3. Problem 3

Question(i):

Problem 3 :-

Question ① :-

Given data points: $\{(x_1, y_1), \dots (x_m, y_m)\}$, $x_j \in \mathbb{R}^n$, $y_j \in \{-1, 1\}$ $j = 1 \dots m$

$S_1 := \{x_j : y_j = 1\}$, $S_2 := \{x_j : y_j = -1\}$.

· Separating hyperplane $H := w^T x + b = 0$.

· SVM training problem:

$$\min \frac{1}{2} \|w\|^2 \quad \text{——} ①$$

$$\text{s.t } 1 - y_j(w^T x_j + b) \leq 0, \; j = 1 \dots m.$$

· KKT Conditions :-

Primal Variables : w, b, $w \in \mathbb{R}^n$, $b \in \mathbb{R}$.

Dual Variable : λ, $\lambda \in \mathbb{R}^n$.

Primal feasibility : $1 - y_j(w^T x_j + b) \leq 0$, $j = 1 \dots m$

Dual feasibility : $\lambda \geq 0$.

Complementary Slackness : $\lambda_j [1 - y_j(w^{*T} x_j + b)] = 0$. — ②

Stationarity : $\left.\begin{array}{l} \nabla_w \mathcal{L}(w^*, b^*, \lambda) = 0 \\ \nabla_b \mathcal{L}(w^*, b^*, \lambda) = 0 \end{array}\right\}$ — ③

where $\mathcal{L}$: Lagrangian,

$$\mathcal{L}(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{j=1}^{m} \lambda_j [1 - y_j (w^T x_j + b)],$$

where $\lambda = [\lambda_1 \dots \lambda_m]^T$.

$$\boxed{\mathcal{L}(w, b, \lambda) = \frac{1}{2} w^T w + \lambda^T (1 - YX^T W) - \lambda^T Y^T b}, \text{ where } Y = \text{diag}(\{y_1 \cdots y_n\})$$

$$\nabla_w \mathcal{L} = w - XY\lambda \qquad \mathcal{L} = \frac{1}{2} w^T w + \lambda^T 1 - w^T (XY\lambda).$$

$$\nabla_w \mathcal{L}\big|_{w=w^*} = 0 \Rightarrow \boxed{W^* = XY\lambda} \quad \text{— } (4)$$

$$\Rightarrow \boxed{w^* = \sum_{i=1}^{m} \lambda_i y_i x_i} \quad \text{— } (5).$$

$$\nabla_b \mathcal{L}(w, b, \lambda) = \nabla_b (\cdots - b^T Y \lambda)$$

$$= -Y\lambda = 0$$

$$\Rightarrow \sum y_i \lambda_i = 0 \Rightarrow \boxed{\lambda^T y = 0} \rightarrow (6).$$

Sub. eq (4) in $\mathcal{L}(w, b, \lambda)$, we get

$$\boxed{g(\lambda) = \min_{w, b} \mathcal{L}(w, b, \lambda)}.$$

$$= \frac{1}{2} (XY\lambda)^T (XY\lambda) + \lambda^T 1 - \lambda^T Y X^T W$$

$$= \lambda^T 1 - \frac{1}{2} (XY\lambda)^T XY\lambda$$

$$= \lambda^T 1 - \frac{1}{2} \lambda^T Y^T X^T X Y \lambda$$

$$\Rightarrow \boxed{g(\lambda) = \lambda^T 1 - \frac{1}{2} \lambda^T (Y \Sigma Y) \lambda} \quad \text{— } \cancel{6}\ 7. \qquad \because Y^T = Y,$$

where $\Sigma = X^T X$,

$$X = [x_1 \cdots x_n] \in \mathbb{R}^{m \times n}$$

Lagrange dual of eq ①, from eq ⑦, ⑥ we get

$$\max \; g(\lambda)$$
$$\text{s.t.} \quad \lambda \geq 0$$
$$\lambda^T y = 0 \leftarrow \text{from eq ⑥}$$

$$\Rightarrow \boxed{\begin{aligned} &\max \; \lambda^T 1 - \tfrac{1}{2}\lambda^T (Y \Sigma Y)\lambda \\ &\text{s.t.} \quad \lambda \geq 0 \\ &\qquad y^T \lambda = 0 \end{aligned}}$$

Obtaining w and b from the optimisation problem:

Obtaining W and b from λ, x, y :-

· Complementary Slackness of KKT :

$$\lambda_j [1 - y_j (w^{T} x_j + b)] = 0 \;\forall i$$

So, $\lambda_j > 0 \Rightarrow 1 - y_j(w^T x_j + b) = 0$. [Active Constraints].

· $\boxed{\text{Support Vectors} : X\_sv = \{x_j : \lambda_j = 0\}. \quad y\_sv = \{y_j : \lambda_j = 0\}.}$

$\lambda_j = 0 \Rightarrow 1 - y_j(w^T x_j + b) < 0$ [Inactive Constraints]

· Now $\boxed{w = \sum_{i=1}^{N\_SV} \lambda_j y_j, x_j}$, $N\_SV := |X\_SV|$

$y_j \in y\_SV$, $x_j \in X\_SV$.

Computing b:-

For support vectors, $1-y_j(\omega^T x_j+b)=0 \ \forall \ y_j, x_j \in y\_sv, X\_sv$

$$\Rightarrow y_j-(\omega^T x_j+b)=0 \quad (\because y_j^2=1 \ \forall \ j)$$

$$\Rightarrow \boxed{b = y_j-\omega^T x_j}$$

· To get a reliable estimate take average of b from all support vectors.

$$\boxed{b=\frac{1}{N_{SV}}\sum_{i=1}^{N\_SV}(y_j-\omega^T x_j)}$$

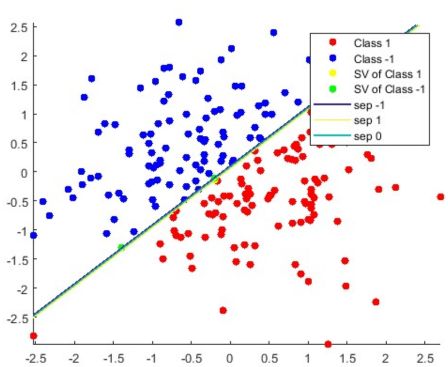
Where $N\_SV := |X\_SV|$

$y_j \in y\_SV,$

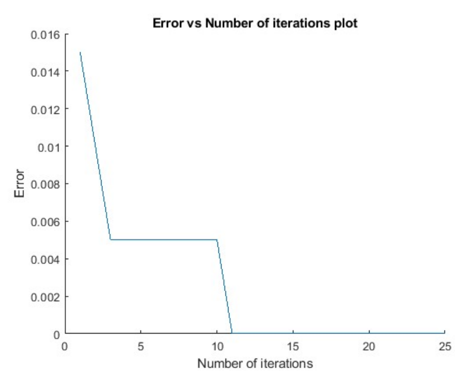
$x_j \in X\_SV.$

Part(ii):

Generated training data and implemented hard margin SVM with its equivalent as given in equation 7 in the question. We get support vectors and separating hyperplane after solving optimisation problem. Obtained misclassification rate and plotted it against number of iterations.



Training data with support vectors and separating hyperplane



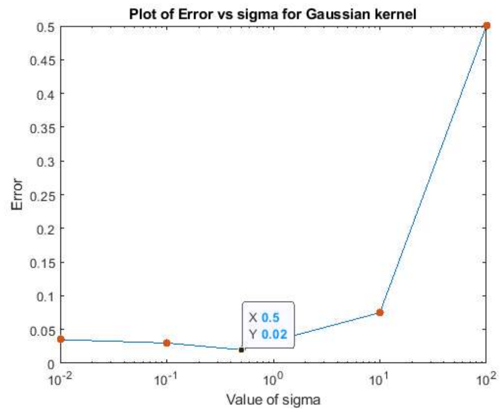
Plot of misclassification error vs iterations

From the plot of error, we can observe that as we increase the number of iterations, error is gradually decreasing and reached optimum value of zero.

Files: BathalaBanuPrasad_SVM.m, BathalaBanuPrasad_Q2.m

Part (iii): SVM with Gaussian Kernel

Used the same training data as before, but used Gaussian kernel. Replaced sigma in the previous SVM method with Gaussian equivalent to get non linear separating hyperplane. Run Gaussian SVM several times with given values of sigma and computed error. Error vs sigma plot as given below.



We can observe that for some sigma = 0.5, we get smallest error of 0.2. Also, as we increase/ decrease sigma beyond this, error is increasing.

Files: BathalaBanuPrasad_SVM_Gau.m, BathalaBanuPrasad_Q3.m

Part (iv): MNIST classification

Part(a): MNIST classification for digits 6 and 8 using eq 5 in the problem.

From the given MNIST data, I have filtered out training data for digits 6 and 8, similarly I have obtaining test data. I have labelled both training and test data corresponding to 6 as 1 and -1 for data corresponding to 8. For finding a separating hyperplane, implemented the optimisation problem given in equation 5. Then, used the training data and training labels for finding the separating hyperplane. Finally, used test data to get test accuracy. Obtained 95.2% accuracy on test data.

Files: BathalaBanuPrasad_Q4.m

Part(b): MNIST classification for digits 1 and 7 using both Hard SVM and SVM with Gaussian kernel

Obtained training and test data with labels as in the previous problem. Now the label corresponding to digit 1 as 1 and -1 for label corresponding to digit 7. Now applied hard SVM algorithm using SVM method implemented in part(ii) and Gaussian Kernel SVM implemented in part (iii) with sigma = 0.5 (as sigma = 0.5 produced lowest error in part 3).

Results:

	SVM without Gaussian kernel	SVM with Gaussian kernel, sigma = 0.5	SVM with Gaussian kernel, sigma = 0.1
Accuracy	0.85	0.9569	0.9569

Files: BathalaBanuPrasad_Q4b.m