Elwyn Cruz
104977892

HW#3

1a) $1 + 3\beta x^T z + 3\beta^2 (x^T z)^2 + \beta^3 (x^T z)^3$

$$1 + 3\beta(x_1 z_1 + x_2 z_2) + 3\beta^2 (x_1 z_1 + x_2 z_2)^2 + \beta^3 (x_1 z_1 + x_2 z_2)^3 = \phi(x)^T \phi(z)$$

$$x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2$$

$$1 + 3\beta x_1 z_1 + 3\beta x_2 z_2 + 3\beta^2 x_1^2 z_1^2 + 6\beta^2 x_1 z_1 x_2 z_2 + 3\beta^2 x_2^2 z_2^2 + \beta^3 x_1^3 z_1^3 + \beta^3 3 x_1^2 z_1^2 x_2 z_2 + \beta^3 3 x_1 z_1 x_2^2 z_2^2 + \beta^3 x_2^3 z_2^3 = \phi(x)^T \phi(z)$$

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

$$K(x,z) = \phi(x)^T \phi(z)$$

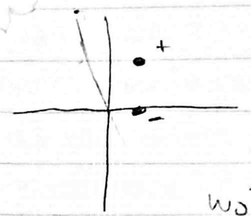
$$K(x,z) = \begin{bmatrix} 1 \\ \sqrt{3\beta}\, x_1 \\ \sqrt{3\beta}\, x_2 \\ \sqrt{3}\beta\, x_1^2 \\ \sqrt{6}\beta\, x_1 x_2 \\ \sqrt{3}\beta\, x_2^2 \\ \sqrt{\beta^3}\, x_1^3 \\ \sqrt{3\beta^3}\, x_1^2 x_2 \\ \sqrt{3\beta^3}\, x_1 x_2^2 \\ \sqrt{\beta^3}\, x_2^3 \end{bmatrix}^T \cdot \begin{bmatrix} 1 \\ \sqrt{3\beta}\, z_1 \\ \sqrt{3\beta}\, z_2 \\ \sqrt{3}\beta\, z_1^2 \\ \sqrt{6}\beta\, z_1 z_2 \\ \sqrt{3}\beta\, z_2^2 \\ \sqrt{\beta^3}\, z_1^3 \\ \sqrt{3\beta^3}\, z_1^2 z_2 \\ \sqrt{3\beta^3}\, z_1 z_2^2 \\ \sqrt{\beta^3}\, z_2^3 \end{bmatrix}$$

1c) $$K(x,z) = (1 + x^T z)^3 = \begin{bmatrix} 1 \\ \sqrt{3}\, x_1 \\ \sqrt{3}\, x_2 \\ \sqrt{3}\, x_1^2 \\ \sqrt{6}\, x_1 x_2 \\ \sqrt{3}\, x_2^2 \\ 1\, x_1^3 \\ \sqrt{3}\, x_1^2 x_2 \\ \sqrt{3}\, x_1 x_2^2 \\ 1\, x_2^3 \end{bmatrix}^T \cdot \begin{bmatrix} 1 \\ \sqrt{3}\, z_1 \\ \sqrt{3}\, z_2 \\ \sqrt{6}\, z_1^2 \\ \sqrt{3}\, z_1 z_2 \\ \sqrt{3}\, z_2^2 \\ 1\, z_1^3 \\ \sqrt{3}\, z_1^2 z_2 \\ \sqrt{3}\, z_1 z_2^2 \\ 1\, z_2^3 \end{bmatrix}$$

For $\beta = 1$, the functions are the same.
As $\beta \to 0$, $K(x,z)$ approaches 1
As $\beta \to \infty$ $K(x,z)$ gets infinitely large.
In the range $0 < \beta < 1$, the terms with β^2 and β^3 become smaller and smaller compared to the β terms.

1c) From this, we can see that β scales the kernel function by some amount. It also weights how much the higher degrees contribute to the value of the kernel function (high β → more contribution, lower β → less)

2a.)



$\min \frac{1}{2}\|w\|^2 \qquad y_n w^T x_n \geq 1$

$\left\| \begin{matrix} w_0 \\ w_1 \end{matrix} \right\| = \sqrt{w_0^2 + w_1^2} \qquad w^T \begin{pmatrix} 1 \\ 1 \end{pmatrix} \geq 1$

$\sqrt{w_0^2 + w_1^2} = 0 \qquad -w^T \begin{pmatrix} 1 \\ 0 \end{pmatrix} \geq 1$

$w_0^2 + w_1^2 = 0 \qquad [w_0 \ \ w_1] \begin{bmatrix} 1 \\ 1 \end{bmatrix} \geq 1$

when $\qquad w_0 + w_1 \geq 1$

$\min \ \frac{1}{2}(w_0^2 + w_1^2) \qquad -w_0 \geq 1$

$\frac{\partial}{\partial w_0} = w_0 \qquad w_0 \leq -1$

$\frac{\partial}{\partial w_1} = w_1 = 1 - w_0 \qquad w_1 \geq 1 - w_0$

minimum is at $w_0 = -1$ and $w_1 = 2$ (bc min will always be @ boundary)

so $w^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

minimize $\frac{1}{2}\|w\|^2$ s.t. $y_n[w^T x_n + b] \geq 1$

$\frac{1}{2}(w_0^2 + w_1^2) \qquad w_0 + w_1 + b \geq 1$

$\frac{\partial}{\partial w_0} = w_0 \qquad -w_0 - b \geq 1$

$\frac{\partial}{\partial w_1} = w_1 = 1 - w_0 - b \qquad w_0 + b \leq -1$

try $w_0 = 0 \rightarrow b \leq -1 \qquad w_1 \geq 1 - w_0 - b$

$w_1 \geq 1 - b$

if $b = -1$, $w_1 \geq 2$ $\quad w_0 = 0$ (since min is @ boundary)

try $w_1 = 0 \rightarrow w_0 + b \leq -1$

$w_1 = 0 \geq 1 - w_0 - b$

$w_0 + b \geq 1$ → not feasible

then $w^* = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$ and $b^* = -1$

Elwyn Cruz
104977892

3.1d) The dimensionality of the feature matrix is n x d where n is the number of tweets and d is the dictionary size (the number of unique words). For our 630 tweets, it is a 630 x 1811 matrix.

3.2b.) By using StratifiedKFold, I am able to maintain the proportions across folds when I split the data. It is useful to maintain this proportion because if it is not maintained, the data could be skewed and inaccurate since it may not be representative of the entire data. Thus, we maintain this proportion so that our classifier can be trained and tested on sets that are representative of all the data.

3.2d)

C	Accuracy	F1-score	AUROC
10^{-3}	0.6635	0.7977	0.5000
10^{-2}	0.7304	0.8192	0.6376
10^{-1}	0.7813	0.8390	0.7387
10^{0}	0.7875	0.8378	0.7557
10^{1}	0.7859	0.8361	0.7557
10^{2}	0.7859	0.8361	0.7557
Best C:	C = 1	C=0.1	C=1

From the looks of it, the performance for all of them seem parabolic, with a maximum around C=1 or C=0.1 (towards the middle C values). F-Score seems to have the best performance, followed by accuracy and then AUROC.

3.3c)

Performance Metric	Score
Accuracy	0.7429
F-Score	0.3448
AUROC	0.6395