Christopher Ngai
40479504

CSM146 PS3

① Kernels

A) Yes, this function is a kernel because if we were to construct a kernel matrix K that contained the number of unique elements in the intersection of two documents for any permutation of documents, we can show that the matrix is positive semi-definite.

Assume $K = \begin{bmatrix} k(x,x) & k(x,z) \\ k(z,x) & k(z,z) \end{bmatrix}$ where $k(x,z)$ represents the # of unique words in both documents x and z

Let $k(x,x) = q$, $q \geq 1$

$k(x,z) = k(z,x) = r$, $r \geq 1$

$k(z,z) = s$, $s \geq 1$

Proof of PSD:

$$(k - \lambda I) = \begin{bmatrix} k(x,x)-\lambda & k(x,z) \\ k(z,x) & k(z,z)-\lambda \end{bmatrix}$$

	$q-\lambda$
s	$qs \mid -s\lambda$
$-\lambda$	$-q\lambda \mid \lambda^2$

$$\det(k-\lambda I) = (k(x,x)-\lambda)(k(z,z)-\lambda) - (k(x,z)\cdot k(z,x))$$
$$= (q-\lambda)(s-\lambda) - r^2$$
$$= \lambda^2 + \lambda(-s-q) + qs - r^2$$

	$-s-q$
$-s$	$s^2 \mid qs$
$-q$	$qs \mid q^2$

$$\lambda = \frac{-(-s-q) \pm \sqrt{(-s-q)^2 - 4(qs-r^2)}}{2}$$

$$= \frac{(s+q) \pm \sqrt{(s^2+2qs+q^2) - 4qs + 4r^2}}{2}$$

$$= \frac{s+q \pm \sqrt{s^2+q^2+4r^2-2qs}}{2}$$

$$\lambda_1 = \frac{s+q+\sqrt{s^2+q^2+4r^2-2qs}}{2}$$

$$\lambda_2 = \frac{s+q-\sqrt{s^2+q^2+4r^2-2qs}}{2}$$

If $q, r, s \geq 1$, then:

$$\lambda_1 = \frac{s+q+\sqrt{s^2+q^2+r^2-2qs}}{2} \geq 0$$

Need to prove $\lambda_2 \geq 0$:

Assume $\lambda_2 \geq 0$:

$$\Rightarrow \frac{s+q-\sqrt{s^2+q^2+r^2-2qs}}{2} \geq 0$$

$$s+q \geq \sqrt{s^2+q^2+r^2-2qs}$$

$$(s+q)^2 \geq s^2+q^2+r^2-2qs$$

$$s^2+2qs+q^2 \geq s^2+q^2+r^2-2qs$$

$$4qs \geq r^2$$

We know $q \geq r$ and $s \geq r$ because the intersection of two different documents will always produce the same or less unique words than two of the same documents. Thus,

$$4qs \geq r^2 \checkmark$$

Subsequently, $\lambda_2 \geq 0$.

Since we proved that the eigenvalues of kernel K are nonnegative, we have shown that K is positive semi-definite. Thus, this function is a kernel.

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

$k(x,z) = x \cdot z$ is a kernel

Scaling $\Rightarrow f(x)\, k(x,z)\, f(z)$ where $f(x) = \frac{1}{\|x\|}$, $f(z) = \frac{1}{\|z\|}$

$\hookrightarrow \frac{1}{\|x\|} k(x,z) \frac{1}{\|z\|}$

$= \frac{x \cdot z}{\|x\|\|z\|} \longrightarrow$ call it $k_1(x,z)$

Let $k_2(x,z) = 1$ be a kernel where the kernel matrix K is an $N \times N$ matrix where $k(x_i, z_j) = 1 \;\; \forall \; i,j \in \{1, \ldots N\}$.
$\Rightarrow$ This is a kernel because its eigenvalues would be ≥ 0 so K is positive semi-definite.

Sum $\Rightarrow$ $k_1(x,z) + k_2(x,z)$

$= \frac{x \cdot z}{\|x\|\|z\|} + 1 \longrightarrow$ call it $k_3(x,z)$

Product $\Rightarrow$ $k_3(x,z)\, k_3(x,z)\, k_3(x,z)$

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

$\Rightarrow$ this is a kernel ✓

c) $x, z \in \mathbb{R}^2$

$$k_\beta(x,z) = (1 + \beta x \cdot z)^3 \quad , \quad \beta > 0$$

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \qquad z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

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

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

$1 + \beta x_1 z_1 + \beta x_2 z_2$

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

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

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

→ we need to find the corresponding feature map $\phi_\beta(\cdot)$ such that $k_\beta = \phi(x)^T \phi(z)$

$$\phi(x) = \begin{pmatrix} 1 \\ \sqrt{3\beta}\, x_1 \\ \sqrt{3\beta}\, x_2 \\ \sqrt{3\beta}\, x_1^2 \\ \sqrt{6}\,\beta x_1 x_2 \\ \sqrt{3\beta}\, x_2^2 \\ \sqrt[3]{\beta^3}\, x_1^3 \\ \sqrt[3]{3\beta^3}\, x_1^2 x_2 \\ \sqrt[3]{3\beta^3}\, x_1 x_2^2 \\ \sqrt[3]{\beta^3}\, x_2^3 \end{pmatrix}$$

The corresponding feature map would be $\phi(x)^T \phi(z) = \Phi_\beta$

The kernel $k_\beta(x,z) = (1+\beta x \cdot z)^3$ is similar to $k(x,z) = (1+x \cdot z)^3$ in the sense that they have similar feature transformation functions where the feature space of $k_\beta(x,z)$, ϕ_β, is just like the feature space of $k(x,z)$ except multiplied by some degree of β in each of the elements of its kernel matrix.

The purpose of the parameter β is to multiply the elements in the feature vector by some degree. For example, some elements are multiplied by $\sqrt[2]{\beta^3}$ whereas others are multiplied by β^2. Depending on the value of β, certain elements will either have a larger or smaller effect on the learning model.

---

(2) SVM

a) $x = (a, e)^T$, $y = -1$

$\rightarrow$ minimize $\frac{1}{2}\|\theta\|^2$, $y_n \theta^T x_n \geq 1$, $n = 1, \ldots, N$

Primal:

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

$$p = \min_\theta \max_\alpha \mathcal{L}(\theta, \alpha)$$

Dual:

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

$$d = \max_\alpha g(\alpha)$$

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

$$\theta = -\alpha(a,e)^T$$

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

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

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

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

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

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

$$= -\alpha(a^2+e^2) + 1 = 0$$

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

$$\Rightarrow \quad \theta = -\alpha(a,e)^T$$

$$\hookrightarrow \boxed{\theta^* = -\frac{1}{a^2+e^2}(a,e)^T}$$

B) $x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

$y_1 = 1$ $y_2 = -1$

$\min_\theta \frac{1}{2}\|\theta\|_2^2$ s.t. $y_n(w^T \phi(x_n) + b) \geq 1$

$\hookrightarrow 1((\theta_1, \theta_2)\begin{pmatrix} 1 \\ 1 \end{pmatrix} + 0) \geq 1$

$\Rightarrow \theta_1 + \theta_2 \geq 1 \rightarrow 1 - \theta_1 - \theta_2 \leq 0$

$\hookrightarrow -1((\theta_1, \theta_2)\begin{pmatrix} 1 \\ 0 \end{pmatrix}) \geq 1$

$-\theta_1 \geq 1$

$\theta_1 \leq -1$ $\theta + 1 \leq 0$

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

Dual problem: $g(\alpha) = \min_\theta \mathcal{L}(\theta, \alpha)$

$\alpha^* = \max_{\alpha_i \geq 0} g(\alpha) = \max_{\alpha_i \geq 0} \min_\theta \mathcal{L}(\theta, \alpha)$

$\min_\theta \mathcal{L}(\theta, \alpha) = \frac{\partial \mathcal{L}(\theta, \alpha)}{\partial \theta}$

$\frac{\partial \mathcal{L}(\theta, \alpha)}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0$ $\frac{\partial \mathcal{L}(\theta, \alpha)}{\partial \theta_2} = \theta_2 - \alpha_1 = 0$

$\theta_1 = \alpha_1 - \alpha_2$ $\theta_2 = \alpha_1$

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

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

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

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

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

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

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

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

$\alpha_2 = \alpha_1 + 1$

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

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

$$2\alpha_1 = \alpha_1 + 2 \qquad \alpha_2 = 3$$

$$\alpha_1 = 2$$

$$\theta_1 = \alpha_1 - \alpha_2 \qquad \theta_2 = \alpha_1$$

$$= 2 - 3 \qquad \theta_2 = 2$$

$$\theta_1 = -1$$

$$\hookrightarrow \quad \theta^* = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

$$\boxed{\theta^* = \begin{pmatrix} -1 \\ 2 \end{pmatrix}}$$

$$\text{margin} = \frac{1}{\|\theta^*\|_2^2} = \frac{1}{\sqrt{(-1)^2 + (2)^2}}$$

$$= \frac{1}{\sqrt{5}} \cdot \frac{\sqrt{5}}{\sqrt{5}} = \frac{\sqrt{5}}{5}$$

$$\boxed{\text{margin} = \frac{\sqrt{5}}{5}}$$

c) $x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $x_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ $b \neq 0$

$y_1 = 1$ $y_2 = -1$

$\min_{\theta} \frac{1}{2}\|\theta\|_2^2$ s.t. $y_n(w^T \varphi(x_n) + b) \geq 1$

$\hookrightarrow 1((\theta_1, \theta_2)\begin{pmatrix} 1 \\ 1 \end{pmatrix} + b) \geq 1$

$\Rightarrow \theta_1 + \theta_2 + b \geq 1 \Rightarrow 1 - \theta_1 - \theta_2 - b \leq 0$

$\hookrightarrow -1((\theta_1, \theta_2)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + b) \geq 1$

$-\theta_1 - b \geq 1$

$\Rightarrow \theta_1 + b \leq -1 \Rightarrow \theta_1 + b + 1 \leq 0$

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

Dual problem $\min_{\theta} \mathcal{L}(\theta, b, \alpha) = \frac{\partial \mathcal{L}(\theta, b, \alpha)}{\partial \theta}$

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

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

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

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

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

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

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

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

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

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

$\alpha_2 = \alpha_1 + 1 + b$

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

$$-\alpha_1 + \alpha_2 = 0$$
$$\alpha_1 = \alpha_2$$

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

$$\alpha_2 = (1-b) + 1 + b$$

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

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

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

$$\theta_1 = \alpha_1 - \alpha_2 \qquad \theta_2 = \alpha_1$$
$$\theta_1 = 2 - 2 \qquad \theta_2 = 2$$
$$\theta_1 = 0$$

$$\hookrightarrow \quad \theta^* = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}$$

$$\boxed{\theta^* = \begin{pmatrix} 0 \\ 2 \end{pmatrix}} \qquad \boxed{b = -1}$$

$$\text{margin} = \frac{1}{\|\theta^*\|_2^2} = \frac{1}{\sqrt{0^2 + 2^2}}$$

$$\boxed{\text{margin} = \frac{1}{2}}$$

Compared to the θ^* and margin without offset, the θ^* vector without offset was a diagonal vector with a positive slope. This produced a margin of $\frac{\sqrt{5}}{5} \approx 0.447$. The θ^* with offset gave us a θ^* vector of $\begin{pmatrix} 0 \\ 2 \end{pmatrix}$ which is a vertically pointing vector that gave us a margin of $\frac{1}{2} = 0.5$, which is greater than the margin of the hyperplane w/o offset. Thus, the θ^* with the biased term b is a better hyperplane.

3 Twitter analysis using SVMs

3.2 Hyperparameter Selection for a Linear-Kernel SVM

3.2b) It is beneficial to maintain class proportions across the folds for cross-validation because sometimes when the proportion of the positive and negative classes are not equal in the training data, the classifier won't have a sufficient amount of training data in order to properly classify the different classes. For example, if there were a lot of positive training examples but only one negative training example, the classifier would produce a valid decision boundary that would correctly classify the training data but it might not do so well for test data. The reason we want to maintain the class proportions is so the classifier can consider a proportional amount of training examples for both classes in each fold.

3.2d)

C	accuracy	F1-score	AUROC	precision	sensitivity	specificity
10^{-3}	0.709	0.830	0.500	0.709	1.000	0.000
10^{-2}	0.711	0.831	0.503	0.710	1.000	0.006
10^{-1}	0.806	0.875	0.719	0.836	0.929	0.508
10^{0}	0.815	0.875	0.753	0.856	0.902	0.605
10^{1}	0.818	0.877	0.759	0.860	0.902	0.617
10^{2}	0.818	0.877	0.759	0.860	0.902	0.617
Best C	0.818	0.877	0.759	0.860	1.000	0.617

For almost all of the performance metrics, with the exception of sensitivity, the 5-fold CV performances gradually increased with its best performance typically being the one with a C value of 10 or 100. For all of the performance metrics, the largest C value of 100 did not provide much of a difference when compared to the C value of 10. Accuracy, AUROC, precision, and specificity all had a relatively sharp increase from 0.001 to 0.1 then tapered off and gradually increased whereas F1-score gradually increased as its C value increased. Finally, sensitivity was the only performance metric that had an inverse relationship between its performance and its C value. As its C value increased, its performance gradually decreased.

3.3 Hyperparameter Selection for an RBF-kernel SVM

3.3a) The role of the additional hyperparameter γ for an RBF-kernel SVM is to determine how much influence a training example has. A small γ value means that a training example will have a farther reach (in terms of radius) or greater influence on the classification of other training data, and a larger γ value means that a training example will have a shorter reach or lesser influence on the classification of other training data. The γ value affects the generalization error because depending on the size of γ, the radius of the area of influence of a training example might be too small where it only includes itself, and thus overfitting will occur. If there's overfitting, the training error will be very low, but the generalization error may be very high.

3.3b) I used a grid in which the C and γ were both [0.001, 0.01, 0.1, 1, 10, 100] because it employed the same C values as I used in Linear SVM Hyperparameter Selection. I also chose the γ values to use these same values because it embodied a wide range of gamma values that would test varying radii of influences and subsequently a wide range of hyperparameter values for RBF SVM.

3.3c)

metric	score	C	γ
accuracy	0.816	10^{2}	10^{-2}
F1-score	0.876	10^{2}	10^{-2}
AUROC	0.754	10^{2}	10^{-2}
precision	0.858	10^{2}	10^{-2}
sensitivity	1.000	10^{-1}	10^{2}
specificity	0.605	10^{2}	10^{-2}

For most of the performance metrics, again with the exception of sensitivity, the CV performance is at its best when the C value is at its highest value and the γ value is at its lowest. On the other hand, sensitivity has its best performance when C is low, but not its lowest, and when γ is at its highest value. Most of the performance metrics produce relatively high performance scores, however specificity did not have the best performance, which is consistent with the linear-kernel SVM. The sensitivity metric gave me the best performance score of 1.000, which is the same as linear-kernel SVM.

3.4 Test Set Performance

3.4a) For the linear-kernel SVM, I chose a C value of 100 because this C value gave me the best performance score for a majority of the performance metrics with the exception of sensitivity. Additionally, for the RBF-kernel SVM, I chose a C value of 100 and a γ value of 0.001 because these hyperparameters also gave me the best performance scores for a majority of the performance metrics with the exception of sensitivity.

3.4c)

metric	Linear-kernel SVM score	RBF-kernel SVM score
accuracy	0.743	0.757
F1-score	0.437	0.452
AUROC	0.626	0.636
precision	0.636	0.700
sensitivity	0.333	0.333
specificity	0.918	0.939

For almost all of the performance metrics, the RBF-kernel SVM score was better than the linear-kernel SVM score by a slight amount. The only exception was the sensitivity in which the two scores were equal to each other. The sensitivity scores were also the lowest out of all the other performance metrics potentially due to overfitting on the training data which led to worse performance when applied to generalized data. On the contrary, the specificity scores for both linear-kernel and RBF-kernel SVM were the highest out of all performance metrics.