1.(a)

let $a = k(x,x)$ (unique words in x)

$c = k(z,z)$ (" " " z)

$b = k(x,z) = k(z,x)$ (unique words in both x & z)

$$K = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad a \geq b \geq 0, \quad c \geq b \geq 0$$

$$K - \lambda I = \begin{bmatrix} a-\lambda & b \\ b & c-\lambda \end{bmatrix}, \quad \det(K-\lambda I) = (c-\lambda)(a-\lambda) - b^2 = 0$$

$$\Rightarrow \lambda^2 - (a+c)\lambda + ac - b^2 = 0$$

$$\lambda = \frac{a+c \pm \sqrt{(a+c)^2 - 4(ac-b^2)}}{2} = \frac{a+c \pm \sqrt{a^2 - 2ac + c^2 + 4b^2}}{2}$$

$$\lambda_1 = \frac{a+c+\sqrt{a^2-2ac+c^2+4b^2}}{2} \geq 0$$

If $\lambda_2 = \dfrac{a+c-\sqrt{(a-c)^2+4b^2}}{2} \geq 0 \Rightarrow (a+c)^2 \geq (a-c)^2 + 4b^2$

$\Rightarrow 4ac \geq 4b^2 \Rightarrow ac \geq b^2$ (since we know $a, c \geq b \geq 0$, it is correct)

λ_1 & $\lambda_2 \geq 0 \Rightarrow K$ is PSD, the function is a kernel.

1.(b)

since $x \cdot z$ is a kernel, then $\dfrac{x \cdot z}{\|x\|\|z\|} = f(x)\, k_1(x,z)\, f(z)$ is also.
(scaling)

$k(x,z) = 1 = \begin{array}{cc} & N \\ N & \begin{bmatrix} \cdot & \cdots & \cdot \\ \vdots & \ddots & \\ \vdots & & \cdot \end{bmatrix} \end{array}$ has eigenvalues ≥ 0 $(\lambda_1 = N, \lambda_2, \cdots, \lambda_N = 0)$

$\Rightarrow$ is also a kernel.

$\Rightarrow$ $1 + \dfrac{x \cdot z}{\|x\|\|z\|}$ is a kernel (sum)

$\left(1 + \dfrac{x \cdot z}{\|x\|\|z\|}\right)\left(1 + \dfrac{x \cdot z}{\|x\|\|z\|}\right)\left(1 + \dfrac{x \cdot z}{\|x\|\|z\|}\right) = k_1(x,z)\, k_1(x,z)\, k_1(x,z)$

is also a kernel (product)

so, $\left(1 + \left(\dfrac{x}{\|x\|}\right) \cdot \left(\dfrac{z}{\|z\|}\right)\right)^3$ is a kernel.

1.(c)

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

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

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

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

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

$k(x,z) = (1 + xz)^3 = (1 + 1 \cdot xz)^3 = k_1(x,z)$

We can know that the similarities between $k_\beta(x,z)$ and $k(x,z)$ is the feature transform function. The different is that $\phi(\cdot)$ of $k(x,z)$ doesn't has any constant β.

β can let different feature have different weight. For example, if $\beta > 1$, then $\beta^{\frac{3}{2}}$ will be larger than just $\beta^{\frac{1}{2}}$ or β, so it gives the higher power terms bigger weight to influence on the model. In contrast, if $1 > \beta > 0$, then $\beta^{\frac{3}{2}}$ will be smaller, so it gives the higher power terms less influence on the model.

2. (a)

$$p^* = \min_\theta \max_\alpha L(\theta, \alpha)$$

since $y = -1$, $x = (a, e)^T$ $\quad y_n \theta^T x_n \geq 1 \Rightarrow -\theta(a,e)^T \geq 1 \Rightarrow 0 \geq 1 + \theta(a,e)^T$

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \alpha(\theta(a,e)^T + 1)$$

$d^* = \max_\alpha \min_\theta L(\theta, \alpha)$, we use dual to solve:

$$\frac{\partial L(\theta,\alpha)}{\partial \theta} = \theta + \alpha(a,e)^T = 0 \Rightarrow \theta = -\alpha(a,e)^T$$

$$L(\theta,\alpha) = \frac{1}{2}\|-\alpha(a,e)^T\|^2 + \alpha(-\alpha(a,e)(a,e)^T + 1)$$

$$= \frac{1}{2}(\alpha^2 a^2 + \alpha^2 e^2) + (-\alpha^2 a^2 - \alpha^2 e^2 + \alpha)$$

$$\frac{L(\theta,\alpha)}{\partial \alpha} = (a^2+e^2)\alpha - 2(a^2+e^2)\alpha + 1 = -(a^2+e^2)\alpha + 1 = 0$$

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

$$\theta^* = -\frac{1}{a^2+e^2}\begin{bmatrix} a \\ e \end{bmatrix}$$

2.(b)

Since $y_1 = 1$, $x_1 = (1,1)^T \Rightarrow y_n \theta^T x_n \geq 1 \Rightarrow \theta^T (1,1)^T \geq 1 \Rightarrow 0 \geq 1 - \theta_1 - \theta_2$

$\|w\|_2 = \sqrt{\theta_1^2 + \theta_2^2}$ $\quad y_2 = -1$, $x_2 = (1,0)^T \Rightarrow y_n \theta^T x_n \geq 1 \Rightarrow -\theta^T (1,0)^T \geq 1 \Rightarrow 0 \geq 1 + \theta_1$

$$L(\theta, \alpha) = \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha_1(1 - \theta_1 - \theta_2) + \alpha_2(1 + \theta_1)$$

$$\frac{\partial L}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0 \Rightarrow \theta_1 = \alpha_1 - \alpha_2$$

$$\frac{\partial L}{\partial \theta_2} = \theta_2 - \alpha_1 = 0 \Rightarrow \theta_2 = \alpha_1$$

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

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

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

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

$$\Rightarrow \alpha_2 = 1 + \alpha_1$$

$$\alpha_1 = \frac{1 + \alpha_1 + 1}{2} \Rightarrow \alpha_1 = 2,\ \alpha_2 = 3,\ \Rightarrow \theta_1 = -1,\ \theta_2 = 2,\ \theta^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

$\gamma = \frac{1}{\|w\|_2}$ $\quad \gamma = \frac{1}{\sqrt{(-1)^2 + 2^2}} = \frac{1}{\sqrt{5}}$

(c)

We have $\quad \theta_1 + \theta_2 + b \geq 1 \Rightarrow 0 \geq 1 - \theta_1 - \theta_2 - b$

$-(\theta_1 + b) \geq 1 \Rightarrow 0 \geq 1 + \theta_1 + b$

Not $\frac{1}{2}(\theta_1^2 + \theta_2^2 + b^2)$

$$L(\theta, b, \alpha) = \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha_1(1 - \theta_1 - \theta_2 - b) + \alpha_2(1 + \theta_1 + b)$$

$$\frac{\partial L}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0 \Rightarrow \theta_1 = \alpha_1 - \alpha_2$$

$$\frac{\partial L}{\partial \theta_2} = \theta_2 - \alpha_1 = 0 \Rightarrow \theta_2 = \alpha_1$$

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

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

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

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

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

$$\alpha_2 = \alpha_1 + b + 1 = 1 - b + b + 1 = 2 \Rightarrow \alpha_1 = 2, \quad b = 1 - \alpha_1 = -1$$

$$\theta^* = \begin{bmatrix} \alpha_1 - \alpha_2 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \quad b^* = -1, \quad \gamma = \frac{1}{\sqrt{0^2 + 2^2}} = \frac{1}{2}$$

The margin is bigger with offset ($\frac{1}{2}$) than without offset ($\frac{1}{\sqrt{5}}$).

3.2 (b)

Since having the same proportion will ensure that we will have enough training data for both categories, we can get a reasonable model by our training data. In contrast, if we don't keep the proportion equal, it is possible to happen that some training set only has one categorical data. If this happen, the model can't be trained succeedly.

(d)

C	accuracy	F1-score	AUROC	precision	sensitivity	specificity
0.001	0.7089	0.8297	0.5	0.7089	1	0
0.01	0.7107	0.8306	0.5031	0.7102	1	0.0063
0.1	0.8060	0.8755	0.7188	0.8357	0.9294	0.5081
1	0.8146	0.8749	0.7531	0.8562	0.9017	0.6045
10	0.8182	0.8766	0.7592	0.8595	0.9017	0.6167
100	0.8182	0.8766	0.7592	0.8595	0.9017	0.6167
best C	100	100	100	100	0.01	100

When C increases, the accuracy and precision also increase until C reach 10. The sensitivity decreases when C increases. The value of specificity is close to the value of C when C is small, and increase until around 0.61 when C increases.

3.3 (a)

If k(xi, xj) is the RBF-kernel, a small gamma means a Gaussian with a large variance so the influence of xj is more, i.e. if xj is a support vector, a small gamma implies the class of this support vector will have influence on deciding the class of the vector xi even if the distance between them is large. If gamma is large, then variance is small implying the support vector does not have wide-spread influence. In conclusion, gamma defines how far the influence of a single training data can reach. Small gamma can reach far training data, and lead to a low bias and high variance model; large gamma can reach close training data, and lead to a high bias and low variance model.

(b)

I use all pairwise combination of C = [0.001, 0.01, 0.1, 1, 10, 100] and gamma = [0.001, 0.01, 0.1, 1, 10, 100]. The setting can cover a wide range of hyperparameter set, so that can let us find the best model.

(c)

	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	0.1	100
specificity	0.6047	100	0.01

Most of the best result got from the same pair C=100, gamma=0.01, which make sense since large C can penalize slack variables and small gamma can lead to a model with low bias. The result of sensitivity shows that overfitting happened since C is too small so the slack variables didn't be limited and gamma is too large that lead to high bias and low variance. Except the result of sensitivity, all the other results come out with a score around 0.6 to 0.9.

3.4(a)

I choose the hyperparameters that came out with the best results (except the results of sensitivity): C=100 for linear-kernel SVM and C=100, gamma=0.01 for RBF-kernel SVM.

(C)

	linear-kernel SVM score	RBF-kernel SVM score
accuracy	0.7428	0.7571
F1-score	0.4375	0.4516
AUROC	0.6258	0.6360
precision	0.6363	0.7
sensitivity	0.3333	0.3333
specificity	0.9183	0.9387

The two sets of results are almost the same, while the score of RBF-kernel SVM are a little higher. The results of sensitivity make sense, since we choose the hyperparameter set which came out with the worst sensitivity. Here we can conclude that the RBF-kernel SVM model is better since it got the scores 1% to 7% higher comparing to scores from linear-kernel SVM in each metric (except the results of sensitivity).