Problem Set 3

Name: Yongqian Li
Student Id #: 004997466

1. Kernals

(a). $K_\beta(x, z) = (1 + \beta x^T z)^3$

$$= 1^3 + 3\beta x^T z + 3\beta^2 z^T x x^T z + \beta^3 x^T z z^T x x^T z$$

$\because x^T z = x_1 z_1 + x_2 z_2$

$\therefore K_\beta(x, z) = 1 + 3\beta(x_1 z_1 + x_2 z_2) + 3\beta^2 (z_1 x_1 + z_2 x_2)^2 + \beta^3 (z_1 x_1 + z_2 x_2)^3$

$= 1 + 3\beta x_1 z_1 + 3\beta x_2 z_2 + 3\beta^2 z_1^2 x_1^2 + 6\beta^2 z_1 x_1 z_2 x_2 + 3\beta^2 z_2 x_2 + \beta^3 (z_1^3 x_1^3 + 3 z_1^2 x_1^2 z_2 x_2$
$\overline{+3 x_1 z_1 z_2^2 x_2^2 + z_2^3 x_2^3)}$

$= 1 + 3\beta x_1 z_1 + 3\beta x_2 z_2 + 3\beta^2 z_1^2 x_1^2 + 6\beta^2 z_1 x_1 z_2 x_2 + 3\beta^2 z_2^2 x_2^2 + \beta^3 z_1^3 x_1^3$
$+ 3\beta^3 z_1^2 z_2 x_1^2 x_2 + 3\beta^3 z_1 z_2^2 x_1 x_2^2 + \beta^3 z_2^3 x_2^3$

(b). From the expanded cubic,

$\because \phi_\beta(x)^T \phi_\beta(z) = K_\beta(x, z)$

$\therefore \phi_\beta(x)^T \phi_\beta(z) = \begin{bmatrix} 1 & \sqrt{3\beta} x_1 & \sqrt{3\beta} x_2 & \sqrt{3}\beta x_1^2 & \sqrt{6}\beta x_1 x_2 & \sqrt{3}\beta x_2^2 & \beta^{\frac{3}{2}} x_1^3 & \sqrt{3}\beta^{\frac{3}{2}} x_1^2 x_2 & \sqrt{3}\beta^{\frac{3}{2}} x_1 x_2^2 & \beta^{\frac{3}{2}} x_2^3 \end{bmatrix}$
$\begin{bmatrix} 1 \\ \sqrt{3\beta} z_1 \\ \sqrt{3\beta} z_2 \\ \sqrt{3}\beta z_1^2 \\ \sqrt{6}\beta z_1 z_2 \\ \sqrt{3}\beta z_2^2 \\ \beta^{\frac{3}{2}} z_1^3 \\ \sqrt{3}\beta^{\frac{3}{2}} z_1^2 z_2 \\ \sqrt{3}\beta^{\frac{3}{2}} z_1 z_2^2 \\ \beta^{\frac{3}{2}} z_2^3 \end{bmatrix}$

$\therefore \phi_\beta(x) = \begin{bmatrix} 1 \\ \sqrt{3\beta} x_1 \\ \sqrt{3\beta} x_2 \\ \sqrt{3}\beta x_1^2 \\ \sqrt{6}\beta x_1 x_2 \\ \sqrt{3}\beta x_2^2 \\ \beta^{\frac{3}{2}} x_1^3 \\ \sqrt{3}\beta^{\frac{3}{2}} x_1^2 x_2 \\ \sqrt{3}\beta^{\frac{3}{2}} x_1 x_2^2 \\ \beta^{\frac{3}{2}} x_2^3 \end{bmatrix}$

(c). When $\beta \to 0$, $K_\beta(x, z) \to 1$

$\beta \to \infty$, $K_\beta(x, z) \to \infty$

$\beta = 1$, $K_\beta(x, z) = K(x, z)$

$\beta \in (0, 1)$, $K_\beta(x, z)$ has more weight on low-order terms $\overbrace{x_1 z_1, x_2 z_2}^{\text{such as}}$

$\beta > 1$, $K_\beta(x, z)$ has more weight on high-order terms

Similarities: The same number of entries and the same computational complexity.

$\therefore$ Comparing with $K(x, z)$, $K_\beta(x, z)$ has the parameter β which weights the terms differently with different choices of β's values. Different β lead to different weighting strategies. scales the entries.

2. SVM

(a). $\begin{cases} y_1 w^T x_1 \geq 1 \\ y_2 w^T x_2 \geq 1 \end{cases} \implies \begin{cases} w_1 + w_2 \geq 1 \\ -(w_1 + 0) \geq 1 \end{cases}$

$\therefore w_1 + w_2 \geq 1$ ①

$\because -w_1 \geq 1$ ②   ② $\implies w_1 \leq -1$

$\therefore w_1 + w_2 \leq -1 + w_2$

if w_1 decrease, w_1^2 would increase, and $\frac{1}{2}\|w\|^2 = w_1^2 + w_2^2$

$\therefore$ let $w_1 = -1$, $-1 + w_2 \geq 1$

$w_2 \geq 2$
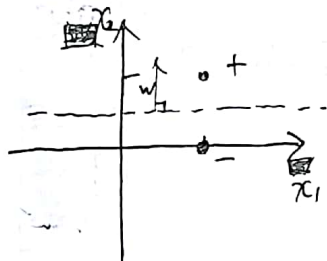
$\therefore \min \frac{1}{2}\|w\|^2 = \frac{1}{2}((-1)^2 + 2^2)$

$= \frac{5}{2}$

$\therefore w^* = (-1, 2)^T$

(b). For part (a), margin $\gamma = \frac{y_1 w^T x_1}{\|w\|} = \frac{1}{\|(-1,2)^T\|} = \frac{1}{\sqrt{5}}$

$\begin{cases} y_1 (w^T x_1 + b) \geq 1 \\ y_2 (w^T x_2 + b) \geq 1 \end{cases}$ $\therefore \begin{cases} w_1 + w_2 + b \geq 1 & ① \\ -(w_1 + 0) - b \geq 1 & ② \end{cases}$

Geometrically, to ~~maxi~~ maximize the margin, we need $w_1 = 0$ ~~s.t. w is vertically~~

$\therefore w = (0, w_2)^T$, $\begin{cases} w_2 + b \geq 1 \\ -b \geq 1 \end{cases}$

$\therefore b \leq -1$,

$w_2 + b \leq w_2 - 1$

$\therefore w_2 - 1 \geq 1$

$\implies w_2 \geq 2$. when $w_2 = 2$, $b = -1$

$\therefore \underline{w^* = (0, 2) \text{ with } b^* = -1}$

$\gamma = \frac{1}{2} > \frac{1}{\sqrt{5}}$

$\therefore$ <u>The margin is larger with offset than without offset.</u>

$\min \frac{1}{2}\|w\|^2 = \frac{1}{2}(2^2) = 2$

3. Twitter analysis using SVMs

3.1.
(a). Implemented

(b). Implemented

(c). Implemented

(d). The feature matrix X has dimentionality : $(630, 1811)$

(Dimentionality of training data is $(560, 1811)$; of test data is $(70, 1811)$)

3.2.
(a). Implemented

(b). It ~~might be~~ is beneficial to maintain class proportions across the folds; because in this way, each ~~split validation~~ split will have training data class proportions ~~have~~ resemble that of the original ~~training~~ data. ~~distribution more than the cases with various proportions.~~ Maintaining class proportions will make the class proportion more similar to that of the total training data, which is ~~also~~ similar to that of the test data, so the results will be representive. ~~(c). Implemented~~ If it is not maintained, there might be some classes in some splits relatively too small, which _{would} lead to strange results.

(c). Implemented

(d).

C	accuracy	F1-score	AUROC
10^{-3}	0.7089	0.8297	0.5000
10^{-2}	0.7107	0.8306	0.5031
10^{-1}	0.8060	0.8757	0.7188
10^{0}	0.8146	0.8749	0.7531
10^{1}	0.8182	0.8766	0.7592
10^{2}	0.8182	0.8766	0.7592
best C	10^{1}	10^{1}	10^{1}

As C increase, scores of all three metrics also increase. In each case, the score stays the same when $C \geqslant 10^{1}$. (roughly)

Generally, ~~accuracy score~~ with the same C, accuracy score is larger than AUROC score, and F1-score is larger than accuracy score.

3.3 (a)- Implemented
   (b)- Indemnted

   (c)、 Choice : $C = \boxed{10}$

	accuracy	F1-score	AUROC
Scores:	0.7429	0.4375	0.6959