1. a) $K(\vec{x}, \vec{z})$ is a Kernel function if it satisfies:

① $K(\vec{x}, \vec{z}) = K(\vec{z}, \vec{x})$

② $K(\vec{x}, \vec{z}) = \Phi(\vec{x})^T \Phi(\vec{z})$, for some function $\Phi(\cdot)$

① This property is satisfied because $K(\vec{x}, \vec{z})$ returns the number of unique words in both $\vec{x}$ and $\vec{z}$. The order of the arguments doesn't have an effect on the intersection of these sets, so $K(\vec{x}, \vec{z}) = K(\vec{z}, \vec{x})$. (size of the ∩)

② We define $\Phi(\cdot)$ to be a vector of dimension N, where N is the number of unique words that can occur in the input vector. In other words, $\Phi(\vec{x}): \vec{x} \in \mathbb{R}^M \rightarrow \vec{y} \in \mathbb{R}^N$ and $M \leq N$. $\vec{y} = [y_1\ y_2 \dots y_N]^T$, with each element $y_i \in \{0, 1\}$. If $y_i = 1$, that means the i^{th} possible word exists in $\vec{x}$, and $y_i = 0$ otherwise.

$K(\vec{x}, \vec{z}) = \Phi(\vec{x})^T \Phi(\vec{z})$. Since $\Phi(\vec{x})$ and $\Phi(\vec{z})$ are vectors that indicate which words occur in each vector (with no repeats), only words that occur in both vectors will contribute to the final value from the inner product. That is, $y_{x,i}\, y_{z,i} = 1$ iff the i^{th} possible word occurs in both $\vec{x}$ and $\vec{z}$, and this contributes to the count of unique words that are in both sets. Therefore, $K(\vec{x}, \vec{z}) = \Phi(\vec{x})^T \Phi(\vec{z})$.

b) $\left(1 + \left(\frac{\vec{x}}{\|\vec{x}\|}\right) \cdot \left(\frac{\vec{z}}{\|\vec{z}\|}\right)\right)^3$

$= \left[1 + \vec{x} \cdot \vec{z} \left(\frac{1}{\|\vec{x}\|}\right)\left(\frac{1}{\|\vec{z}\|}\right)\right]^3$

Let $f_1(\vec{x}) = \frac{1}{\|\vec{x}\|}$, then

$= \left[1 + K(\vec{x}, \vec{z}) f_1(\vec{x}) f_1(\vec{z})\right]^3$

By (scaling), we define $K_1(\vec{x}, \vec{z}) = K(\vec{x}, \vec{z}) f_1(\vec{x}) f_1(\vec{z})$ as a valid kernel.

$= \left[1 + K_1(\vec{x}, \vec{z})\right]^3$

$= \left[K_1(\vec{x}, \vec{z})\right]^3 + 3\left[K_1(\vec{x}, \vec{z})\right]^2 + 3\left[K_1(\vec{x}, \vec{z})\right] + 1$

(by binomial expansion)

$\left[K_1(\vec{x}, \vec{z})\right]^3 = \left[K_1(\vec{x}, \vec{z})\right]\left[K_1(\vec{x}, \vec{z})\right]\left[K_1(\vec{x}, \vec{z})\right]$

- This term is a kernel by (product)

$3\left[K_1(\vec{x}, \vec{z})\right]^2 = 3\left[K_1(\vec{x}, \vec{z})\right]\left[K_1(\vec{x}, \vec{z})\right]$

- This term is a kernel by (product) and (scaling)
- let $f(\vec{x}) = \sqrt{3}$

$3\left[K_1(\vec{x}, \vec{z})\right]$

- This term is a kernel by (scaling), let $f(\vec{x}) = \sqrt{3}$

1

- 1 is a kernel function. We can define $K_2(\vec{x}, \vec{z}) = 1$ for all $\vec{x}, \vec{z}$

This satisfies both conditions for a kernel:

① $K_2(\vec{x}, \vec{z}) = K_2(\vec{z}, \vec{x}) = 1$

② let $\Phi(\vec{x}) = [1]$. $K_2(\vec{x}, \vec{z}) = \Phi(\vec{x})^T \Phi(\vec{z}) = [1][1] = 1$

Therefore,

$$[K_1(\vec{x}, \vec{z})]^3 + 3[K_1(\vec{x}, \vec{z})]^2 + 3[K_1(\vec{x}, \vec{z})] + 1$$

is a kernel, and so we have shown

$$\left(1 + \left(\frac{\vec{x}}{\|\vec{x}\|}\right) \cdot \left(\frac{\vec{z}}{\|\vec{x}\|}\right)\right)^3$$

is a kernel.

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

The role of the parameter β is to scale each element of the feature vector $\phi(\vec{x})$. For smaller combinations, the β is raised to a smaller power, and vice versa for larger combinations.

If $0 < \beta < 1$, it can be used to punish larger combinations in the feature vector.

2. a) $\min_{\vec{\theta}} \frac{1}{2}\|\vec{\theta}\|_2^2$ s.t. $y_n \vec{\theta}^T \vec{x}_n \geq 1$, $n = 1, \ldots, N$

in this case → $\vec{\theta}^T \begin{bmatrix} a \\ e \end{bmatrix} + 1 \leq 0$

$= a\theta_1 + e\theta_2 + 1 \leq 0$

$$L(\vec{\theta}, \alpha) = \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha[a\theta_1 + e\theta_2 + 1]$$

$$g(\alpha) = \min_{\vec{\theta}} L(\vec{\theta}, \alpha)$$

$$d^* = \max_{\alpha} g(\alpha) = \max_{\alpha} \min_{\vec{\theta}} \frac{1}{2}[\theta_1^2 + \theta_2^2] + \alpha[a\theta_1 + e\theta_2 + 1]$$

$$\left.\begin{aligned} \frac{\partial L(\vec{\theta}, \alpha)}{\partial \theta_1} &= \theta_1 + \alpha a, \quad \theta_1 = -\alpha a \\ \frac{\partial L(\vec{\theta}, \alpha)}{\partial \theta_2} &= \theta_2 + \alpha e, \quad \theta_2 = -\alpha e \end{aligned}\right\} \vec{\theta}^* = \begin{bmatrix} -\alpha a \\ -\alpha e \end{bmatrix}$$

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

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

$$= \alpha - \frac{1}{2}\alpha^2[a^2 + e^2]$$

$$\frac{dg(\alpha)}{d\alpha} = 1 - \alpha[a^2 + e^2] = 0, \quad \alpha = \frac{1}{a^2 + e^2}$$

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

b) $\min_{\vec{\theta}} \frac{1}{2}\|\vec{\theta}\|_2^2$ s.t. $(+1)(\theta_1+\theta_2) \geq 1$

$(-1)(\theta_1) \geq 1$

$\hookrightarrow \quad 1-\theta_1-\theta_2 \leq 0$

$\theta_1 + 1 \leq 0$

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

$$g(\vec{\alpha}) = \min_{\vec{\theta}} L(\vec{\theta},\vec{\alpha})$$

$$d^* = \max_{\vec{\alpha}} \min_{\vec{\theta}} \frac{1}{2}[\theta_1^2+\theta_2^2] + \alpha_1(1-\theta_1-\theta_2) + \alpha_2(\theta_1+1)$$

$$\frac{\partial L(\vec{\theta},\vec{\alpha})}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0, \quad \theta_1 = \alpha_1 - \alpha_2$$

$$\frac{\partial L(\vec{\theta},\vec{\alpha})}{\partial \theta_2} = \theta_2 - \alpha_1 = 0, \quad \theta_2 = \alpha_1$$

$$g(\vec{\alpha}) = \frac{1}{2}[\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2 + \alpha_1^2] + \alpha_1(1-2\alpha_1+\alpha_2) + \alpha_2(\alpha_1-\alpha_2+1)$$

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

$$\frac{\partial g(\vec{\alpha})}{\partial \alpha_1} = -2\alpha_1 + \alpha_2 + 1 = 0, \quad \alpha_1 = \frac{\alpha_2+1}{2}$$

$$\frac{\partial g(\vec{\alpha})}{\partial \alpha_2} = -\alpha_2 + 1 + \alpha_1, \quad \alpha_2 = \alpha_1 + 1$$

$\alpha_1 = 2, \alpha_2 = 3$

$\theta_1 = -1, \theta_2 = 2, \vec{\theta}^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$

$\gamma = \frac{1}{\|\vec{\theta}\|_2^2} = \frac{1}{\sqrt{5}}$

c) $\min_{\vec{\theta}} \frac{1}{2}\|\vec{\theta}\|_2^2$ s.t $\theta_1 + \theta_2 + b \geq 1$

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

$\Rightarrow (1-b) - \theta_1 - \theta_2 \leq 0$

$\theta_1 + (1+b) \leq 0$

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

$+ \alpha_2([1+b] + \theta_1)$

$g(\vec{\alpha}, b) = \min_{\vec{\theta}} L(\vec{\theta}, \vec{\alpha})$

$\frac{\partial L(\vec{\theta}, \vec{\alpha})}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 \overset{0}{=}, \quad \theta_1 = \alpha_1 - \alpha_2$

$\frac{\partial L(\vec{\theta}, \vec{\alpha})}{\partial \theta_2} = \theta_2 - \alpha_1 = 0, \quad \theta_2 = \alpha_1$

$g(\vec{\alpha}, b) = \frac{1}{2}[\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2 + \alpha_1^2] + \alpha_1([1-b] - 2\alpha_1 + \alpha_2)$

$+ \alpha_2([1+b] + \alpha_1 - \alpha_2)$

$g(\vec{\alpha}, b) = -\alpha_1^2 - \frac{1}{2}\alpha_2^2 + \alpha_1\alpha_2 + \alpha_1[1-b] + \alpha_2[1+b]$

$\frac{\partial g(\vec{\alpha}, b)}{\partial \alpha_1} = -2\alpha_1 + \alpha_2 + 1 - b = 0, \quad \alpha_1 = \frac{\alpha_2 + 1 - b}{2}$

$\frac{\partial g(\vec{\alpha}, b)}{\partial \alpha_2} = -\alpha_2 + \alpha_1 + 1 + b = 0, \quad \alpha_2 = \alpha_1 + 1 + b$

$$\frac{\partial g(\vec{\alpha}, b)}{\partial b} = b(\alpha_2 - \alpha_1) = 0,$$

case $b = 0$: $\vec{\theta^*}$ same as $2b$.

case $\alpha_2 - \alpha_1 = 0$:

$$\alpha_2 = \alpha_1$$

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

$$2\alpha_1 = \alpha_1 + 1 - b$$

$$\alpha_1 = 1 - b$$

$$\alpha_1 = \alpha_1 + 1 + b$$

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

$$\theta_1 = \alpha_1 - \alpha_2 = \emptyset, \quad \theta_2 = \alpha_1 = 2$$

$$\vec{\theta^*} = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \quad b^* = -1 \quad \gamma = \frac{1}{2}$$

w/o offset, $\vec{\theta^*} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$, $\gamma = \frac{1}{\sqrt{5}}$

By having an offset, we were able to achieve a larger margin with our decision boundary

Henry (Ziheng) Yang
204584728

CS 188 - Pset 3 Coding

3.2

[B] It's beneficial to maintain class proportions across folds because each model needs a reasonable amount of data from either class to learn a proper decision boundary. In an extreme case, in one fold it is possible to have no negative labels, so the model would not even be aware of an entire class for classification. Keeping the class proportion across the folds also allows for better comparison by the performance metrics, as we can assume the model are trained on reasonably similar datasets.

C	Accuracy	F1-score	AUROC	precision	sensitivity	specificity
10^{-3}	0.7089	0.8297	0.5	0.7089	1.0	0.0
10^{-2}	0.7107	0.8306	0.5031	0.7102	1.0	0.0063
10^{-1}	0.8060	0.8755	0.7188	0.8357	0.9294	0.5081
10^{0}	0.8146	0.8749	0.7531	0.8562	0.9017	0.6045
10^{1}	0.8182	0.8766	0.7592	0.8595	0.9017	0.6167
10^{2}	0.8182	0.8766	0.7592	0.8595	0.9017	0.6167
Best C	0.8182	0.8766	07592	0.8595	1.0	0.6167

[D] We know that the hyperparameter C is used to determine how much to "punish" larger slack variables. In our CV performance across most performance metrics, a larger C performed better. This can be understood as minimizing the slack variables, thus reducing the margins and classifying more strictly on each sample. Sensitivity worked better with less punishment on the slack variables, which makes sense as sensitivity is a measure of how the true positive / positive, and having more slack allows for all of the samples to be classified closely to their labels -- essentially memorizing instead of generalizing.

3.3

[A] Gamma is the parameter attached to the RBF kernel function, which is equivalent to ½*variance. A large gamma implies a small variance, and a small gamma implies a large variance. If there is a small variance, it means that the support vector involved in the kernel function does not have widespread influence. Essentially, the gamma value controls how much

influence the support vector has in deciding on the class of the input vector, meaning a small gamma value will increase the amount of influence from the support vectors.

[B] The grid search used to determine the best pair of hyperparameters (C, gamma) was essentially a nested for loop that explored every gamma value for each C value. All possible pairings were exhausted in this manner to determine the best combination of hyperparameters, and this was done for each performance metric.

[C] The CV performance is similar to that of the Linear-Kernel SVM. Other than sensitivity, the best pairing of hyperparameters (C, gamma) was (100, 0.01). The rationale is the same as that of the Linear-Kernel SVM. The reason sensitivity had a completely different set of best hyperparameters is that it is a poor metric -- it returns a high score for a model that simply memorizes instead of generalizing. In addition, the high gamma hyperparameter essentially prevents the support vectors from influencing the labels, which contributes to increased "memorization" of the training labels.

metric	score	C	gamma
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	0.1	100.0
specificity	0.6047	100	0.01

3.4

[A] The following hyperparameters were chosen as they produced the best scores in 3.2 and 3.3. In other words, these hyperparameters seemed to tune our model the best given our performance metrics, and we will be using the same performance metrics to evaluate their performance on the test dataset.

	Linear-Kernel SVM	RBF-Kernel SVM
Hyperparameters	C = 100	C = 100, gamma = 0.01

[C] The following performance metrics demonstrate that the RBF-Kernel SVM performed better than the Linear-Kernel SVM. This is seen pretty much across the board on every performance metric. A possible reason for this increase in performance for the RBF-Kernel is that it can correspond to a feature space of infinite dimension.

Metric	Linear-Kernel SVM Score	RBF-Kernel SVM Score
accuracy	0.7473	0.7571
f1-score	0.4375	0.4516
auroc	0.6259	0.6361
precision	0.6364	0.7
sensitivity	0.3333	0.3333
specificity	0.9184	0.9387