Q1

$Q_1(a)$

$$L_p = \frac{1}{2} w^T S w - \nu\rho + \frac{1}{N}\sum_t \xi^t$$

$$\text{subj to } \begin{cases} r^t(w^T x^t + w_0) \geq \rho - \xi^t \\ \xi^t \geq 0 \\ \rho \geq 0, \ \nu \in [0,1] \end{cases}$$

$$L_p = \frac{1}{2} w^T S w - \nu\rho + \frac{1}{N}\sum_t \xi^t - \sum_t \alpha^t [r^t(w^T x^t + w_0) - \rho + \xi^t] - \sum_t \mu^t \xi^t - \eta\rho$$

$\because S$ is symmetric positive definite

$$\begin{cases} \frac{\partial L_p}{\partial w} = Sw - \sum_t \alpha^t r^t x^t = 0 \Rightarrow Sw = \sum_t \alpha^t r^t x^t \\ \frac{\partial L_p}{\partial w_0} = -\sum_t \alpha^t r^t = 0 \Rightarrow \sum \alpha^t r^t = 0 \quad \Rightarrow \text{cancel } w_0 \\ \frac{\partial L_p}{\rho} = -\nu - \sum_t \alpha^t (-1) - \eta = 0 \Rightarrow \nu = \sum_t \alpha^t - \eta \Rightarrow \text{cancel } \rho \\ \frac{\partial L_p}{\partial \xi^t} = \frac{1}{N} - \alpha^t - \mu^t = 0 \Rightarrow \mu^t = \frac{1}{N} - \alpha^t \quad \Rightarrow \text{cancel } \xi^t \end{cases}$$

$$L_p = \frac{1}{2} w^T S w - \left(\sum_t \alpha^t - \eta\right)\rho + \frac{1}{N}\sum_t \xi^t - \sum_t \alpha^t [r^t(w^T x^t + w_0) - \rho + \xi^t] - \sum_t \left(\frac{1}{N} - \alpha^t\right)$$

$$= \frac{1}{2} w^T S w - \sum_t \alpha^t r^t w^T x^t$$

$$= \frac{1}{2} w^T \sum_t \alpha^t r^t x^t - \sum_t \alpha^t r^t w^T x^t$$

$$= -\frac{1}{2} \sum \alpha^t r^t w^T x^t$$

$$= -\frac{1}{2} \sum \alpha^t r^t w^T S^T S^{-1} x^t$$

$$= -\frac{1}{2} \sum \alpha^t r^t (Sw)^T S^{-1} x^t$$

$$= -\frac{1}{2} \sum_t \alpha^t r^t \left[\sum_k \alpha^k r^k (x^k)^T\right] S^{-1} x^t$$

$$L_p = -\frac{1}{2} \sum_t \sum_k \alpha^t \alpha^k r^t r^k (x^k)^T S^{-1} x^t$$

$$\text{subj. } \sum \alpha^t r^t = 0, \ 0 \leq \alpha^t \leq \frac{1}{N}, \ \sum_t \alpha^t \geq \nu$$

Q2.

2. (b)

$$L_p = \frac{1}{2} w^T S w + \sum_t C^t \xi^t - \sum \alpha^t \left[r^t (w^T x^t + w_0) - 1 + \xi_t \right] - \sum \mu^t \xi^t$$

subj. $\xi^t \geq 0$, $r^t (w^T x^t + w_0) \geq 1 - \xi^t$

$\because S$ is positive definite

$$\begin{cases} \frac{\partial L_p}{\partial w} = Sw - \sum \alpha^t r^t x^t = 0 \Rightarrow Sw = \sum \alpha^t r^t x^t \\ \frac{\partial L_p}{\partial w_0} = -\sum \alpha^t r^t = 0 \Rightarrow \sum \alpha^t r^t = 0 \Rightarrow \text{cancel } w_0 \\ \frac{\partial L_p}{\partial \xi^t} = C^t - \alpha^t - \mu^t = 0 \Rightarrow \mu^t = C^t - \alpha^t \Rightarrow \text{cancel } \xi^t \end{cases}$$

$$L_p = \frac{1}{2} w^T S w + \sum_t C^t \xi^t - \sum_t \alpha^t \left[r^t (w^T x^t + w_0) - 1 + \xi_t \right] - \sum_t (C^t - \alpha^t) \xi^t$$

$$= \frac{1}{2} w^T \sum_t \alpha^t r^t x^t - \sum_t \alpha^t r^t w^T x^t + \sum_t \alpha^t$$

$$= \sum_t \alpha^t - \frac{1}{2} w^T \cdot \sum_t \alpha^t r^t x^t$$

$$w^T = \sum_t \alpha^t r^t (x^t)^T \cdot S^{-1}$$

$$L_p = \sum \alpha^t - \frac{1}{2} \sum_k \alpha^k r^k (x^k)^T S^{-1} \sum_t \alpha^t r^t x^t$$

$$= \sum \alpha^t - \frac{1}{2} \sum_K \sum_t \alpha^K \alpha^t r^K r^t (x^K)^T S^{-1} x^t$$

subj $\sum \alpha^t r^t = 0$, $0 \leq \alpha^t \leq C^t$

---

2. For perceptron the update rule is

$w = 0$

$f(x) = \langle w, x \rangle + b$

for k epoch = 1, 2, ..., T

  for all (x_i, r_i), where $i = 1, \dots n$

    if $f(x_i) \cdot r_i \leq 0$

      $w \leftarrow w + r_{ik} x_{ik}$

$$w = 0 + \alpha_1 r_1 x_1 + \alpha_2 r_2 x_2 + \cdots + \alpha_n r_n x_n$$

$$= \sum_{i=1}^{n} \alpha_i r_i x_i$$

α_i is number of times perceptron made mistakes on ith observation within entirely T loops.

New update rule with respect to α_i

$\alpha = 0$

$f(x) = \sum_j \alpha_j r_j \langle x_j, x \rangle + b$ → replace → $\sum_j \alpha_j r_j \phi(x_j, x) + b$

for all (x_i, r_i), where $i = 1, \dots n$

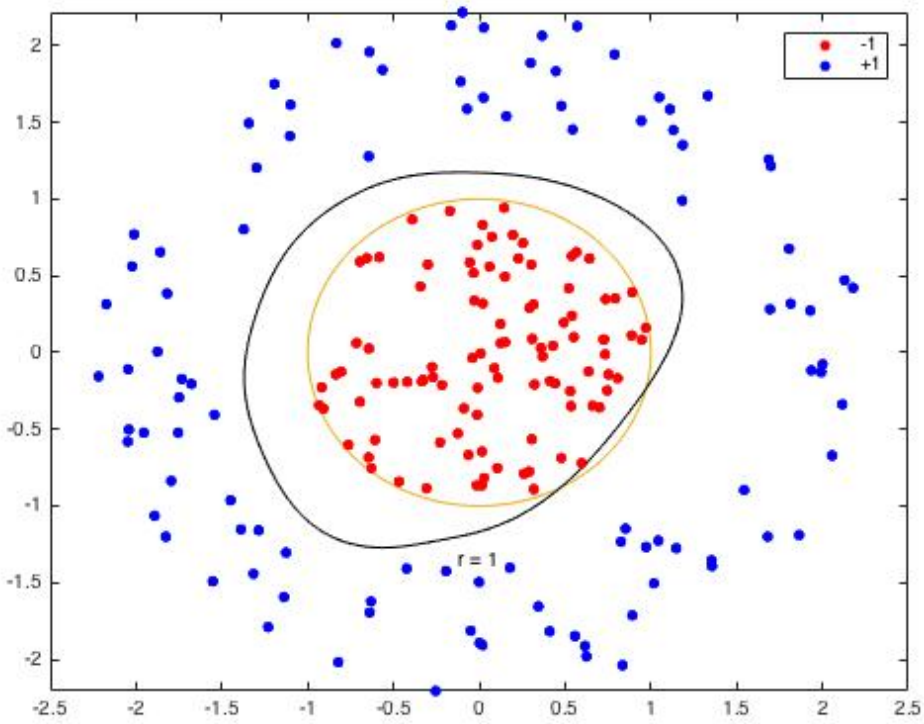
  if $f(x_i) \cdot r_i \leq 0$

    $\alpha_i = \alpha_i + 1$

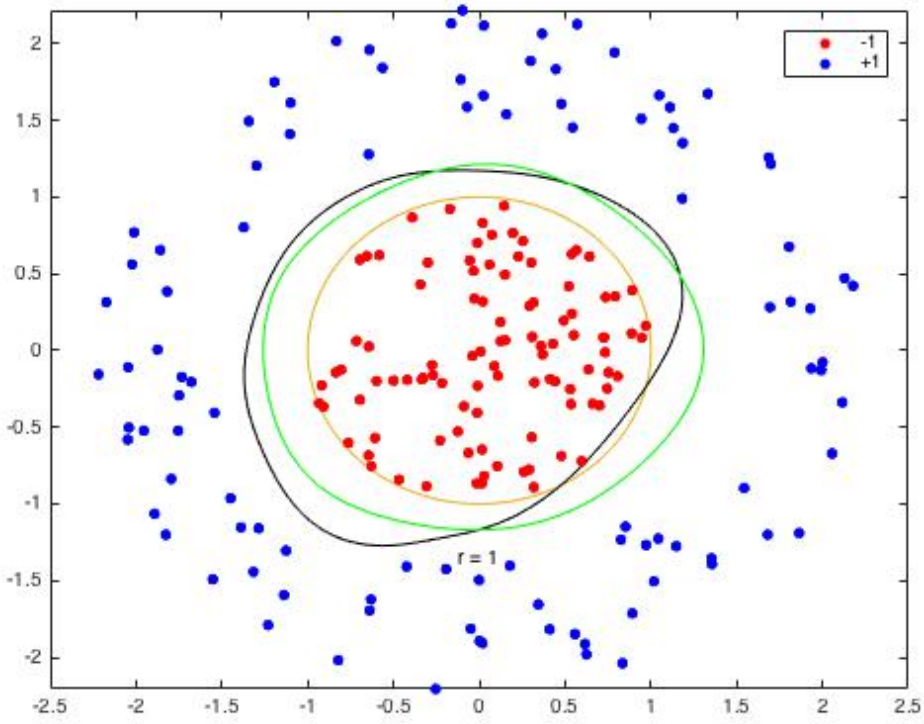
If we replace inner product with kernel function $\phi(x_j, x)$, so $f(x) = b + \sum_j \alpha_j r_j \phi(x_j, x)$

Q3

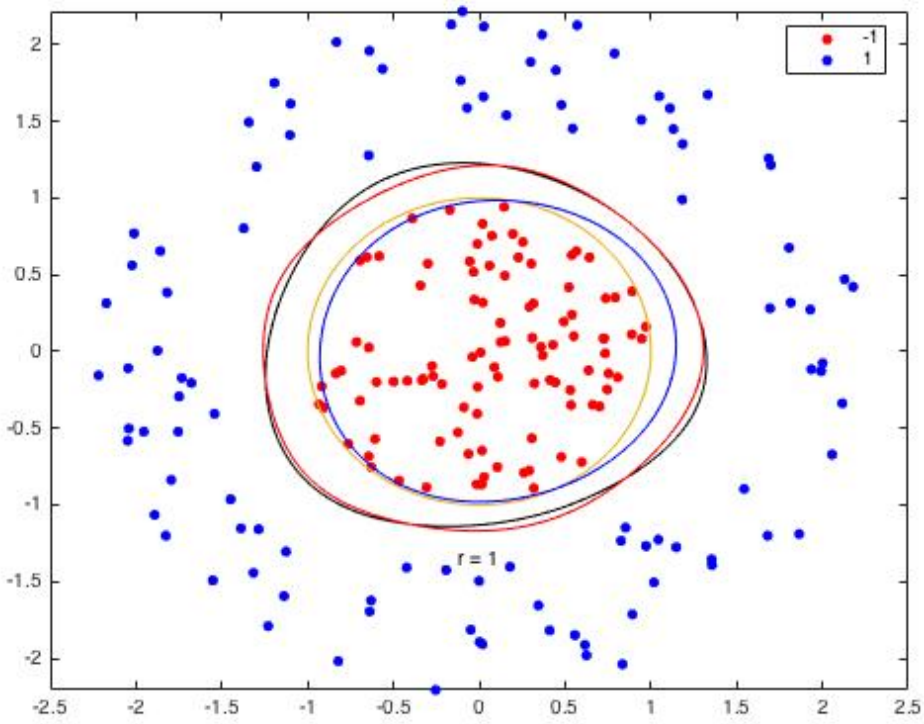
For data3, the error rate =0. The decision boundary for GD kernel is the black curve:



The decision boundary for built-in SVM is the green curve. The decision boundary for GD is the black curve. We can discover that margin around the green curve(SVM) is larger than black curve(perceptron).



When I play with the box constraint parameter, the smaller the box constraint is, the larger the boundary is. The blue curve represents box constraint=0.01 and black represents 1 while the red 100. The box constraint means the value of C in object function $1/2*<w,w>+C*\Sigma_i s_i$, where C is penalty we choose for slack variable. If box constraint or C is large, it means we penalize the severely for slack and we can observe less margin-violating observations and less support vectors. So you can see it as soft margin.



In addition,
On the digits49_train data, the error rate= 0.0047
On the digits49_test data, the error rate= 0.0282
On the digits79_train data, the error rate= 0.0035
On the digits79_test data, the error rate= 0.0035