Problem 1

a) We can map x and z to vectors that represent the presense of a word. Each vector is the length of all possible words that appear in any document and is index matched to a word. If that word appears in x, it will have a 1 at that index and a 0 otherwise. Thus to find the number of unique words, we can take the dot product of vectors x and z. This is a kernal function as ϕ is 1 if the word is present in this document.

b)

$K(x,z) = x \cdot z$

$f(x) = \frac{1}{\|x\|}$

$f(z) = \frac{1}{\|z\|}$

so $\frac{x}{\|x\|} \cdot \frac{z}{\|z\|}$ is a kernel function

1 is a kernel function if $\phi = 1$ so the dot product is 1.

Thus $1 + \frac{x}{\|x\|} \cdot \frac{z}{\|z\|}$ is kernel function.

Because of the product rule,

$1 + \frac{x}{\|x\|} \cdot \frac{z}{\|z\|}$ times by itself 3 times is also a kernel function

Thus $\left(1 + \frac{x}{\|x\|} \cdot \frac{z}{\|z\|}\right)^3$ is a kernel function.

c) $x = \langle x_1, x_2 \rangle$

$z = \langle z_1, z_2 \rangle$

$$K_B(x,z) = (1 + B(x_1 z_1 + x_2 z_2))^3$$

$$= B^3 x_1^3 z_1^3 + 3B^3 x_1^2 z_1^2 x_2 z_2 + 3B^2 x_1^2 z_1^2 + 3B^3 x_1 z_1 x_2^2 z_2^2$$

$$+ 6B^2 x_1 z_1 x_2 z_2 + 3B x_1 z_1 + B^3 x_2^3 z_2^3 + 3B^2 x_2^2 z_2^2 + 3B x_2 z_2 + 1$$

$$\phi_B = \langle B^{3/2} x_1^3, \sqrt{3} B^{3/2} x_1^2 x_2, \sqrt{3} B x_1^2, \sqrt{3} B^{3/2} x_1 x_2^2, \sqrt{6} B x_1 x_2, \sqrt{3B} x_1,$$

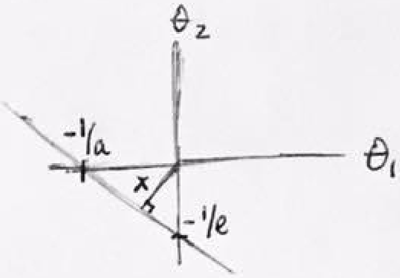
$$B^{3/2} x_2^3, \sqrt{3} B x_2^2, \sqrt{3B} x_2, 1 \rangle$$

$\cancel{K_B(x,z) = (1+(Bx_1z_1 + Bx_2z_2))^3}$

B scales each term to some degree so it can scale or decrease the magnitude of the kernel function.

Problem 2

a) minimizing $\frac{1}{2}(\theta_1^2 + \theta_2^2)$



constraints: $-(a\theta_1 + e\theta_2) \geq 1$

$a\theta_1 + e\theta_2 \leq -1$

slope orthogonal to line:

$\frac{1/e}{-1/a} = -\frac{a}{e} \rightarrow \frac{e}{a}$ orthogonal to line

$\bar{v} = a\hat{\theta}_1 + e\hat{\theta}_2$ (in direction $\frac{e}{a}$)

$\frac{v}{\|v\|} = \frac{a\hat{\theta}_1}{\sqrt{a^2+e^2}} + \frac{e\hat{\theta}_2}{\sqrt{a^2+e^2}}$

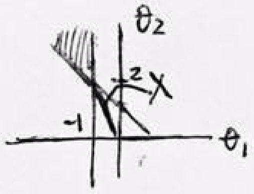
To find x, project θ_1 onto $\bar{v}$:

$x = \frac{\hat{\theta}_1}{-a} \cdot \left(\frac{a\hat{\theta}_1}{\sqrt{a^2+e^2}} + \frac{e\hat{\theta}_2}{\sqrt{a^2+e^2}}\right)$

$= \frac{1}{\sqrt{a^2+e^2}}$

$\theta_1 = \frac{-a}{a^2+e^2} \qquad \theta_2 = \frac{-e}{a^2+e^2}$

b) $\theta_1 + \theta_2 \geq 1$



$-\theta_1 \geq 1 \rightarrow \theta_1 \leq -1$

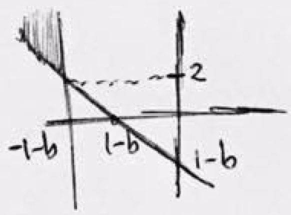
$\bar{v} = \hat{\theta}_1 + \hat{\theta}_2$

$\theta_1^* = -1$

$\theta_2^* = 2$

$\gamma = \frac{1}{\sqrt{1^2+2^2}} = \frac{1}{\sqrt{5}}$

c) $\theta_1 + \theta_2 + b \geq 1$, $-(\theta_1 + b) \geq 1$ $\rangle$ $\theta_1 + \theta_2 \geq 1 - b$, $\theta_1 \leq -1 - b$



$b = -1$ where line is closest to the origin

$\theta_1^* = 0$

$\theta_2^* = 2$

$b^* = -1$

$\gamma = 1/\sqrt{2^2} = \frac{1}{2}$

Problem 3

ii.d)

C	accuracy	F1-score	AUROC	precision	sensitivity	specificity
10^{-3}	0.7089	0.8297	0.5	0.7089	1	0
10^{-2}	0.7107	0.8306	0.5031	0.7102	1	0.0062
10^{-1}	0.8060	0.8755	0.7187	0.8356	0.9294	0.5081
10^{0}	0.8146	0.8748	0.7531	0.8561	0.9016	0.6095
10^{1}	0.8182	0.8765	0.7591	0.8595	0.9016	0.6166
10^{2}	0.8182	0.8765	0.7591	0.8595	0.9016	0.6166
best C	10.0	10.0	10.0	10.0	0.001	10.0

For accuracy, F1-score, AUROC, and precision the performance increases with increased C. As for sensitivity, performance decreases with higher C. Specificity also increases performance with higher C.

iii.a)

γ defines how far the influence of a single training example reaches. A large gamma leads to high bias and low variance while a small gamma leads to low bias and high variance.

iii.b) My grid search went over all combinations of C and γ which had possible values from 10^{-3} to 10^{2} like part 3.2.

iii.c)

metric	score	C	γ
accuracy	0.8165	100	0.01
F1-Score	0.8763	100	0.01
AUROC	0.7545	100	0.01
precision	0.8583	100	0.01
sensitivity	1	0.001	0.001
specificity	0.6047	100	0.01

A higher C and lower γ gave better performance for all metrics except for sensitivity. 100 as C and $\gamma = 0.01$ seemed to be the best parameters.

iv.e)

metric	linear score	rbf score
accuracy	0.7429	0.7571
f1-score	0.4375	0.4516
auroc	0.6259	0.6360
precision	0.6363	0.7
sensitivity	0.333	0.3333
specificity	0.9184	0.9388

The rbf classifier SVM performs better than the linear SVM.