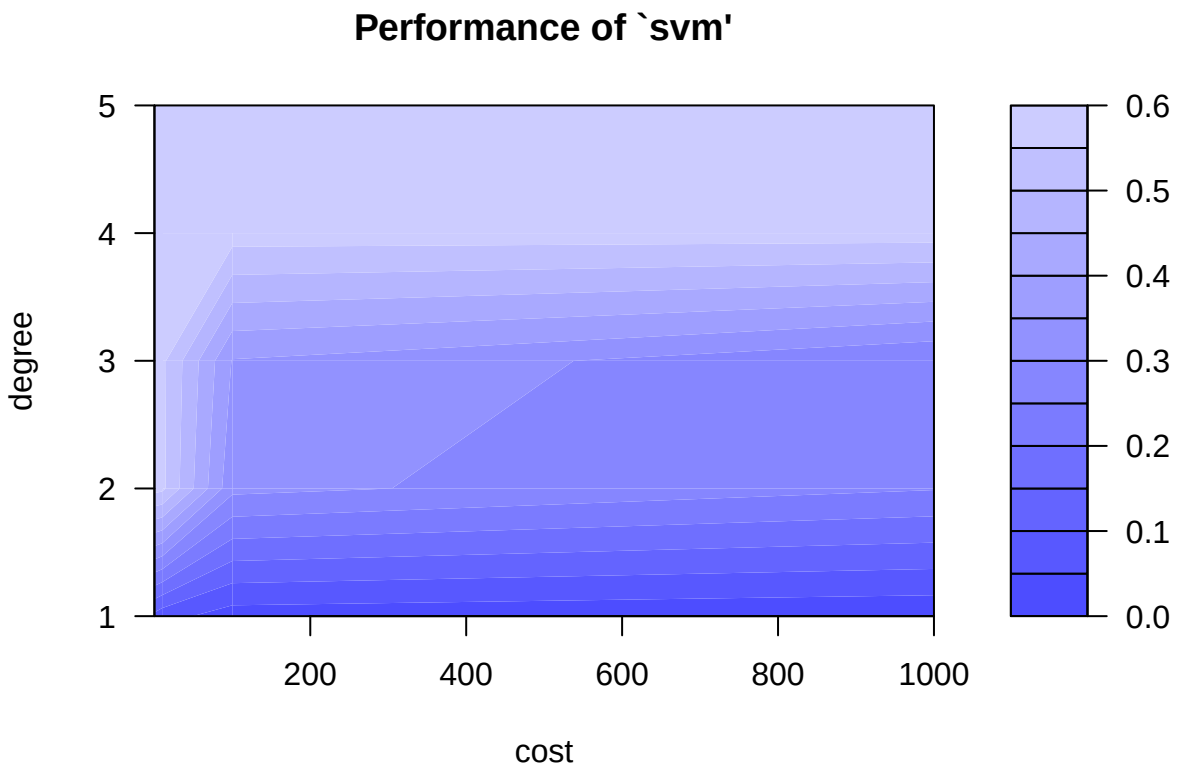
- The training CV error is lowest for a radial model with cost=1 and gamma=0.5, but the value is around 4x higher than for the linear model.

- The best polynomial model is with degree=1 and cost=1000. The lowest training CV errors are given by the linear SVM and polynomial with degree=1, and this suggest the true decision boundary is linear. We would have to test these models on a test set to properly ascertain which of the models is the best.


Performance of `svm'
error
0.02
0.04
0.06
0.08
0.10
0
200
400
600
800
1000
cost

Part (d)


Performance of `svm'
gamma
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
200
400
600
800
1000
cost
0.1
0.2
0.3
0.4
0.5



- SVM with Linear Kernel: Tends to be simpler and faster to train, suitable for linearly separable data. The performance hinges on the right choice of cost.
- SVM with Radial Basis Kernel: More flexible and capable of modeling complex relationships but can be overfitted if not properly tuned. The choice of gamma and cost is critical.
- SVM with Polynomial Kernel: Offers a different form of flexibility than the radial kernel by adjusting the polynomial degree, which can model more complex boundaries than linear SVMs but also risks overfitting with higher degrees.

Question 3

Part (a and b)

- There are a large number of support vectors, so the margin is very wide.

Part (c)

- Training set error rate: $129/800 = 0.16$
- Test set error rate: $50/270 = 0.185$

Part (d and e)

- Training error rate: $128/800 = 0.16$
- Test error rate: $48/270 = 0.178$
- Using the optimal value of cost improves the test error rate slightly.

Part (f)

- Training error rate: $119/800 = 0.149$
- Test error rate: $52/270 = 0.193$

Part (g)

- Training error rate: $109/800 = 0.136$
- Test error rate: $58/270 = 0.215$

Part (h)

- The optimal radial and polynomial models both have a lower training and higher test error rate than the linear SVM. This suggests that both models are over fitting the training set when compared to the linear SVM.
- The linear SVM with optimal cost has a test error rate that is slightly above its training error rate. The small increase is normal behaviour, and the fact that it is still below the radial and polynomial error rates strongly supports the linear SVM being the best model.