CS 4641 Homework 2

Budi Ryan — GTID: 903266310

1 Problem 1

Having a constant value of α_k means that we have no control while running the algorithm and will have to wait for a while until we know whether it converges or diverges.
Having α_k as a function of k means that we have a better control in preventing the gradient descent algorithm from diverging.

2 Problem 2

(a) Initially, we map those points into the provided feature vector. We get:

$$\phi(x_1) = [1, 0, 0]^T$$

$$\phi(x_2) = [1, 2, 2]^T$$

In order to get a vector which is parallel to $\vec{w}$, we can simply substract the two 3D vectors due to the fact that there are only 2 vectors overall and the decision boundary must be perpendicular to the line created by linking these 2 vectors. Finally we get the answer:

$$\phi(x_1) - \phi(x_2) = [0, -2, -2]^T$$

The result we got here is the vector which is parallel to $\vec{w}$ we will find later.

(b) Since there are only 2 vectors, the value of the margin achieved by w is the distance between the two vectors $\phi(x_1)$ and $\phi(x_2)$. The margin is given by:

$$margin = \sqrt{(0)^2 + (-2)^2 + (-2)^2}$$

$$margin = 2\sqrt{2}$$

(c) Based on question (a), we have already got a vector which is parallel to the real $\vec{w}$. We have also got the margin value from question (b). Hence we can let $\vec{w}$ to be $[0, k, k]^T$ and calculate k.
Solving for k:

$$margin = \frac{2}{||w||_2}$$

$$2\sqrt{2} = \frac{2}{\sqrt{2}.k}$$

$$k = \frac{1}{2}$$

(d) Using one of the constraints, for example constraint (2), we can calculate w_0.

$$([0, 0.5, 0.5] \cdot [1, 0, 0]^T + W_0) \cdot -1 = 1$$

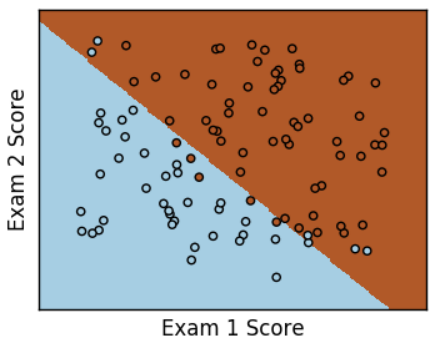
$$W_0 = -1$$

(e) After getting the results from previous parts, we can construct the equation to be:

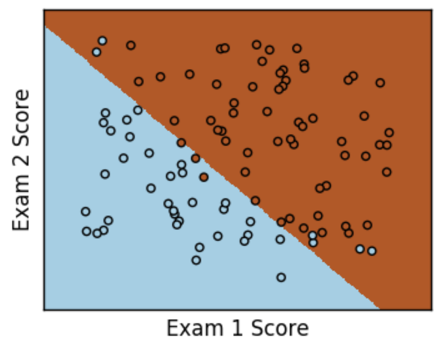
$$h(x) = -1 + \frac{1}{\sqrt{2}} \cdot x + \frac{1}{2} \cdot x^2$$

3 Section 2.3

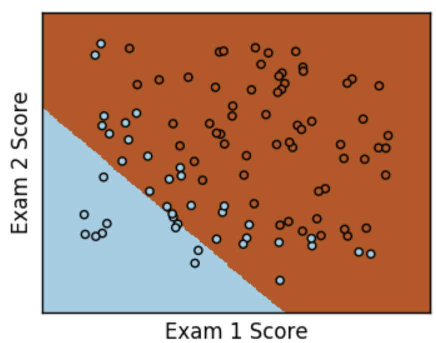
Here are the result of the experiments I did with varying regularization constant λ:



(a) $\lambda = 0.00000001$



(b) $\lambda = 1$



(c) $\lambda = 100$

Explanation:
Figure (a) is using the original regularization constant, while figure (b) and (c) are results of the experiment. From the figures above, we can conclude that regularization reduces over-fitting. However, too much regularization can also reduce the prediction capability, see figure (c).

4 Section 3.3 and 3.4

As C increased, I observed that lesser misclassification happens and increases the chance of overfitting. However, the runningtime of the SVM algorithm also increased.

In gaussian kernel, the higher the sigma is, bias becomes higher and variance becomes lower, and vice-versa.

In polynomial kernel,increasing d(degree of polynomial) increases the chance of overfitting, thus increasing the variance.

5 Section 3.5

Based on the result of my experiment, the optimum values are:
$C = 700$ and $\sigma = 15.00$
The mean accuracy using the built-in function is: 0.984615384615

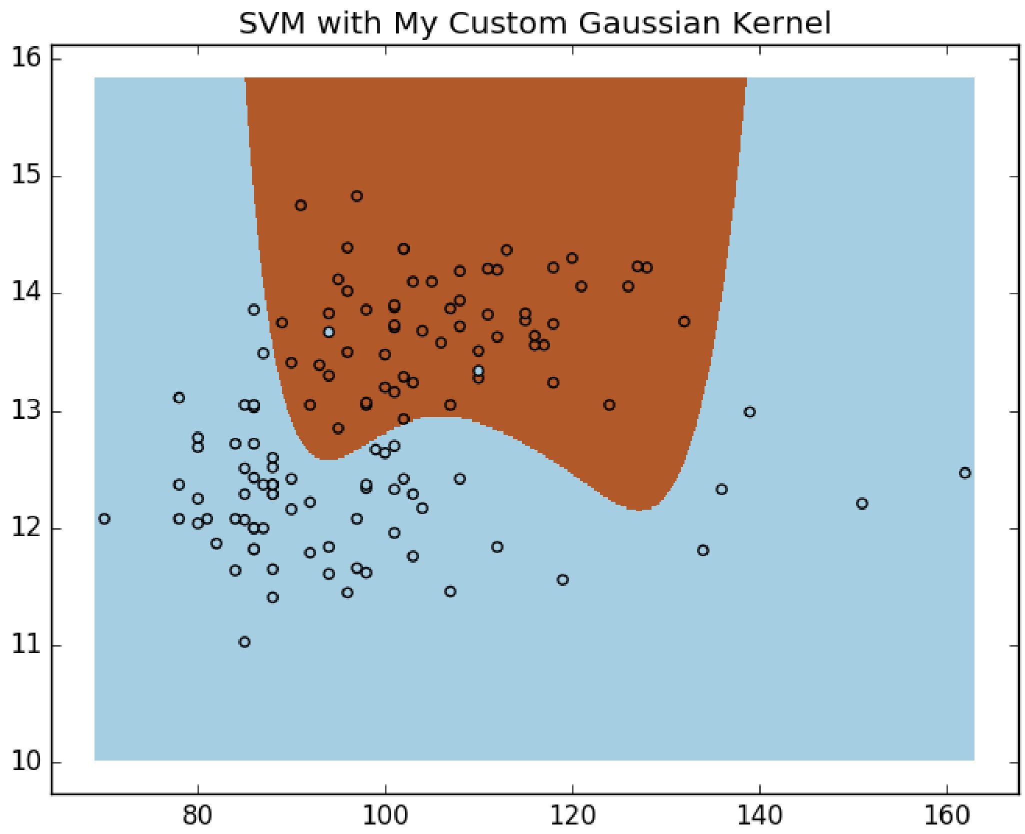


Figure 2: $C = 700$ and $\sigma = 15.00$