CS M146 | Fall 2019

Problem Set 3

Name: Yongqian Li
Student Id #: 004997466

1. Kernals

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

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

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

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

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

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

(b). From the expanded cubic,

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

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

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

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

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

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

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

based on their orders

$\beta \in (0,1)$, $K_\beta(x,z)$ has more weight on low-order terms ~~$x_1 z_1$, $x_2 z_2$~~ since $\beta > \beta^2 > \beta^3 > 0$ ~~such as x_1, x_2.~~

$\beta > 1$, $K_\beta(x,z)$ has more weight on high-order terms since $\beta < \beta^2 < \beta^3$

$\therefore$ 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). $\begin{cases} y_1 W^T x_1 \geq 1 \\ y_2 W^T x_2 \geq 1 \end{cases} \Rightarrow \begin{cases} W_1 + W_2 \geq 1 \\ -(W_1 + 0) \geq 1 \end{cases}$

$\therefore \begin{matrix} W_1 + W_2 \geq 1 & ① \\ -W_1 \geq 1 & ② \end{matrix} \qquad ② \Rightarrow W_1 \leq -1$

$\therefore W_1 + W_2 \leq -1 + W_2$

if W_1 decrease, W_1^2 would increase, and $\frac{1}{2}\|W\|^2 = W_1^2 + W_2^2$

$\therefore$ let $W_1 = -1$, $-1 + W_2 \geq 1$

$W_2 \geq 2$

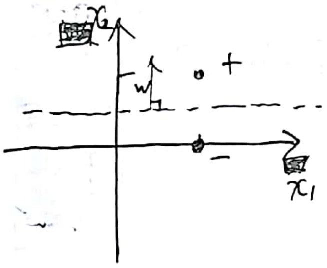
$\therefore \min \frac{1}{2}\|W\|^2 = \frac{1}{2}((-1)^2 + 2^2)$

$= \frac{5}{2}$

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

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

$\begin{cases} y_1 (W^T x_1 + b) \geq 1 \\ y_2 (W^T x_2 + b) \geq 1 \end{cases} \qquad \therefore \begin{matrix} W_1 + W_2 + b \geq 1 & ① \\ -(W_1 + 0) - b \geq 1 & ② \end{matrix}$



Geometrically, to ~~minim~~ maximize the margin, we need $W_1 = 0$ ~~s.t. w is vertically~~

$\therefore w = (0, W_2)^T$, $\begin{cases} W_2 + b \geq 1 \\ -b \geq 1 \end{cases}$

$\therefore b \leq -1$,

$W_2 + b \leq W_2 - 1$

$\therefore W_2 - 1 \geq 1$

$\Rightarrow W_2 \geq 2$. when $W_2 = 2$, $b = -1$

$\therefore W^* = (0, 2)$ with $b^* = -1$

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

$\therefore$ The margin is larger with offset than without offset.

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

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 ~~cross validation~~ split will have training data class proportions ~~[illegible]~~ resemble that of the original ~~to 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~~ 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	0.8759	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
~~10^{3}~~			
best C	~~10^{2}~~ 10^{1}	~~10^{3}~~ 10^{1}	~~10^{3}~~ 10^{1}

As C increase, scores of ~~each~~ all three metrics also increase. In each case, the score stays roughly the same when $C \geq 10^1$.

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)、Implemented

(c)、Choice : C = 10

	accuracy	F1-score	AUROC
Scores:	0.7429	0.4375	0.6259

[redacted]