DATE

Q1a)

$$J(w,b) = \frac{1}{n}\sum (y_i - w^T x_i - b)^2 + \lambda \|w\|^2$$

$$J(\theta) = \frac{1}{n}\sum (y_i - X\theta)$$

$$J(\theta) = (Y - X\theta)^T (Y - X\theta) + \lambda n \|A\theta\|^2$$

$$\theta^T A^T A \theta$$

$$\theta = [b \;\; w_1 \;\; \cdots \;\; w_d]^T \qquad A = \begin{bmatrix} 0 & \cdots & 0 \\ 0 & & 1 \end{bmatrix}$$

$$\frac{dJ(\theta)}{d\theta} = -2X^T Y + 2(X^T X)\theta + 2\lambda n A^T A \theta = 0$$

$$\Rightarrow \quad \theta (X^T X + \lambda n A^T A) = X^T Y$$

$$\therefore \quad \theta = (X^T X + \lambda n A^T A)^{-1} X^T Y$$

& $y = \theta^T x$

Q1 a) No, they won't be equivalent. If we center the data in input space & then apply KRR without offset, the form of kernel will be $\langle \phi(u-\bar{u}), \phi(v-\bar{v}) \rangle$ whereas in case of KRR without offset, kernel will be $\langle (\phi(u) - \phi(\bar{u})), (\phi(v) - \phi(\bar{v})) \rangle$. And these two expressions are different.
Hence they will be different.

b) $$b = \bar{y} - w^T \bar{x}$$

$$= \bar{y} - \frac{1}{\mu} \tilde{y}^T \left[X - \tilde{K} (\tilde{K} + \mu I)^{-1} X \right] \bar{x}$$

$$= \bar{y} - \frac{\tilde{y}^T}{\mu} \left[1 - \tilde{K} (\tilde{K} + \mu I)^{-1} \right] k(x)$$

where $$k(x) = \begin{bmatrix} \langle \tilde{x}_1, \bar{x}_0 \rangle \\ \langle \tilde{x}_2, \bar{x} \rangle \\ \vdots \\ \langle \tilde{x}_n, \bar{x} \rangle \end{bmatrix} = \begin{bmatrix} \langle x_1 - \bar{x}, \bar{x} \rangle \\ \langle x_2 - \bar{x}, \bar{x} \rangle \\ \vdots \\ \langle x_n - \bar{x}, \bar{x} \rangle \end{bmatrix}$$

Typo

DATE

c) $\tilde{K}'$

The ij element of $\tilde{K}'$ is

$$\langle x_i - \bar{x}, x_j' - \bar{x} \rangle = \langle x_i, x_j' \rangle - \frac{1}{n}\sum_{r=1}^{n}\langle x_i, x_r \rangle - \frac{1}{n}\sum_{s=1}^{n}\langle x_s, x_j' \rangle + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n}\langle x_r, x_s \rangle$$

$$= K' - K O_{n\times m} - O_{n\times n} K' + O_{n\times n} K O_{n\times m}$$

where K is the train-train kernel matrix & each element of O is $1/n$ & dimensions of O matrix are ~~different from prior given~~ specified wherever they are mentioned.

Also, for feature mapping $\phi(x)$ first replace x_i with $\phi(x_i)$

For predictions.

$$\hat{f}(x) = \bar{y} + \tilde{X}' \hat{\omega}$$

where $\bar{Y}$ is a vector $\begin{bmatrix} \bar{y} \\ \bar{y} \\ \vdots \\ \bar{y} \end{bmatrix}$ $\bar{y}$ = mean of y for train samples

& $\tilde{X}' = \begin{bmatrix} x_1' - \bar{x} \\ x_2' - \bar{x} \\ \vdots \\ x_m' - \bar{x} \end{bmatrix}$ $\bar{x}$ = mean vector of x_i's for training samples.

$$\therefore \hat{f}(x) = \bar{Y} + \tilde{X}' \left[\frac{1}{\mu} X^T \left[I - (\tilde{K} + \mu I)^{-1} \tilde{K} \right] \tilde{y} \right]$$

$$= \bar{Y} + \frac{1}{\mu}(\tilde{X}' X^T) \left[I - (\tilde{K} + \mu I)^{-1} \tilde{K} \right] \tilde{y}$$

$$= \bar{Y} + \frac{1}{\mu}(\tilde{K}') \left[I - (\tilde{K} + \mu I)^{-1} \tilde{K} \right] \tilde{y}$$

Typo

DATE

d) MSE_{Train} = 10.3465

MSE_{Test} = 49.2640

b = 15.9750

e) MSE_{Train} = 16.4636

MSE_{Test} = = 161.0847.

The KRR without offset is performing worse than KRR with offset. This also seems logical because KRR without offset doesn't provide the best possible fit as it is constrained to pass through origin ~~is the feature not~~.

Typo

$\underset{\omega, \xi_i^+, \xi_i^-}{}$

Q2 $J_{SVR}(\omega, b, \xi^+, \xi^-) = \min_{\omega, b, \xi^+, \xi^-} \frac{1}{2} \|\omega\|^2 + \frac{C}{n} \sum (\xi_i^+ + \xi_i^-)$

DATE

$$\text{st} \quad y_i - \omega^T x_i - b \leq \epsilon + \xi_i^+ \quad \forall i \quad -(1)$$

$$\& \quad \omega^T x_i + b - y_i \leq \epsilon + \xi_i^- \quad \forall i \quad -(2)$$

$$\xi_i^+ \geq 0 \quad \forall i$$

$$\xi_i^- \geq 0 \quad \forall i$$

From (1) & (2)

$$\xi_i^+ \geq y_i - \omega^T x_i - b - \epsilon \quad \& \quad \xi_i^+ \geq 0$$

$$\xi_i^- \geq \omega^T x_i + b - y_i - \epsilon \quad \xi_i^- \geq 0$$

Let

$$(y_i - \omega^T x_i - b) \leq -\epsilon ,$$

$$\therefore \quad \xi_i^+ \geq 0 \geq y_i - \omega^T x_i - b - \epsilon$$

$$\therefore \quad \xi_i^+ = 0 \quad -(3)$$

&

$$\xi_i^- \geq \omega^T x_i + b - y_i - \epsilon \geq -\epsilon + \epsilon$$

$$\xi_i^- \geq \omega^T x_i + b - y_i - \epsilon \geq 0$$

$$\xi_i^- = \omega^T x_i + b - y_i - \epsilon \quad -(4)$$

Or $\xi_i^- = 0$

Let $(y_i - \omega^T x_i - b) \geq \epsilon$

$$\xi_i^+ \geq y_i - \omega^T x_i - b - \epsilon \geq \epsilon - \epsilon \geq 0$$

$$\therefore \quad \xi_i^+ = y_i - \omega^T x_i - b - \epsilon \quad -(5)$$

$$\& \quad \xi_i^- \geq \omega^T x_i + b - y_i - \epsilon$$

$$\geq -\epsilon - \epsilon$$

$$\geq 0 \geq -2\epsilon$$

$$\because \xi_i^- \geq 0$$

$$\therefore \quad \xi_i^- = 0 \quad -(6)$$

Typo

For ③ & ④

If $y_i - \omega^T x_i - b \leq -\epsilon$

$$\left.\begin{aligned} \xi_i^+ &= 0 \\ \xi_i^- &= -y_i + \omega^T x_i + b + \epsilon \\ &= \omega^T x_i + b - y_i - \epsilon \end{aligned}\right\} \text{⑦}$$

From ⑤ & ⑥

If $y_i - \omega^T x_i - b \geq \epsilon$

$$\left.\begin{aligned} \xi_i^+ &= y_i - \omega^T x_i - b - \epsilon \\ \xi_i^- &= 0 \end{aligned}\right\} \text{⑧}$$

If $-\epsilon \leq y_i - \omega^T x_i - b \leq +\epsilon$

$$\xi_i^+ \geq y_i - \omega^T x_i - b - \epsilon$$
$$\geq 0 \geq y_i - \omega^T x_i - b - \epsilon$$

$\therefore \xi_i^+ = 0$

$$\xi_i^- \geq \omega^T x_i + b - y_i - \epsilon$$
$$\geq 0 \geq y_i - \omega^T x_i - b - \epsilon$$

$\therefore \xi_i^- = 0$ ⑨

$\therefore$ From ⑦, ⑧ & ⑨

$$\xi_i^+ + \xi_i^- = \begin{cases} y_i - \omega^T x_i - b - \epsilon & ; \; y_i - \omega^T x_i - b \geq \epsilon \\ 0 & ; \; -\epsilon \leq y_i - \omega^T x_i - b \leq \epsilon \\ \omega^T x_i + b - y_i - \epsilon & ; \; y_i - \omega^T x_i - b \leq -\epsilon \end{cases}$$

or $\xi_i^+ + \xi_i^- = \max\{0, |y - (\omega^T x_i + b)| - \epsilon\}$

$= L_\epsilon(y, \omega^T x_i + b)$

$\therefore J(\omega, b, \xi^-, \xi^+) = \min \frac{1}{2}\|\omega\|^2 + \frac{C}{n} \sum (\xi_i^+ + \xi_i^-)$

Typo

DATE

$$= \min \frac{1}{2}\|w\|_2^2 + \frac{c}{n}\sum l_\epsilon(y_i, w^T x_i + b)$$

$$= \min \frac{1}{n}\sum l_\epsilon(y_i, w^T x_i + b) + \frac{1}{2c}\|w\|_2^2$$

$\therefore$ Let $\lambda = 1/2c$

$$= \min \frac{1}{n}\sum l_\epsilon(y_i, w^T x_i + b) + \lambda \|w\|_2^2$$

b. $L(w, b, \xi^-, \xi^+, \alpha, \beta, \delta) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum(\xi_i^- + \xi_i^+)$

$$+ \sum \alpha_i (y_i - w^T x_i - b - \epsilon - \xi_i^+)$$
$$+ \sum \beta_i (w^T x_i + b - y_i - \epsilon - \xi_i^-)$$
$$- \sum \gamma_i (\xi_i^+) + \sum \delta_i (\xi_i^-)$$

The dual problem is

$$\max_{\alpha \geq 0; \beta \geq 0; \gamma \geq 0; \delta \geq 0} L_D(\alpha, \beta, \gamma, \delta)$$

where

$$L_D(\alpha, \beta, \gamma, \delta) = \min_{w, b, \xi^-, \xi^+} L(w, b, \xi^-, \xi^+, \alpha, \beta, \gamma, \delta)$$

$$\frac{\partial L}{\partial w} = w + \sum \alpha_i (-x_i) + \sum \beta_i x_i = 0$$

$$\therefore w = \sum \alpha_i x_i - \sum \beta_i x_i \quad -(1)$$
$$= \sum (\alpha_i - \beta_i) x_i$$

$$\frac{\partial L}{\partial b} = \sum(-\alpha_i) + \sum \beta_i = 0 \quad -(2)$$

$$\frac{\partial L}{\partial \xi_i^-} = \frac{c}{n} - \beta_i - \delta_i = 0 \quad \forall i \quad -(3)$$

$$\frac{\partial L}{\partial \xi_i^+} = \frac{c}{n} - \alpha_i - \gamma_i = 0 \quad \forall i \quad -(4)$$

Typo

DATE

Using ③ & ④, we can eliminate ξ_i^+ & ξ_i^-

$$L = \frac{1}{2}\|\omega\|^2 + \sum \alpha_i (y_i - \omega^T x_i - b - \epsilon) + \sum \beta_i (\omega^T x_i + b - y_i - \epsilon)$$

$$= \frac{1}{2}\|\omega\|^2 + \sum (y_i - \omega^T x_i - b)(\alpha_i - \beta_i) + \sum \epsilon (-\alpha_i - \beta_i)$$

$$= \frac{1}{2}\|\omega\|^2 + \sum (y_i - \omega^T x_i)(-\beta_i + \alpha_i) + b\{\sum (\beta_i) - \sum \alpha_i\}$$

$$- \epsilon \{\sum \alpha_i + \sum \beta_i\}$$

$$= \frac{1}{2}\|\omega\|^2 + \sum (y_i - \omega^T x_i)(-\beta_i + \alpha_i) - \epsilon \sum \alpha_i$$

Note: $\omega^T x_i = \{\sum_{j=1}^{n} (\alpha_j - \beta_j) x_j^T\} x_i$

$$\|\omega\| = \sum (\alpha_j - \beta_j) x_i^T x (\alpha_i - \beta_i)$$

$$\alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix}$$

$$\|\omega\| = (\alpha - \beta)^T X X^T (\alpha - \beta)$$

$n \times 1 \quad 1 \times n \quad n \times p$

$$= \frac{1}{2}\left(\sum (\alpha_i - \beta_i) x_i\right)\left(\sum (\alpha_j - \beta_j) x_j\right) + \sum y_i (-\beta_i + \alpha_i)$$

$$- \sum (\alpha_j - \beta_j) x_j^T x_i (-\beta_i + \alpha_i) - 2\epsilon \sum \alpha_i$$

$$= \frac{-1}{2} \sum (\alpha_i - \beta_i)(\alpha_j - \beta_j) x_j^T x_i + \sum y_i (\alpha_i - \beta_i) - 2\epsilon \sum \alpha_i$$

$$\therefore L_D(\alpha, \beta, \xi, \xi^*) = \frac{-1}{2} \sum (\alpha_i - \beta_i)(\alpha_j - \beta_j) \langle x_i, x_j \rangle + \sum y_i (\alpha_i - \beta_i) - 2\epsilon \sum \alpha_i$$

Typo

DATE

$$\sum (\beta_i - \alpha_i) = 0 \quad \text{s.t.}$$

$$C/n - \beta_i - \gamma_i = 0 \quad \forall i$$
$$C/n - \alpha_i - \gamma_i = 0 \quad \forall i$$

or $0 \le \alpha_i \le C/n$ & $0 \le \beta_i \le C/n$ $\forall i$

$$\therefore L_D(\alpha, \beta) = -\frac{1}{2} \sum (\alpha_i - \beta_i)(\alpha_j - \beta_j) \langle x_i, x_j \rangle + \sum y_i (\alpha_i - \beta_i) - 2\epsilon \sum \alpha_i$$

s.t.
$$\sum (\beta_i - \alpha_i) = 0$$
& $0 \le \alpha_i \le C/n$ & $0 \le b_i \le C/n$

Given that the strong duality holds. Therefore using the KKT necessity theorem & its conditions,

(1) From first KKT condition,

$$w^* - \sum \alpha_i x_i + \sum \beta_i x_i = 0$$

or

$$w^* = \sum_{i=1}^{n} (\alpha_i - \beta_i) x_i$$

(2) From complementary slackness,

$$\alpha_i^* (y_i - w^T x_i - b - \epsilon - \xi_i^+) = 0$$

If x_i satisfies above eq,

$$y_i - w^T x_i - b - \epsilon - \xi_i^+ = 0 \quad \text{①}$$
$$y_i - (w^T x_i + b) = \epsilon + \xi_i^+$$

Typo

DATE

&

$$\beta_i^* (\omega^T x_i + b - y_i - \epsilon - \xi_i^-) = 0$$

If x_i satisfies above eq. then,

$$\omega^T x_i + b - y_i = \epsilon + \xi_i^- \quad \text{--- (2) or}$$

∴ All x_i which satisfies (1) & (2) will be called support vectors.

∵ if x_i is not a SV, then $\alpha_i^* = 0$ or $\beta_i^* = 0$

ω^* depends on SVs only,

$$\therefore \omega^* = \sum_{\substack{\text{support} \\ \text{vectors}}} (\alpha_i^* - \beta_i^*) x_i$$

(3) From complimentary slackness

$$\gamma_i^* \xi_i^* = 0$$

& we know,

$$\gamma_i^* + \alpha_i^* = C/n$$

ie if $\alpha_i^* < C/n$, then $\gamma_i^* > 0$ & $\xi_i^* = 0$

Ily

~~if $\beta_i^* < C/n$~~

$$\delta_i^* \xi_i^- = 0$$

(we know,

$$\delta_i^* + \beta_i^* = C/n$$

ie if $\beta_i^* < C/n$, then $\delta_i^* > 0$ & $\xi_i^- = 0$

∴ Considering any i s.t $0 < \alpha_i^* < C/n$ & $0 < b_i^* < C/n$

Then

$$\omega^T x_i + b - y_i = \epsilon \Rightarrow b^* = y_i - \omega^T x_i + \epsilon$$

$$\text{Or } y_i - (\omega^T x_i + b) = \epsilon \rightsquigarrow b^* = y_i - \omega^T x_i - \epsilon$$

Typo

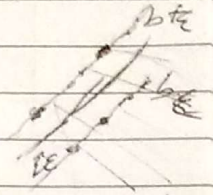
DATE

So, the possible values for b^* are

$y_i - w^T x_i + \epsilon$ when $\alpha_i < C/n$ or $y_i - w^T x_i - \epsilon$ when $\beta_i < C/n$

&

$$w^* = \sum_{\text{support vectors}} (\alpha_i^* - \beta_i^*) x_i$$



c) The SVR is

$$\hat{f}(x) = w^{*T} x + b^*$$
$$= \sum (\alpha_i^* - \beta_i^*) x_i^T x + b^*$$
$$= \sum (\alpha_i^* - \beta_i^*) K(x_i, x) + b^*$$

where $\alpha_i^* = [\alpha_1^*, \ldots, \alpha_n^*]$ & $\beta_i^* = [\beta_1^*, \ldots, \beta_n^*]$ is the sol. of

$$\max \frac{-1}{2} \sum (\alpha_i - \beta_i)(\alpha_j - \beta_j) \langle x_i, x_j \rangle + \sum y_i (\alpha_i - \beta_i) - 2\epsilon \sum \alpha_i$$

st

$$\sum (\beta_i - \alpha_i) = 0$$

& $0 \le \alpha_i \le C/n$ & $0 \le \beta_i \le C/n$

& b^* is given by

$y_i - w^T x_i + \epsilon$ where $\alpha_i < C/n$ OR $y_i - w^T x_i - \epsilon$ when $\beta_i < C/n$

d) Only the training examples which are part of support vector will ~~be considered~~ determine the final predictor because w^* won't be depending on parameters

Support vectors are the x_i's which satisfies either of the below condition

ie

$$w^T x_i + b \mp y_i - \epsilon = \xi_i \ge 0$$

Typo

DATE

or $y_i - (w^T x_i + b) - \epsilon = \xi^* \geq 0$

ie all the points which satisfies

$$|y_i - (w^T x_i + b)| \geq \epsilon$$

ie all the points for which predictor error is greater than or equal to ϵ.

```
clear all;
close all;
clc;
load bodyfat_data.mat;

% Creating test and train data sets
X_train = X(1:150,:);
X_test  = X(151:end,:);
Y_train = y(1:150);
Y_test  = y(151:end);

sigma = 1.5;
lambda = 0.003;
[n, p] = size(X_train);

%%%%%%%%%%%%%% PART D %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
K_train = exp(-1/(2 * sigma ^ 2) * dist2(X_train, X_train));
O_mat   = ones(150, 150) * 1/n;
K_train_tilda = K_train - K_train * O_mat - O_mat * K_train + O_mat * K_train * O_mat;

[m, p] = size(X_test);
K_test = exp(-1/(2 * sigma ^ 2) * dist2(X_train, X_test));
O_mat_test = ones(n, m) * 1/n;
K_test_tilda = K_test - K_train * O_mat_test - O_mat * K_test + O_mat * K_train *
O_mat_test;

% train mean squared error
u = n * lambda;
y_mean = mean(Y_train);
B = (eye(n) - (K_train_tilda + u * eye(n)) \ K_train_tilda);
y_mean_vec = ones(n, 1) * y_mean;
Y_train_centered = Y_train - y_mean_vec;
y_train_pred = y_mean_vec + K_train_tilda' * B * Y_train_centered / u;
mean_squared_error_train = (y_train_pred - Y_train)' * (y_train_pred - Y_train) / n; %
10.3465

% test mean squared error
y_mean_vec = ones(m, 1) * y_mean;
y_test_pred = y_mean_vec + K_test_tilda' * B * Y_train_centered / u;
mean_squared_error_test = (y_test_pred - Y_test)' * (y_test_pred - Y_test) / m; %
49.2640

% b
x_mean = mean(X_train, 1);
K_b = exp(-1/(2 * sigma ^ 2) * dist2(X_train, x_mean));
B_b = (eye(n) - K_train_tilda * inv(K_train_tilda + u * eye(n)));
k_b_xmean = K_b - ones(n,1);
b = y_mean - Y_train_centered' * B_b * K_b / u; % 15.9750


%%%%%%%%%%%%%% PART E %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
K_train = exp(-1/(2 * sigma ^ 2) * dist2(X_train, X_train));
K_test = exp(-1/(2 * sigma ^ 2) * dist2(X_train, X_test));
B = (eye(n) - (K_train + u * eye(n)) \ K_train);
y_train_pred = K_train' * B * Y_train / u;
MSE_train_off = (y_train_pred - Y_train)' * (y_train_pred - Y_train) / n; % 16.4636
```

```
% test mean squared error
y_test_pred = K_test' * B * Y_train / u;
MSE_test_off = (y_test_pred - Y_test)' * (y_test_pred - Y_test) / m; % 161.0847
```