CS M146, Winter 2019
Problem Set 3: SVM and Kernels
Due March 2, 2019

Tian Ye
Collaborators: Derek Chu, Austin Guo

March 2, 2019

1 Kernels

(a) **Solution:** We will define $k(x, z)$ as $k(x, z) = \phi(x)^T \phi(z)$. Since we are counting the number of unique words in both x and z, we will iterate over all the words in the English dictionary and evaluate whether it exists in both x and z. $k(x, z) = k(z, x)$ is very easy to see because the intersection for x and z is the same as the intersection for z and x: ie. it is commutative.

The second condition is $\phi(x)^T \phi(z) = \phi(z)^T \phi(x)$. If $\phi(x)$ is a function that creates a vector with each index representing the entire dictionary of possible words, and an entry is 1 if it is present in the document and 0 if it is not. Therefore the dot product is commutative and tells the union of the two documents of unique words.

Hence this function is a kernel since we can express it as $\phi(x)^T \phi(z)$.

(b) **Solution:**

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

$$= \left(1 + \frac{1}{||x||} \cdot \frac{1}{||z||}(x \cdot z)\right)^3 \tag{2}$$

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

$$\text{Where } f(x) = \frac{1}{||x||} \text{ and } f(z) = \frac{1}{||z||}, \Rightarrow (1 + k_1(x, z))^3 \tag{4}$$

$$= (1 + 2k_1(x, z) + k_1(x, z)^2)(1 + k_1(x, z)) \tag{5}$$

$$= 1 + 3k_1(x, z) + 3k_1(x, z)^2 + k_1(x, z)^3 \tag{6}$$

$$= 1 + 3k_1(x, z) + 3k_1(x, z) \cdot k_1(x, z) + k_1(x, z) \cdot k_1(x, z) \cdot k_1(x, z) \tag{7}$$

$$= 1 + 3k_1(x, z) + 3k_2(x, z) \cdot k_1(x, z) + k_2(x, z) \cdot k_1(x, z) \tag{8}$$

$$= 1 + k_4(x, z) + k_5(x, z) + k_6(x, z) \tag{9}$$

$$\text{If we define } k_7(x, z) = x^0 \cdot z^0 = 1, \tag{10}$$

$$= k_7(x, z) + k_4(x, z) + k_5(x, z) + k_6(x, z) \tag{11}$$

$$= k_8(x, z) + k_9(x, z) \tag{12}$$

$$= k_{10}(x, z) \tag{13}$$

Hence, $(1 + \frac{x}{||x||} \cdot \frac{z}{||z||})^3$ is a kernel.

(c) **Solution:** We know that $k_\beta(x,z) = (1+\beta x \cdot z)^3$ for any $\beta > 0$, where $x \cdot z = x_1 z_1 + x_2 z_2$. From this we yield the following:

$$\begin{aligned}&(1+\beta(x_1z_1+x_2z_2))^3\\&= (1+\beta(x_1z_1+x_2z_2))(1+\beta(x_1z_1+x_2z_2))(1+\beta(x_1z_1+x_2z_2))\end{aligned} \tag{14}$$

$$\begin{aligned}&= (2\beta x_1z_1 + \beta^2x_1^2z_1^2 + 2\beta^2x_1z_1x_2z_2 + 2\beta x_2z_2 + \beta^2x_2^2z_2^2 + 1)\\&\qquad (1+\beta(x_1z_1+x_2z_2))\end{aligned} \tag{15}$$

$$\begin{aligned}&= \beta^3x_1^3z_1^3 + 3\beta^2x_1^2z_1^2 + 3\beta^3x_1^2z_1^2x_2z_2 + 3\beta x_1z_1 + 3\beta^3x_1z_1x_2^2z_2^2 +\\&\qquad 6\beta^2x_1z_1x_2z_2 + \beta^3x_2^3z_2^3 + 3\beta^2x_2^2z_2^2 + 3\beta x_2z_2 + 1\end{aligned} \tag{16}$$

Therefore,

$$\phi_\beta(x) = \begin{bmatrix} 1 \\ \sqrt{3\beta}x_1 \\ \sqrt{3\beta}x_2 \\ \sqrt{3}\beta x_1^2 \\ \sqrt{3}\beta x_2^2 \\ \beta^{\frac{3}{2}}x_1^3 \\ \beta^{\frac{3}{2}}x_2^3 \\ \sqrt{3}\beta^{\frac{3}{2}}x_1x_2^2 \\ \sqrt{3}\beta^{\frac{3}{2}}x_1^2x_2 \\ \sqrt{6}\beta x_1x_2 \end{bmatrix} \tag{17}$$

Since β is not a part of $\phi(x)$, β essentially performs the role of acting as a weight to each term within $\phi_\beta(x)$. This results in the overall kernel function to be scaled by a certain amount since $\phi(x)$ is $\phi_\beta(x)$ without any β terms.

2 SVM

(a) **Solution:** $x = (a, e)^T, y = -1$, $\min \frac{1}{2}||\theta||^2$

$-1 \cdot \theta^T \begin{bmatrix} a \\ e \end{bmatrix} \geq 1 \Rightarrow -\theta_1 a - \theta_2 e \geq 1 \Rightarrow 1 + \theta_1 a + \theta_2 e \geq 0$

Lagrangian: $\mathcal{L}(\theta, \alpha) = \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha(1 + \theta_1 a + -\theta_2 e)$

Dual problem is maximizing α where $g(\alpha) = \min\ \theta\ \mathcal{L}(\theta, \alpha)$

Optimizing θ^*:

$$\frac{\partial \mathcal{L}(\theta, \alpha)}{\partial \theta_1} = \theta_1 + \alpha_a = 0 \tag{18}$$

$$\frac{\partial \mathcal{L}(\theta, \alpha)}{\partial \theta_2} = \theta_2 + \alpha_e = 0 \tag{19}$$

$$\theta_1 = -\alpha a \tag{20}$$

$$\theta_2 = -\alpha e \tag{21}$$

Now, $g(\alpha) = \min\ \theta\ \mathcal{L}(\theta, \alpha) = \mathcal{L}(\theta^*, \alpha)$

$$= \frac{1}{2}(\alpha^2 a^2 + \alpha^2 e^2) - \alpha(1 - \alpha a^2 - \alpha e^2) \tag{22}$$

$$= \frac{1}{2}\alpha^2 a^2 + \frac{1}{2}\alpha^2 e^2 - \alpha + \alpha^2 a^2 + \alpha^2 e^2 \tag{23}$$

$$= \frac{3}{2}\alpha^2 a^2 + \frac{3}{2}\alpha^2 e^2 - \alpha \tag{24}$$

Maximizing g:

$$\frac{\partial g(\alpha)}{\partial \alpha} = 2\alpha a^2 + 2\alpha e^2 + 1 = 0 \tag{25}$$

$$\Rightarrow \alpha(a^2 + e^2) = 1 \tag{26}$$

$$\alpha^* = \frac{1}{(\alpha^2 + e^2)} \tag{27}$$

Hence:

$$\theta^* = \begin{bmatrix} -\frac{a}{(\alpha^2 + e^2)} \\ -\frac{e}{(\alpha^2 + e^2)} \end{bmatrix} \tag{28}$$

(b) **Solution:** $x_1 = (1,1)^T$, $x_2 = (1,0)^T$, $y_1 = 1$, $y_2 = -1$, min θ $\frac{1}{2}||\theta||^2$

$\Rightarrow 1 \cdot \theta^T \begin{bmatrix} 1 \\ 1 \end{bmatrix} \geq 1$ and $-1 \cdot \theta^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} \geq 1$

$\Rightarrow \theta_1 + \theta_2 \geq 1$ and $-\theta_1 \geq 1$

$\Rightarrow 1 - \theta_1 - \theta_2 \geq 0$ and $1 + \theta_1 \geq 0$

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

Dual problem is maximizing $\alpha_1\alpha_2$ where $g(\alpha_1, \alpha_2) = \min\ \theta\ \mathcal{L}(\theta, \alpha_1, \alpha_2)$

Optimizing θ^*:

$$\frac{\partial \mathcal{L}(\theta, \alpha_1, \alpha_2)}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0 \qquad \Rightarrow \theta_1 = \alpha_1 - \alpha_2 \tag{29}$$

$$\frac{\partial \mathcal{L}(\theta, \alpha_1, \alpha_2)}{\partial \theta_2} = \theta_2 - \alpha_1 = 0 \qquad \Rightarrow \theta_2 = \alpha_1 \tag{30}$$

Now we have $g(\alpha_1, \alpha_2) = \min\ \theta\ \mathcal{L}(\theta, \alpha_1, \alpha_2) = \mathcal{L}(\theta^*, \alpha_1, \alpha_2)$

Simplifying:

$$\mathcal{L}(\theta^*, \alpha_1, \alpha_2)$$

$$= \frac{1}{2}(\alpha_1^2 - 2\alpha_1\alpha_2 + \alpha_2^2 + \alpha_1^2) + \alpha_1(1 - \alpha_1 + \alpha_2 - \alpha_1) + \alpha_2(1 + \alpha_1 - \alpha_2) \tag{31}$$

$$= \frac{1}{2}(2\alpha_1^2 - 2\alpha_1\alpha_2 + 2\alpha_2^2) + \alpha_1(1 - 2\alpha_1 + \alpha_2) + \alpha_2(1 + \alpha_1 - \alpha_2) \tag{32}$$

Maximizing g:

$$\frac{\partial g(\alpha_1, \alpha_2)}{\partial \alpha_1} = \frac{1}{2}(4\alpha_1 - 2\alpha_2) + 1 - 4\alpha_1 + \alpha_2 + \alpha_2 \tag{33}$$

$$\Rightarrow 0 = 2\alpha_1 - \alpha_2 + 1 - 4\alpha_1 + \alpha_2 + \alpha_2 \tag{34}$$

$$0 = -2\alpha_1 + \alpha_2 + 1 \tag{35}$$

$$\frac{\partial g(\alpha_1, \alpha_2)}{\partial \alpha_2} = \frac{1}{2}(-2\alpha_1 + 2\alpha_2) + \alpha_1 + 1 + \alpha_1 - 2\alpha_2 \tag{36}$$

$$\Rightarrow 0 = -\alpha_1 + \alpha_2 + \alpha_1 + 1 + \alpha_1 - 2\alpha_2 \tag{37}$$

$$0 = -\alpha_2 + \alpha_1 + 1 \tag{38}$$

Combining (35) and (38):

$$-\alpha_1 + 2 = 0 \Rightarrow \alpha_1 = 2 \tag{39}$$

$$-\alpha_2 + 3 = 0 \Rightarrow \alpha_2 = 3 \tag{40}$$

Therefore $\theta^* = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$ and the margin $\gamma = \frac{1}{\sqrt{(-1^2)+(2)^2}} = \frac{\sqrt{5}}{5}$

(c) **Solution:** $x_1 = (1,1)^T$, $x_2 = (1,0)^T$, $y_1 = 1$, $y_2 = -1$, min θ $\frac{1}{2}||\theta||^2$, $y_n(\theta^T x_n + b) \geq 1$

$\Rightarrow 1 \cdot \theta^T \begin{bmatrix} 1 \\ 1 \end{bmatrix} + b \geq 1$ and $-1 \cdot \theta^T \begin{bmatrix} 1 \\ 0 \end{bmatrix} - b \geq 1$

$\Rightarrow \theta_1 + \theta_2 + b \geq 1$ and $-\theta_1 - b \geq 1$

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

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

Dual problem is maximizing $\alpha_1\alpha_2$ where $g(\alpha_1, \alpha_2) = \min\ \theta\ \mathcal{L}(\theta, \alpha_1, \alpha_2)$

Optimizing θ^*:

$$\frac{\partial \mathcal{L}(\theta, \alpha_1, \alpha_2)}{\partial \theta_1} = \theta_1 - \alpha_1 + \alpha_2 = 0 \qquad \Rightarrow \theta_1 = \alpha_1 - \alpha_2 \tag{41}$$

$$\frac{\partial \mathcal{L}(\theta, \alpha_1, \alpha_2)}{\partial \theta_2} = \theta_2 - \alpha_1 = 0 \qquad \Rightarrow \theta_2 = \alpha_1 \tag{42}$$

Now we have $g(\alpha_1, \alpha_2) = \min\ \theta\ \mathcal{L}(\theta, \alpha_1, \alpha_2) = \mathcal{L}(\theta^*, \alpha_1, \alpha_2)$

Simplifying:

$$\mathcal{L}(\theta^*, \alpha_1, \alpha_2)$$

$$= \frac{1}{2}(\alpha_1^2-2\alpha_1\alpha_2+\alpha_2^2+\alpha_1^2)+\alpha_1(1-\alpha_1+\alpha_2-\alpha_1-b)+\alpha_2(1+\alpha_1-\alpha_2+b) \tag{43}$$

$$= \frac{1}{2}(2\alpha_1^2-2\alpha_1\alpha_2+2\alpha_2^2)+\alpha_1(1-2\alpha_1+\alpha_2-b)+\alpha_2(1+\alpha_1-\alpha_2+b) \tag{44}$$

Maximizing g:

$$\frac{\partial g(\alpha_1, \alpha_2)}{\partial \alpha_1} = \frac{1}{2}(4\alpha_1 - 2\alpha_2) + 1 - 4\alpha_1 + \alpha_2 + \alpha_2 - b \tag{45}$$

$$\Rightarrow 0 = -2\alpha_1 + \alpha_2 + 1 - b \tag{46}$$

$$\Rightarrow \alpha_1 = \frac{\alpha_2 + 1 - b}{2} \tag{47}$$

$$\frac{\partial g(\alpha_1, \alpha_2)}{\partial \alpha_2} = \frac{1}{2}(-2\alpha_1 + 2\alpha_2) + \alpha_1 + 1 + \alpha_1 - 2\alpha_2 + b \tag{48}$$

$$\Rightarrow 0 = -2\alpha_2 + \alpha_1 + 1 + b \tag{49}$$

$$\Rightarrow \alpha_2 = \alpha_1 + 1 + b \tag{50}$$

$$\frac{\partial g(\alpha_1, \alpha_2)}{\partial b} = -\alpha_1 + \alpha_2 \tag{51}$$

$$\Rightarrow \alpha_2 = \alpha_1 \tag{52}$$

Combining (47) (50) and (52), we yield:

$$\alpha_1 = 2, \alpha_2 = 2, b = -1 \tag{53}$$

Therefore,

$$(\theta^*, b^*) = (\begin{bmatrix} 0 \\ 2 \end{bmatrix}, -1) \tag{54}$$

$$\gamma = \frac{1}{\sqrt{0^2 + 2^2}} = \frac{1}{2} \tag{55}$$

3 Twitter Analysis using SVMs

3.1 Feature Extraction

(a) **Solution:** Completed.

(b) **Solution:** Completed.

(c) **Solution:** Completed.

3.2 Hyperparameter Selection for Linear-Kernel SVM

(a) **Solution:** Completed.

(b) **Solution:** We maintain class proportions across folds to have a more even distribution in responses. This prevents data from being skewed as if we did not maintain proportions, we would have one fold that is more skewed than the rest, resulting in inaccurate training and test data sets. This in turn would cause our classifier to be inaccurate as different proportions will naturally result in different predictions. Hence, the proportions are kept constant to help ensure that the classifier can be trained and tested on a representative set of data.

(c) **Solution:** Completed.

(d) **Solution:**

C	Accuracy	F1-Score	AUROC	Precision	Sensitivity	Specificity
10^{-3}	0.7089	0.8297	0.5000	0.7089	1.0000	0.0000
10^{-2}	0.7107	0.8306	0.5031	0.7102	1.0000	0.0063
10^{-1}	0.8060	0.8755	0.7188	0.8357	0.9294	0.5081
10	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	$10^{1}/10^{2}$	$10^{1}/10^{2}$	$10^{1}/10^{2}$	$10^{1}/10^{2}$	10^{-3}	$10^{1}/10^{2}$

As a general trend, the 5-fold CV's performance increases as C increases in value, as all categories with the exception of sensitivity increase as C increases in size. However, sensitivity's performance decreases as C increases due to the fact that C trades off correct classification of training examples to maximize the margin. Hence, the greater C results in more false negatives, reducing the performance of sensitivity while increasing the performance of the remaining categories.

3.3 Hyperparameter Selection for an RBF-kernel SVM

(a) **Solution:** Gamma serves as a marker of the degree of importance that we assign to a given training point. Since the SVM searches for the maximal margin between two different classes, gamma essentially serves as the inverse of hte radius of influence of samples selected by the model as support vectors. Hence, as gamma increases towards infinity, the radius only includes the support vector, resulting in overfitting. As gamma goes to 0, the model becomes too constrained is therefore unable to capture the complexity of the data. A small gamma however allows for good generalization behavior as it appropriately weighs the impact of a given training point. As a result, the gamma in an RBF-kernel SVM allows for good generalization of the problem; however, it does not achieve nearly as good a performance metric as the linear SVM for the training set.

(b) **Solution:** According to scikit-learn.org, it is said that a grid from 10^-3 to 10^3 is usually sufficient. Given that our C has already been tested in that range, I used this grid to test for optimal values of C and γ

(c) **Solution:**

Metric	Score	C	γ
Accuracy	0.8165	100.0	0.01
F1-Score	0.8763	100.0	0.01
AUROC	0.7561	1000.0	0.001
Precision	0.8586	100.0	0.01
Sensitivity	1.0000	0.001	0.001
Specificity	0.6106	1000.0	0.001

In general, the performance is better with high values of C and low values of gamma: with every metric with the exception of sensitivity, C is either 100.0 or 1000.0 and γ is either 0.01 or 0.001. Sensitivity on the other hand perfers small values of C of 0.001. It can also be seen here that the RBF-kernel SVM has worse performance on the training set than the linear SVM.

3.4 Test Set Performance

(a) **Solution:** For the linear SVM, I chose C = 100.0 as 5 out of the 6 metrics had the best performance with a C value of 100.0. For the RBF-kernel SVM, I chose C = 100.0 and $\gamma = 0.01$ as 3 of the 6 metrics had the best performance with these hyperparameters, with two other metrics having their optimal hyperparameters with values similar to those as well.

(b) **Solution:** Completed.

(c) **Solution:**

Metric	Linear SVM Performance	RBF-kernel SVM Performance
Accuracy	0.7429	0.7571
F1-Score	0.4375	0.4516
AUROC	0.6259	0.6361
Precision	0.6364	0.7000
Sensitivity	0.3333	0.3333
Specificity	0.9184	0.9388

It can be seen from this table that RBF-kernel SVM tends to perform overall better than the linear SVM; this can be chalked up to the fact that RBF-kernel SVM has an additional hyperparamterthat permits the definition of the influence of a single hyperparameter, which results in a classifier that is better at generalization. Hence, the RBF-kernel SVM has better performance on the test set than the linear SVM despite its worse performance on the training set.