Dhruv Chakraborty
UID: 204-962-098
Prof. Sankararaman

CM146: PSET 3

1) Kernels

(a) $k(x,z)$ is a kernel. We can see that it is symmetric since $k(x,z) = k(z,x)$ as the intersection of the words remains the same. Now we show positive semi-definitiness.

set $K = \begin{bmatrix} k(x,x) & k(x,z) \\ k(z,x) & k(z,z) \end{bmatrix}$ $\because k(x,z) = k(z,x)$

$$\det(K - \lambda I) = (k(x,x) - \lambda)(k(z,z) - \lambda) - k(x,z)^2 = 0$$

$$\Rightarrow \lambda^2 - \lambda(k(x,x) + k(z,z)) + k(x,x)k(z,z) - k(x,z)^2 = 0$$

Using quadratic formula to solve for λ,

$$\lambda = \frac{k(x,x) + k(z,z) \pm \sqrt{(k(x,x)+k(z,z))^2 - 4(k(x,x)k(z,z) - k(x,z)^2)}}{2}$$

Clearly, $\frac{k(x,x) + k(z,z) + \sqrt{(k(x,x)+k(z,z))^2 - 4(k(x,x)k(z,z) - k(x,z)^2)}}{2}$

is positive so $\lambda_1 > 0$ (as k is +ve and $\sqrt{x} > 0 \ \forall x \in \mathbb{R}$)

Now considering $\lambda_2 = \frac{k(x,x) + k(z,z) - \sqrt{(k(x,x)+k(z,z))^2 - 4(k(x,x)k(z,z) - k(x,z)^2)}}{2}$

We want $k(x,x) + k(z,z) \geq \sqrt{(k(x,x)+k(z,z))^2 - 4(k(x,x)k(z,z) - k(x,z)^2)}$

$$\Updownarrow$$

$$\cancel{(k(x,x)+k(z,z))^2} \geq \cancel{(k(x,x)+k(z,z))^2} - 4(k(x,x)k(z,z) - k(x,z)^2)$$

$$\Updownarrow$$

$$0 \geq -4(k(x,x)k(z,z) - k(x,z)^2)$$

$$\Updownarrow$$

$$k(x,x)k(z,z) \geq k(x,z)^2$$

But we know this since the intersection of a document with itself is necessarily greater than/equal to a different document i.e. $k(x,x) \geq k(x,z)$ and $k(z,z) \geq k(x,z)$.

Since K is symmetric and has only non-ve eigenvalues it is positive semi-definite.

$\therefore$ $k(x,z)$ is a kernel.

(b) Since $k(x,z) = x \cdot z$ is a kernel, we use scaling ~~to~~ with $f(x) = \frac{1}{\|x\|}$ to get that $\frac{1}{\|x\|} x \cdot z \frac{1}{\|z\|}$ is also a kernel.

Also since $k^*(x,z) = 1$ is clearly a kernel, $1 + \frac{x}{\|x\|} \cdot \frac{z}{\|z\|}$ is a kernel through the sum rule.

Multiplying this kernel by itself thrice and using the product rule, $\underline{\left(1 + \left(\frac{x}{\|x\|}\right) \cdot \left(\frac{z}{\|z\|}\right)\right)^3}$ is a kernel.

(c) $k_\beta(x,z) = (1 + \beta \vec{x} \cdot \vec{z})^3$

$$= \cancel{(1+\beta(x_1 z_1} (1 + \beta(x_1 z_1 + x_2 z_2))^3 \quad \because x, z \in \mathbb{R}^2$$

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

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

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

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

So $\varphi_\beta(x) = \begin{bmatrix} 1 \\ \sqrt{3\beta}\, x_1 \\ \sqrt{3}\beta x_1^2 \\ \beta^{3/2} x_1^3 \\ \sqrt{3\beta}\, x_2 \\ \sqrt{6}\beta x_1 x_2 \\ \sqrt{3}\beta^{3/2} x_1^2 x_2 \\ \sqrt{3}\beta x_2^2 \\ \sqrt{3}\beta^{3/2} x_1 x_2^2 \\ \beta^{3/2} x_2^3 \end{bmatrix}$

→ The differance between $k_\beta(x,z)$ and $k(x,z)$ is that k_β has the scaling factor β that scales the vector. It could potentially be used to regularize ~~the kernel~~

β scales the vector.

2) SVM

(a) Want to minimize $\frac{1}{2}\|\theta\|^2$ with $y_n\theta^T x_n \geq 1$ i.e. $-\theta^T(a,e)^T \geq 1$
$\Rightarrow 1 + a\theta_1 + e\theta_2 \leq 0$.

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

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

Now maximizing over α:

$$\max_\alpha \frac{1}{2}\left\| -\alpha\begin{pmatrix} a \\ e\end{pmatrix}\right\|^2 + \alpha\left(-\alpha\begin{pmatrix} a \\ e\end{pmatrix}(a\ e) + 1\right)$$

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

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

$$\frac{\partial}{\partial \alpha} = -\alpha(a^2+e^2) + 1 = 0 \Rightarrow \alpha^* = \frac{1}{a^2+e^2}$$

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

(b) Given $x_1 = \begin{pmatrix} 1 \\ 1\end{pmatrix}$, $x_2 = \begin{pmatrix} 1 \\ 0\end{pmatrix}$, $y_1 = 1$ and $y_2 = -1$ we must satisfy $\theta_1 + \theta_2 \geq 0$, $\theta_1 \leq 0$ and
$\theta_1 + \theta_2 \geq 1$, $\theta_1 \leq -1$ for $y_n \theta^T x_n \geq 1$.
$\theta_1 = -1$ and $\theta_2 = 2$ satisfy these constraints

$$\therefore \theta^* = \begin{pmatrix} -1 \\ 2 \end{pmatrix} \text{ and } \gamma = \frac{1}{\sqrt{1^2+2^2}} = \frac{1}{\sqrt{5}}$$

(c) The constraints with the offset term are $\theta_1 + \theta_2 \geq 1 - b$ & $\theta_1 \leq 1 - b$
These are satisfied by $\theta = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$ and $b = -1$.

The margin $\gamma = \frac{1}{\sqrt{2^2}} = \frac{1}{2}$, which is larger than the margin without the offset $\left(\frac{1}{\sqrt{5}}\right)$.

2) Twitter Analysis using SVMs

3.2) Hyperparameter Selection for a Linear-Kernel SVM

(b) Since we wish to find hyperparameters that generalize best to unseen data, specifically measured using training/test data, we want the dataset we are testing on to have a similar distribution. A fold that doesn't have the same proportions could lead to overfitting and ~~is an~~ would be an outlier. Moreover, a fold with mostly +ve / -ve samples teaches the model nothing.

(d)

C	accuracy	F1-score	AUROC	precision	sensitivity	specificity
10^{-3}	70.89%	82.97%	50%	70.89%	100%	0%
10^{-2}	71.07%	83.06%	50.31%	71.02%	100%	0.63%
10^{-1}	80.60%	87.55%	71.88%	83.57%	92.94%	50.81%
10^{0}	81.46%	87.49%	75.31%	85.62%	90.17%	60.45%
10^{1}	81.82%	87.66%	75.92%	85.95%	90.17%	61.67%
10^{2}	81.82%	87.66%	75.92%	85.95%	90.17%	61.67%
best C	10	10	10	10	0.001	10

On all performance metrics C=10, ~~100~~ and C=100 give the same value, ~~but~~ which is actually the best for all metrics outside sensitivity. For sensitivity, smaller values like $C=10^{-3}$ and $c=10^{-2}$ perform better, but this is an ~~outl~~ ~~anator~~ outlier and susceptible to overfitting. We use C=10 as our best C.

3.3) Hyperparameter Selection for an RBF-Kernel SVM

(a) The γ parameter "defines how far the influence of a single training example reaches, lower values meaning far and higher values meaning close." (from sklearn docs). Basically as γ increases, ~~the~~ each support vectors ~~the~~ radius of influence decreases. We can tune it to avoid overfitting and increase generalization.

(b) In my grid, both C and γ range from 10^{-3} to 10^{2} in powers of 10. We used these values for C for the linear kernel and they performed quite well. Moreover, this grid allows for a very large range of values for both parameters.

(c)

metric	score	C	γ
accuracy	79.34%	100	0.01
F1-score	87.63%	100	0.01
AUROC	75.45%	100	0.01
precision	85.83%	100	0.01
sensitivity	100%	0.001	0.001
specificity	60.47%	100	0.01

Just as in the linear kernel, sensitivity is an outlier with $C = 0.001$ and $\gamma = 0.001$ performing best for it. However, for every other performance metric, $C = 100$ and $\gamma = 0.01$ are the best hyperparameters in our grid.

3.4) Test Set Performance

(a) We use $C = 10$ for the linear kernel SVM as it performed the best on all metrics ~~at~~ except sensitivity (as did $C = 100$). Similarly, as $C = 100$ & $\gamma = 0.01$ did the best for almost all metrics for the RBF kernel SVM, we use these hyperparameters to train it.

(c)

metric	linear	RBF
accuracy	74.29%	75.71%
F1-score	43.75%	45.16%
AUROC	62.59%	63.61%
precision	63.64%	70.00%
sensitivity	33.33%	33.33%
specificity	91.84%	93.88%

As we can see the RBF kernel classifier performs slightly better than the linear kernel SVM on all metrics (except sensitivity). We would ideally deploy this model over the linear one.