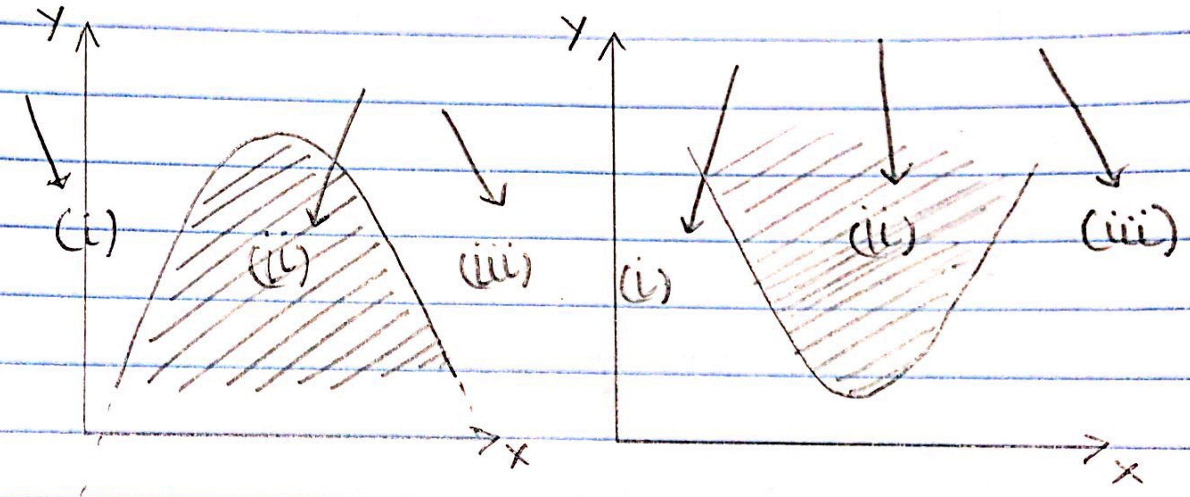
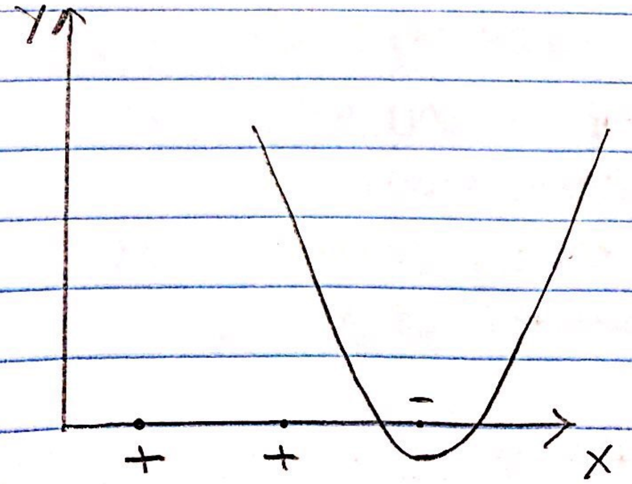
Question 1:

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

The parabolic model, H, has the following general geometric shape:



Therefore, since there are three clear areas, the model can shatter any combination of three points. An example:

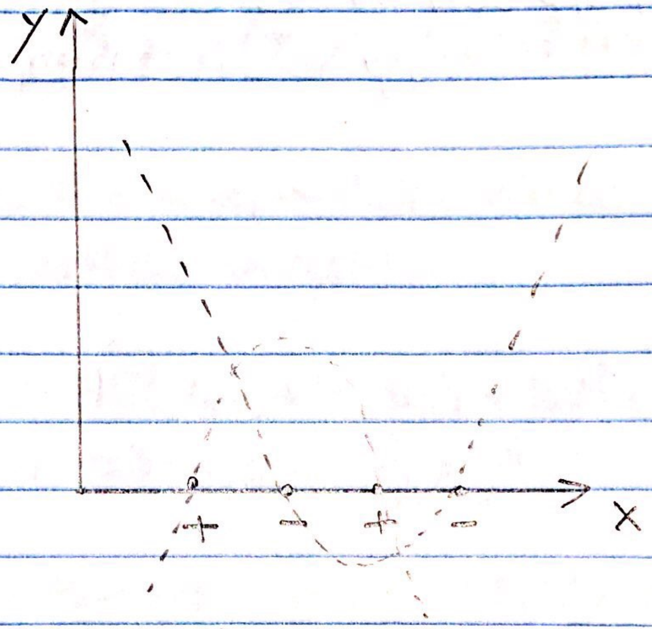


Since the model can shatter any combination of 3 pts.

$VC(H) \geq 3$.

For 4 pts, we can have two cases:

(i) All 4 pts have different positions: In 1D, this can be seen ~~theta~~ as:



By alternating signs, the model is unable to shatter the four pts.

(ii) Atleast 2 pts have the same position: In this case, the pair of pts can be labelled as − and + respectively. Clearly, since they are the same point with different labels, it cannot be shattered.

Therefore, since $VC(H) < 4$ and $VC(H) \geq 3$, we prove that $VC(H) = 3$.

□

Question 2:

When we expand $K_\beta(x, z) = (1 + \beta x . z)^3$, we get:

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

Since $x . z = x_1 z_1 + x_2 z_2$ when $x, z \in R^2$, we can further simplify:

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

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

Therefore, we can get the feature map:

$$\phi_\beta(\cdot) = (1, \sqrt{3\beta} x_1, \sqrt{3\beta} x_2, \sqrt{3}\beta x_1^2, \sqrt{3}\beta\sqrt{2} x_1 x_2, \sqrt{3}\beta x_2^2, \sqrt{\beta^3} x_1^3, \sqrt{\beta^3}\sqrt{3} x_1 x_2^2, \sqrt{3}\sqrt{\beta^3} x_1^2 x_2, \sqrt{\beta^3} x_2^3)$$

$$= (1, \sqrt{3\beta} x_1, \sqrt{3\beta} x_2, \sqrt{3}\beta x_1^2, \sqrt{6}\beta x_1 x_2, \sqrt{3}\beta x_2^2, \sqrt{\beta^3} x_1^3, \sqrt{3\beta^3} x_1 x_2^2, \sqrt{3\beta^3} x_1^2 x_2, \sqrt{\beta^3} x_2^3)$$

Clearly, the feature map is very similar to the feature map of $K(x, z) = (1 + x . z)^3$. However

the n^{th} order term ($x_1^{n}, x_1^{n}x_2, \dots$) will be scaled by $\beta^{n/2}$. This can be used to give more weight to higher-order terms (if $\beta > 1$) or more weight to lower-order terms (if $\beta < 1$).
$\beta > 0$

For $\beta = 1$, $\phi(\cdot) = \phi_\beta(\cdot)$.

□

Question 3:

(a) Let $w^* = \begin{bmatrix} x \\ y \end{bmatrix}$.

Since there are two examples in training, both will be marked as support vectors. Therefore; for both, $y w^T x = 1$:

$$1 \cdot \begin{bmatrix} x \\ y \end{bmatrix} \cdot [1 \; 1] = 1$$

$$\Rightarrow x + y = 1 \quad \text{———} \quad ①$$

$$-1 \cdot \begin{bmatrix} x \\ y \end{bmatrix} \cdot [1 \; 0] = 1$$

$$\Rightarrow -x = 1 \quad \text{———} \quad ②$$

Therefore, we have:

$$x = -1 \Rightarrow -1 + y = 1, \; y = 2$$

$$\therefore w^* = [-1, 2]^T$$

(b) Again, both data points as support vectors will give us:

$$x + y + b = 1 \quad \text{———} \quad ①$$

$$-x + b = 1 \quad \text{———} \quad ②$$

From a geometrical point of view, the w^* vector must be a horizontal line.

$$x = 0$$

$$b = -1 \quad \text{———} \quad \text{from } ②$$

$$0 + y - 1 = 1 = 2 \quad \text{———} \quad \text{from } ①$$

Therefore, we get:

$$w^* = [0 \ \ 2]^T, \ b^* = -1$$

For offset, we get the margin as:

$$\frac{1}{2}\|w^*\|^2 = \frac{1}{2}(0^2 + 2^2) = \underline{\underline{2}}$$

Without offset, we get margin:

$$\frac{1}{2}\|w^*\|^2 = \frac{1}{2}(-1^2 + 2^2) = \underline{\underline{2.5}}$$

Therefore, the margin with offset is less than margin without offset.

Question 4:

4.1.(d) Implemented feature extraction and generated training and testing sets

4.2.(b) If class proportions aren't maintained across the folds, the training set can be a poor representation of the underlying distribution. By stratifying the folds, the training set becomes a better sample of the dataset and the test set accuracy can also be improved.

4.2.(d)

C	**Accuracy**	**F1-score**	**AUROC**
0.001	0.7089419539640778	0.8296828227419593	0.8105494821634063
0.01	0.7107437557658796	0.8305628004640422	0.8110783467587265
0.1	0.8060326761654195	0.875472682955829	0.8575527426160339
1.0	0.8146271113085273	0.8748648327495685	**0.8712327387802071**
10	**0.8181827370986664**	**0.876562152886752**	0.8695790180283852
100	**0.8181827370986664**	**0.876562152886752**	0.8695790180283852
Best *C*	**10/100**	**10/100**	**1.0**

The performance for accuracy and F1-score increase gradually until they reach a plateau at C = 10, where the performance metric stays the same. For AUROC, the performance metric increases until C = 1.0. It then decreases as C increases to 10, where it plateaus similar to the other metrics.

For the next few questions, I have assumed C = 10 as the best hyperparameter for the first two metrics.

4.3.(c)

C	**Metric**	**Performance on test set**
10	Accuracy	**0.7428571428571429**
10	F1-score	**0.43749999999999994**
1.0	AUROC	**0.7405247813411079**

The accuracy and AUROC measures were pretty similar with a value of ~0.74. However, the F1-score reduced drastically from the training to the test set.