CIS 519: Introduction to Machine Learning
Homework 2

Archith Shivanagere Muralinath(PennID: 82629708, PennKey: archith)

Problem 1:

Constant value for α_k: In this case, as we get closer to the minimum, we are slowing down the gradient descent.
α_k as a function of k: In this case, the gradient descent moves at a constant speed towards the minimum.

Problem 2:

(a)

Dataset with only 2 points in 1D: $(x_1 = 0, y_1 = -1)$ and $(x_2 = \sqrt{2}, y_2 = +1)$
Feature Vector $\phi(x) = [1, \sqrt{2}x, x^2]^T$

3D form of $\phi(x_1) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ and

3D form of $\phi(x_2) = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix}$

Vector that is parallel to the optimal vector $\mathbf{w}$ is given by $\phi(x_2) - \phi(x_1)$

$$= \begin{pmatrix} 0 \\ 2 \\ 2 \end{pmatrix}$$

(b)

The given dataset has only 2 points. Margin between these 2 points is the same as the distance between them. Therefore, this is the same as norm of the vector $\phi(x_2) - \phi(x_1)$.

$$= \sqrt{0^2 + 2^2 + 2^2}$$

Value of the margin that is achieved by $\mathbf{w} = 2\sqrt{2}$

(c)

Vector $\mathbf{w}$ is parallel to the vector $\phi(x_2) - \phi(x_1)$, as it is already shown in *problem 2a.*

Using the fact that the margin is equal to $\frac{2}{\|\mathbf{w}\|_2}$, it can be deduced that $\|\mathbf{w}\| = \frac{\sqrt{2}}{2}$.

Since it is parallel to $\phi(x_2) - \phi(x_1)$, we can deduce that $\mathbf{w} = \begin{pmatrix} 0 \\ \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0.5 \\ 0.5 \end{pmatrix}$

(d)

Given the maximum margin classifier with constraints, solving for w_0 by substituting the values of $y_1, y_2, \mathbf{w}, \phi(x_1), \phi(x_2)$ gives us $w_0 \leq -1$ and $w_0 \geq -1$. Therefore, $w_0 = -1$.

(e)

Plugging in values of w_0, w^T,$\phi(x)$ in the equation $h(x) = w_0 + \mathrm{w}^T\phi(x)$ we get

$h(x) = \frac{x^2}{2} + \frac{\sqrt{2}x}{2} - 1$

Problem 3:

(a) Given that hyper-dimensional sphere centered at origin as a classifier, all instances inside the sphere of any radius should be positive. Suppose a positive instance is farther away from origin than a negative instance, the hyper-dimensional sphere does not separate them and fails in this case. Therefore, the VC dimension of this classifier is 1.

(b) If we are able to change the direction of the classification surface, so that we could have anything inside the sphere be predicted positive or everything inside the sphere be predicted negative, then the VC dimension of this classifier is 2.

This is because 2 different instances will always fit but in the case, suppose from the origin there is a negative instance, then a positive instance and then again a negative instance in the same order as distance, we will not be able to fit this in. Therefore, the VC dimension of this classifier is 2.

Part 2: Programming Exercises

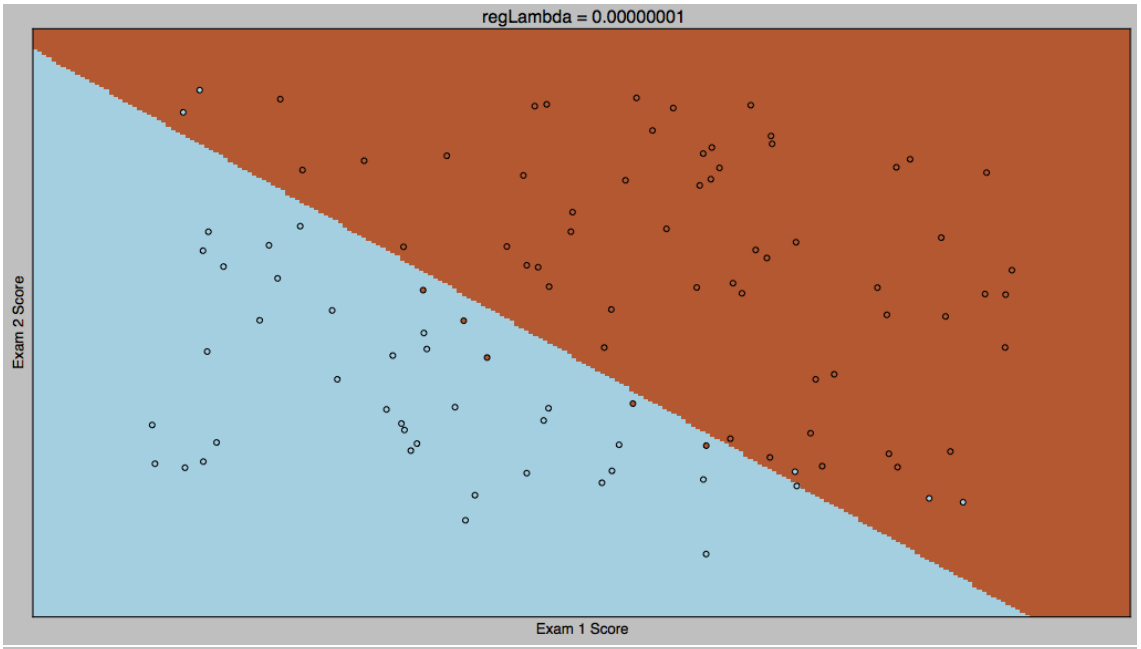
(3.4) Implementing the Gaussian Radial Basis Function Kernel

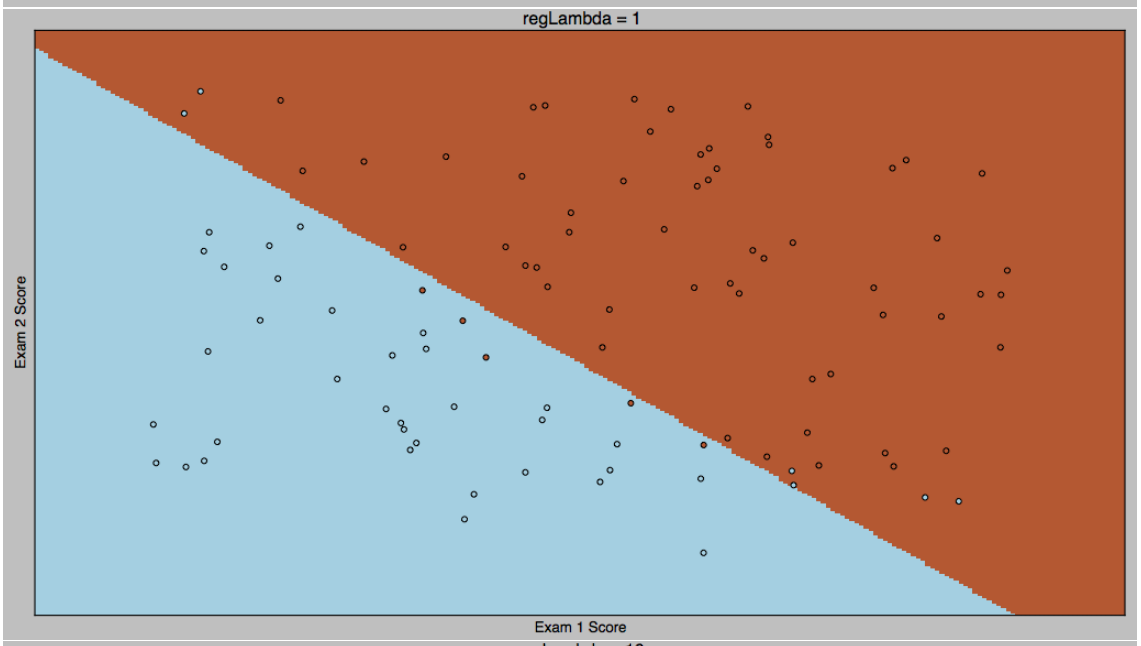
As C increased, I observed that the polynomial kernel classifies more instances correctly in the training data. As d increased, I observed that it took more complicated shape for better accommodation according to test data. In case d is extremely big, the training error is very close to 0 but there is overfitting. As C increased, I observed that the Gaussian kernel, keeping sigma constant, the class in blue color started to expand, and classifier was doing a better job for this set but it was reducing the generalizing capability of classifier. As sigma increased, I observed that the different classes (red, brown, blue) took more straighter shapes i.e, the boundaries of these classes were straightening. For values of sigma less than 1, classifier was getting better and better, and training error was reducing.

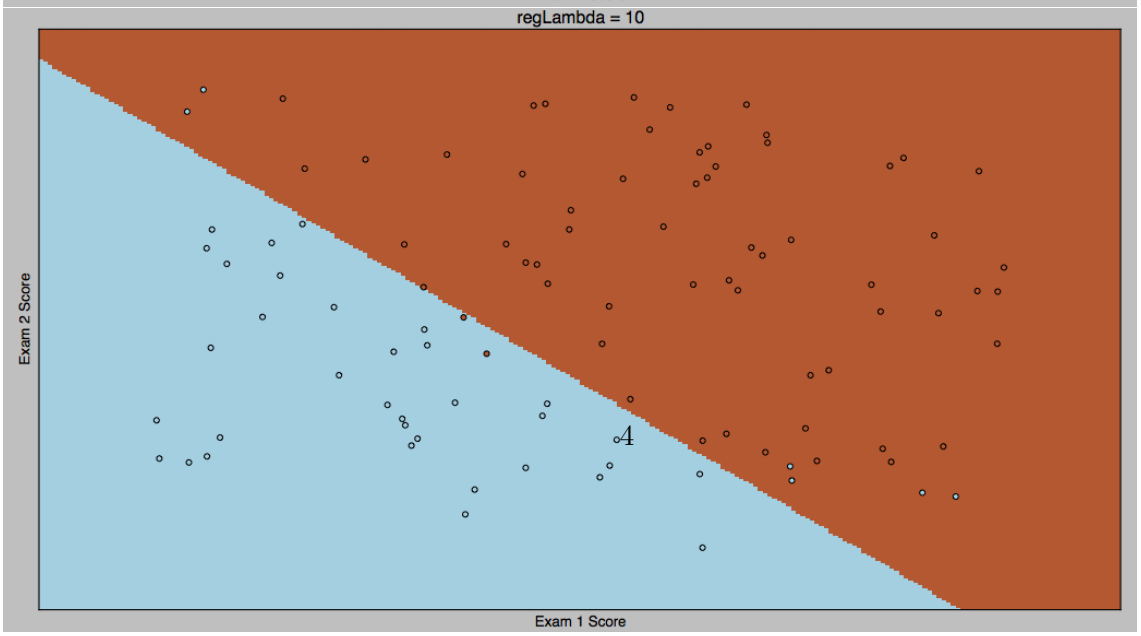
(2.3) Analysis of the regularization parameter

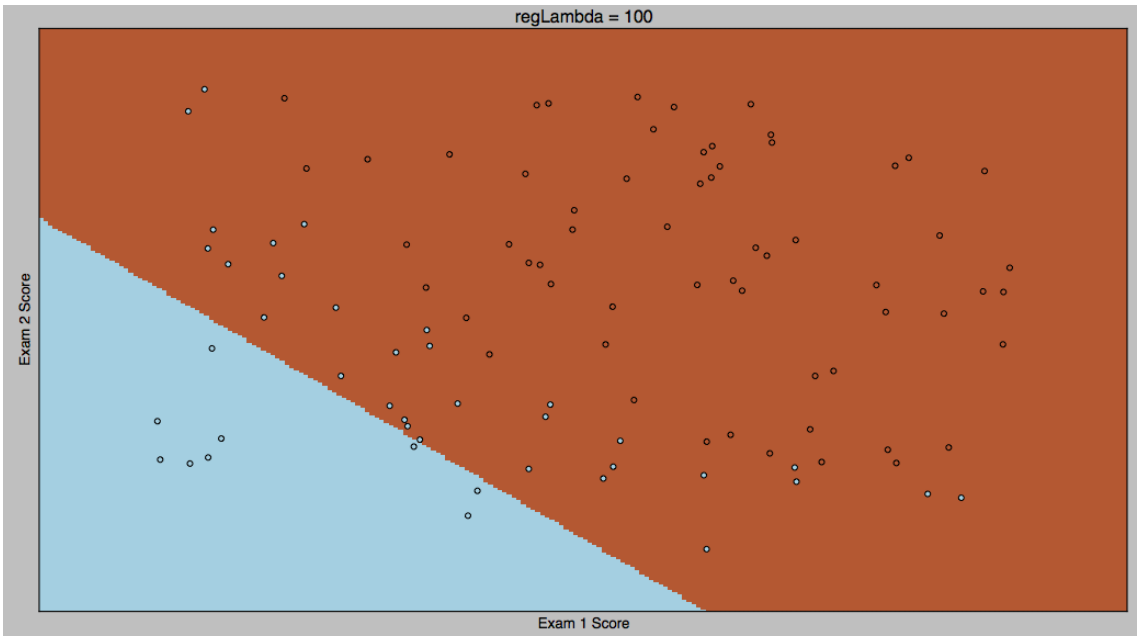
As I increased the value of λ, I observed that the decision surface moves towards one of the classes, and shifts significantly for big values of λ. This is because of increased dependency of gradient function on λ, due to increase in the value of λ. Therefore, as λ increases, more regularized values are generated even though the probabilities scope is wider with more uncertain values.

Please find below different plots with increasing λ.


regLambda = 0.00000001
Exam 2 Score
Exam 1 Score


regLambda = 1
Exam 2 Score
Exam 1 Score


regLambda = 10
Exam 2 Score
Exam 1 Score


regLambda = 100
Exam 2 Score
Exam 1 Score