Tianyang Zhang

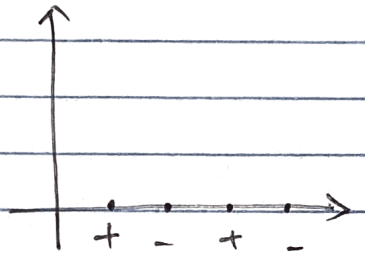
404-743-024

1. $H = \{ \text{sgn}(ax^2+bx+c);\ a,b,c \in R \}$

where $\text{sgn}(ax^2+bx+c) = 1$ if $ax^2+bx+c > 0$

$\text{sgn}(ax^2+bx+c) = 0$ otherwise.

Since quadratic function could change sign only twice. it cannot shatter into alternating 4 VC dimension. The VC dimension of H is 3.



This cannot be shattered.

2. $K_\beta(x,z) = (1+\beta x \cdot z)^3$

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

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

$\therefore \phi_\beta(\vec{x}) = \begin{bmatrix} 1 \\ \vdots \\ \sqrt{3}\, x_i \sqrt{\beta} \\ \vdots \\ \sqrt{3}\, x_i^2 \sqrt{\beta} \\ \vdots \\ \sqrt{6}\, x_i x_j \beta \\ \vdots \\ x_i^3 \sqrt{\beta^3} \\ \vdots \\ \sqrt{3}\, x_i^2 x_j \sqrt{\beta^3} \\ \vdots \\ \sqrt{3}\, x_i x_j^2 \sqrt{\beta^3} \\ \vdots \end{bmatrix}$

The difference between K_β and K is that K_β has the parameter β to scale each term of ϕ_β. Therefore, β plays the role of adjusting the magnitude of the kernel function.

3. a). $x_1 = (1,1)^T$, $x_2 = (1,0)^T$

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

$1 - w_1 - w_2 \leq 0$, $1 + w_1 \leq 0$

$L(w, \alpha) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha_1(1 - w_1 - w_2) + \alpha_2(1 + w_1)$

$d^* = \max_\alpha \min_w L(w, \alpha)$

$\frac{\partial L}{\partial w_1} = w_1 - \alpha_1 + \alpha_2 = 0$,

$\frac{\partial L}{\partial w_2} = w_2 - \alpha_1 = 0$

$\therefore w_1 = \alpha_1 - \alpha_2$

$w_2 = \alpha_1$

Sub w_1, w_2 into $L(w, \alpha)$.

$L(w, \alpha) = \frac{1}{2}((\alpha_1 - \alpha_2)^2 + \alpha_1^2) + \alpha_1(1 - \alpha_1 + \alpha_2 - \alpha_1) + \alpha_2(1 + \alpha_1 - \alpha_2)$

$= \frac{1}{2}(2\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2) + (\alpha_1 - 2\alpha_1^2 + \alpha_1\alpha_2) + (\alpha_2 + \alpha_1\alpha_2 - \alpha_2^2)$

$\frac{\partial L}{\partial \alpha_1} = 2\alpha_1 - \alpha_2 + 1 - 4\alpha_1 + \alpha_2 + \alpha_2 = -2\alpha_1 + \alpha_2 + 1 = 0$

$\frac{\partial L}{\partial \alpha_2} = -\alpha_1 + \alpha_2 + \alpha_1 + 1 + \alpha_1 - 2\alpha_2 = \alpha_1 - \alpha_2 + 1 = 0$

$\therefore \alpha_1 = \frac{\alpha_2 + 1}{2}$

$\alpha_2 = \alpha_1 + 1$

$\therefore \alpha_1 = \frac{\alpha_1 + 2}{2}$

$\therefore \alpha_1 = 2$, $\alpha_2 = 3$.

Since $w_1 = \alpha_1 - \alpha_2$

$w_2 = \alpha_1$

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

b). With offset b be non-zero, we get:

$w_1 + w_2 + b \geq 1 \Rightarrow 1 - w_1 - w_2 - b \leq 0$

$-(w_1 + b) \geq 1 \qquad\qquad 1 + w_1 + b \leq 0$

$\therefore L(w, \alpha, b) = \frac{1}{2}(w_1^2 + w_2^2) + \alpha_1(1 - w_1 - w_2 - b) + \alpha_2(1 + w_1 + b)$

$\frac{\partial L}{\partial w_1} = w_1 - \alpha_1 + \alpha_2 = 0$

$\frac{\partial L}{\partial w_2} = w_2 - \alpha_1 = 0$

$\therefore w_1 = \alpha_1 - \alpha_2$

$w_2 = \alpha_1$

$\Rightarrow L(w, \alpha, b) = \frac{1}{2}(2\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2) + (\alpha_1 - 2\alpha_1^2 + \alpha_1\alpha_2 - \alpha_1 b)$
$+ (\alpha_2 + \alpha_1\alpha_2 - \alpha_2^2 + \alpha_2 b)$

$\frac{\partial L}{\partial \alpha_1} = -2\alpha_1 + \alpha_2 + 1 - b = 0$

$\frac{\partial L}{\partial \alpha_2} = -\alpha_2 + \alpha_1 + 1 + b = 0$

$\therefore \alpha_1 = \frac{\alpha_2 + 1 - b}{2}$

$\alpha_2 = \alpha_1 + 1 + b$

$\alpha_1 = \frac{\alpha_1 + 1 + b - b + 1}{2}, \quad \alpha_1 = 2, \quad \alpha_2 = \alpha_1 = 2$

$\therefore b = -1, \quad w_1 = 0, \quad w_2 = 2$

$\therefore (w^*, b^*) = \left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, -1\right), \quad r = \frac{1}{2}$

with $b = 0$, $r = \frac{1}{\sqrt{5}}$

Therefore, we could have larger margin within decision boundary with offset.

4. 1

d). The dimensionality of the feature matrix is:

dimensionality of the all the tweet: (630, 1140930)

dimensionality of training data: (560, 1014160)

dimensionality of test data: (70, 126770)

4.2

b). It is beneficial to maintain class proportions across folds since the amount of data for each model needs to be reasonable, in order to make the model learn a proper decision boundary. If not doing this, in some extreme case, the model may even not be aware of one of the class for classification.

C	**accuracy**	**F1-score**	**AUROC**
$\mathbf{10^{-3}}$	0.7089	0.8297	0.5
$\mathbf{10^{-2}}$	0.7107	0.8306	0.5031
$\mathbf{10^{-1}}$	0.8060	0.8755	0.7188
$\mathbf{10^{0}}$	0.8146	0.8749	0.7531
$\mathbf{10^{1}}$	0.8182	0.8766	0.7592
$\mathbf{10^{2}}$	0.8182	0.8766	0.7592
Best C	100	100	100

Parameter C is used to decide the boundary, and it could adjust larger slack variables. Larger value C would give a relevantly small margin to avoid misclassification. Smaller value of C would instead give a large margin which may include more misclassification. In our case, the larger C leads to much better performance for our model.

4.3

a). The best C value is 100.

c).

Performance Metric	**Linear-Kernel SVM score**
Accuracy	0.7429
F1_score	0.4375
AUROC	0.6259