20W-COMSCIM146 ps3

JACOB KAUFMAN

TOTAL POINTS

41 / 42

QUESTION 1

Kernel 8 pts

1.1 (a) **2 / 2**

✓ **- 0 pts Correct**

- 0.5 pts You need to prove that any kernel matrix is PSD rather than a 2*2 matrix.

- 2 pts your answer is not correct.

1.2 (b) **3 / 3**

✓ **- 0 pts Correct**

- 1 pts The scaling rule is not correctly used.

- 3 pts wrong answer

1.3 **2.5 / 3**

- 0 pts Correct

- 1 pts wrong mapping function

- 1 pts wrong or no comparison

- 1 pts wrong or no argument about beta

- 0.5 pts incomplete comparison

✓ **- 0.5 pts incomplete argument about role of beta**

- 0.5 pts incomplete mapping function

- 3 pts blank

QUESTION 2

SVM 8 pts

2.1 (a) **2 / 2**

✓ **- 0 pts Correct**

- 1 pts wrong or partial final answer

- 2 pts blank

2.2 (b) **3 / 3**

✓ **- 0 pts Correct**

- 1 pts Wrong gamma

- 1 pts wrong theta

- 3 pts not attempted / not found

2.3 (c) **3 / 3**

✓ **- 0 pts Correct**

- 1 pts gamma incorrect

- 1 pts theta incorrect

- 1 pts b incorrect

- 3 pts not attempted / not found

QUESTION 3

26 pts

3.1 **2 / 2**

✓ **- 0 pts Correct**

3.2 **10 / 10**

✓ **- 0 pts Correct**

- 1 pts 3.2b) briefly describe why it might be beneficial to maintain class proportions across folds

- 1 pts How does the 5-fold CV performance vary with C and the performance metric?

- 5 pts table of results

- 10 pts no solution

3.3 **8 / 8**

✓ **- 0 pts Correct**

- 8 pts no answer/wrong

3.4 **5.5 / 6**

- 0 pts Correct/minor mistakes

- 0 pts Sensitivity should be 1 for both cases

- 3 pts Very wrong values perhaps your implementation and settings are not right but hard to say from the results what was done wrong exactly in your code won't be able to explain.

- 0 pts Did not report your best Cs. Or even what

values you tried.

- **1.5 pts** Wrong RBF results

- **1.5 pts** Wrong linear results

- **1 pts** Totally wrong f1 score for both linear and RBF

- **0 pts** Looks like your RBF is wrong. But it's in the ball park.

- **6 pts** No Ans

✓ - **0.5 pts** **RBF accuracy/specificity looks very wrong.**

- **3 pts** No idea, what is the results for each metric of linear model what is the results for RBF.

- **0.5 pts** Some linear/RBF results are also wrong.

- **1 pts** Sensitivity n Specifity looks wrong for both RBF+Linear

CM146, Winter 2020
Problem Set 2: SVM and Kernels
Due March 1, 2020 at 11:59 pm

Jacob Kaufman

03/01/2020

1 Kernels [8pts]

Solution:

(a) We want to find $\mathbf{x}$, $\mathbf{z}$ such that $\phi(\mathbf{x})\phi^T(\mathbf{z}) = k(\mathbf{x}, \mathbf{z})$
Let the. dimension of. $\phi(\cdot)$ be the number of words in the English dictionary ($= N$). Define

$$\phi(\mathbf{x}) = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{pmatrix}$$

to be the mapping function where $\phi_i = 1$ corresponds to the i^{th} word of the English dictionary appearing in the document $\mathbf{x}$, and $\phi_i = 0$ corresponds to the word not appearing in the document. Thus $\phi(\mathbf{x})\phi^T(\mathbf{z})$ represents the size of the intersection of the sets of words in the two documents, or the number of unique words that appear in both documents. So,

$$\phi(\mathbf{x})\phi^T(\mathbf{z}) = k(\mathbf{x}, \mathbf{z})$$

Also, it is clear that $k(\mathbf{x}, \mathbf{z}) = k(\mathbf{z}, \mathbf{x})$ because the number of similar words among $\mathbf{x}$ and $\mathbf{z}$ is equal to the number of similar words among $\mathbf{z}$ and $\mathbf{x}$.

(b) It will be helpful later to know that

$$\phi(\mathbf{x}) = 1 \implies \phi(\mathbf{x})\phi^T(\mathbf{z}) = 1 \implies 1 \text{ is a kernel.}$$

Let $f(\mathbf{x}) = \frac{1}{||\mathbf{x}||}$

1.1 (a) **2 / 2**

✓ **- 0 pts Correct**

- 0.5 pts You need to prove that any kernel matrix is PSD rather than a 2*2 matrix.

- 2 pts your answer is not correct.

CM146, Winter 2020
Problem Set 2: SVM and Kernels
Due March 1, 2020 at 11:59 pm

Jacob Kaufman

03/01/2020

1 Kernels [8pts]

Solution:

(a) We want to find $\mathbf{x}$, $\mathbf{z}$ such that $\phi(\mathbf{x})\phi^T(\mathbf{z}) = k(\mathbf{x}, \mathbf{z})$
Let the. dimension of. $\phi(\cdot)$ be the number of words in the English dictionary ($= N$). Define

$$\phi(\mathbf{x}) = \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_N \end{pmatrix}$$

to be the mapping function where $\phi_i = 1$ corresponds to the i^{th} word of the English dictionary appearing in the document $\mathbf{x}$, and $\phi_i = 0$ corresponds to the word not appearing in the document. Thus $\phi(\mathbf{x})\phi^T(\mathbf{z})$ represents the size of the intersection of the sets of words in the two documents, or the number of unique words that appear in both documents. So,

$$\phi(\mathbf{x})\phi^T(\mathbf{z}) = k(\mathbf{x}, \mathbf{z})$$

Also, it is clear that $k(\mathbf{x}, \mathbf{z}) = k(\mathbf{z}, \mathbf{x})$ because the number of similar words among $\mathbf{x}$ and $\mathbf{z}$ is equal to the number of similar words among $\mathbf{z}$ and $\mathbf{x}$.

(b) It will be helpful later to know that

$$\phi(\mathbf{x}) = 1 \implies \phi(\mathbf{x})\phi^T(\mathbf{z}) = 1 \implies 1 \text{ is a kernel.}$$

Let $f(\mathbf{x}) = \frac{1}{||\mathbf{x}||}$

$$
\begin{aligned}
& k(\mathbf{x}, \mathbf{z}) = \mathbf{x} \cdot \mathbf{z} \text{ is a kernel.} \\
\implies & f(\mathbf{x})(\mathbf{x} \cdot \mathbf{z}) f(\mathbf{z}) \text{ is a kernel} \\
\implies & \left(\frac{1}{||\mathbf{x}||}\right)(\mathbf{x} \cdot \mathbf{z})\left(\frac{1}{||\mathbf{x}||}\right) \text{ is a kernel} \\
\implies & \left(\frac{1}{||\mathbf{x}||}\right)(\mathbf{x} \cdot \mathbf{z})\left(\frac{1}{||\mathbf{x}||}\right) + 1 \text{ is a kernel} \\
\implies & \left(\left(\frac{\mathbf{x}}{||\mathbf{x}||}\right) \cdot \left(\frac{\mathbf{z}}{||\mathbf{z}||}\right) + 1\right) \text{ is a kernel} \\
& \left(\left(\frac{\mathbf{x}}{||\mathbf{x}||}\right) \cdot \left(\frac{\mathbf{z}}{||\mathbf{z}||}\right) + 1\right)^3 \text{ is a kernel}
\end{aligned}
$$

(c) Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$. Let $k_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta \mathbf{x} \cdot \mathbf{z})^3$ for any $\beta > 0$. We want to find $\phi(\cdot)$. We do this by expanding the kernel function and rewriting it as $\phi^T(\mathbf{x})\phi(\mathbf{z})$. We do this below.

$$
\begin{aligned}
(1 + \beta(\mathbf{x} \cdot \mathbf{z}))^3 &= 1 + 3\beta(\mathbf{x} \cdot \mathbf{z}) + 3\beta^2(\mathbf{x} \cdot \mathbf{z})^2 + \beta^3(\mathbf{x} \cdot \mathbf{z})^3 \\
&= 1 + 3\beta(x_1 z_1 + x_2 z_2) + 3\beta^2(x_1 z_1 + x_2 z_2)^2 + \beta^3(x_1 z_1 + x_2 z_2)^3 \\
&= 1 + 3\beta x_1 z_1 + 3\beta x_2 z_2 + \\
&\quad 3\beta^2(x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2) + \\
&\quad \beta^3(x_1^3 z_1^3 + 3x_1^2 z_1^2 x_2 z_2 + 3x_1 z_1 x_2^2 z_2^2 + x_2^2 z_2^2)
\end{aligned}
$$

which is equal to $\phi^T(\mathbf{x})\phi(\mathbf{z})$ when

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

The transformations are equivalent, with the exception that each entry in the vector $\phi_\beta(\cdot)$ will be scaled by some power of β. This is a special

1.2 (b) **3 / 3**

✓ **- 0 pts** **Correct**

- 1 pts The scaling rule is not correctly used.

- 3 pts wrong answer

$$
\begin{aligned}
& k(\mathbf{x}, \mathbf{z}) = \mathbf{x} \cdot \mathbf{z} \text{ is a kernel.} \\
\implies & f(\mathbf{x})(\mathbf{x} \cdot \mathbf{z}) f(\mathbf{z}) \text{ is a kernel} \\
\implies & \left(\frac{1}{||\mathbf{x}||}\right)(\mathbf{x} \cdot \mathbf{z})\left(\frac{1}{||\mathbf{x}||}\right) \text{ is a kernel} \\
\implies & \left(\frac{1}{||\mathbf{x}||}\right)(\mathbf{x} \cdot \mathbf{z})\left(\frac{1}{||\mathbf{x}||}\right) + 1 \text{ is a kernel} \\
\implies & \left(\left(\frac{\mathbf{x}}{||\mathbf{x}||}\right) \cdot \left(\frac{\mathbf{z}}{||\mathbf{z}||}\right) + 1\right) \text{ is a kernel} \\
& \left(\left(\frac{\mathbf{x}}{||\mathbf{x}||}\right) \cdot \left(\frac{\mathbf{z}}{||\mathbf{z}||}\right) + 1\right)^3 \text{ is a kernel}
\end{aligned}
$$

(c) Let $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$. Let $k_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta \mathbf{x} \cdot \mathbf{z})^3$ for any $\beta > 0$. We want to find $\phi(\cdot)$. We do this by expanding the kernel function and rewriting it as $\phi^T(\mathbf{x})\phi(\mathbf{z})$. We do this below.

$$
\begin{aligned}
(1 + \beta(\mathbf{x} \cdot \mathbf{z}))^3 &= 1 + 3\beta(\mathbf{x} \cdot \mathbf{z}) + 3\beta^2(\mathbf{x} \cdot \mathbf{z})^2 + \beta^3(\mathbf{x} \cdot \mathbf{z})^3 \\
&= 1 + 3\beta(x_1 z_1 + x_2 z_2) + 3\beta^2(x_1 z_1 + x_2 z_2)^2 + \beta^3(x_1 z_1 + x_2 z_2)^3 \\
&= 1 + 3\beta x_1 z_1 + 3\beta x_2 z_2 + \\
&\quad 3\beta^2(x_1^2 z_1^2 + 2x_1 z_1 x_2 z_2 + x_2^2 z_2^2) + \\
&\quad \beta^3(x_1^3 z_1^3 + 3x_1^2 z_1^2 x_2 z_2 + 3x_1 z_1 x_2^2 z_2^2 + x_2^2 z_2^2)
\end{aligned}
$$

which is equal to $\phi^T(\mathbf{x})\phi(\mathbf{z})$ when

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

The transformations are equivalent, with the exception that each entry in the vector $\phi_\beta(\cdot)$ will be scaled by some power of β. This is a special

case of $k_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta \mathbf{x} \cdot \mathbf{z})^3$, where $\beta = 1$. As stated before, β scales the coordinates of the vector $\phi_\beta(\mathbf{x})$.

2 SVM [8pts]

Solution:

(a) We are trying to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y\theta^T \mathbf{x} \geq 1$. The constraint can be simplified to

$$(-1)(a\theta_1 + e\theta_2) \geq 1 \implies a\theta_1 + e\theta_2 + 1 \leq 0$$

. We use Lagrange multipliers to solve this problem.

$$\begin{aligned}
\mathcal{L}(\theta, \alpha) &= \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha(1 + a\theta_1 + e\theta_2) \\
\frac{\partial \mathcal{L}}{\partial \theta_1} &= \theta_1 + a\alpha = 0 \implies \theta_1 = -a\alpha \\
&\text{and} \\
\frac{\partial \mathcal{L}}{\partial \theta_2} &= \theta_2 + e\alpha = 0 \implies \theta_2 = -e\alpha \\
\implies \mathcal{L}(\alpha) &= \frac{1}{2}(a^2\alpha^2 + e^2\alpha^2) + \alpha(1 - a^2\alpha - e^2\alpha) \\
\frac{\partial \mathcal{L}}{\partial \alpha} &= \alpha(a^2 + e^2) + 1 - 2a^2\alpha - 2e^2\alpha \\
&= 1 - a^2\alpha - e^2\alpha = 0 \\
&\implies \alpha = \frac{1}{a^2 + e^2} \\
&\implies \theta_1 = -\frac{a}{a^2 + e^2} \\
&\text{and} \\
\theta_2 &= -\frac{e}{a^2 + e^2} \\
&\implies \theta^* = \begin{pmatrix} -\frac{a}{a^2+e^2} \\ -\frac{e}{a^2+e^2} \end{pmatrix}
\end{aligned}$$

(b) In this problem, we aim to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y_1\theta^T x_1 \geq 1$ and $y_2\theta^T x_2 \geq 1$. Thus we want to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that

1.3 **2.5 / 3**

- **0 pts** Correct
- **1 pts** wrong mapping function
- **1 pts** wrong or no comparison
- **1 pts** wrong or no argument about beta
- **0.5 pts** incomplete comparison

✓ **- 0.5 pts incomplete argument about role of beta**

- **0.5 pts** incomplete mapping function
- **3 pts** blank

case of $k_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta \mathbf{x} \cdot \mathbf{z})^3$, where $\beta = 1$. As stated before, β scales the coordinates of the vector $\phi_\beta(\mathbf{x})$.

2 SVM [8pts]

Solution:

(a) We are trying to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y\theta^T \mathbf{x} \geq 1$. The constraint can be simplified to

$$(-1)(a\theta_1 + e\theta_2) \geq 1 \implies a\theta_1 + e\theta_2 + 1 \leq 0$$

. We use Lagrange multipliers to solve this problem.

$$\begin{aligned}
\mathcal{L}(\theta, \alpha) &= \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha(1 + a\theta_1 + e\theta_2) \\
\frac{\partial \mathcal{L}}{\partial \theta_1} &= \theta_1 + a\alpha = 0 \implies \theta_1 = -a\alpha \\
&\text{and} \\
\frac{\partial \mathcal{L}}{\partial \theta_2} &= \theta_2 + e\alpha = 0 \implies \theta_2 = -e\alpha \\
\implies \mathcal{L}(\alpha) &= \frac{1}{2}(a^2\alpha^2 + e^2\alpha^2) + \alpha(1 - a^2\alpha - e^2\alpha) \\
\frac{\partial \mathcal{L}}{\partial \alpha} &= \alpha(a^2 + e^2) + 1 - 2a^2\alpha - 2e^2\alpha \\
&= 1 - a^2\alpha - e^2\alpha = 0 \\
&\implies \alpha = \frac{1}{a^2 + e^2} \\
&\implies \theta_1 = -\frac{a}{a^2 + e^2} \\
&\text{and} \\
\theta_2 &= -\frac{e}{a^2 + e^2} \\
&\implies \theta^* = \begin{pmatrix} -\frac{a}{a^2+e^2} \\ -\frac{e}{a^2+e^2} \end{pmatrix}
\end{aligned}$$

(b) In this problem, we aim to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y_1\theta^T x_1 \geq 1$ and $y_2\theta^T x_2 \geq 1$. Thus we want to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that

2.1 (a) **2 / 2**

✓ **- 0 pts Correct**

- 1 pts wrong or partial final answer

- 2 pts blank

case of $k_\beta(\mathbf{x}, \mathbf{z}) = (1 + \beta \mathbf{x} \cdot \mathbf{z})^3$, where $\beta = 1$. As stated before, β scales the coordinates of the vector $\phi_\beta(\mathbf{x})$.

2 SVM [8pts]

Solution:

(a) We are trying to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y\theta^T \mathbf{x} \geq 1$. The constraint can be simplified to

$$(-1)(a\theta_1 + e\theta_2) \geq 1 \implies a\theta_1 + e\theta_2 + 1 \leq 0$$

. We use Lagrange multipliers to solve this problem.

$$\begin{aligned}
\mathcal{L}(\theta, \alpha) &= \frac{1}{2}(\theta_1^2 + \theta_2^2) + \alpha(1 + a\theta_1 + e\theta_2) \\
\frac{\partial \mathcal{L}}{\partial \theta_1} &= \theta_1 + a\alpha = 0 \implies \theta_1 = -a\alpha \\
&\text{and} \\
\frac{\partial \mathcal{L}}{\partial \theta_2} &= \theta_2 + e\alpha = 0 \implies \theta_2 = -e\alpha \\
\implies \mathcal{L}(\alpha) &= \frac{1}{2}(a^2\alpha^2 + e^2\alpha^2) + \alpha(1 - a^2\alpha - e^2\alpha) \\
\frac{\partial \mathcal{L}}{\partial \alpha} &= \alpha(a^2 + e^2) + 1 - 2a^2\alpha - 2e^2\alpha \\
&= 1 - a^2\alpha - e^2\alpha = 0 \\
&\implies \alpha = \frac{1}{a^2 + e^2} \\
&\implies \theta_1 = -\frac{a}{a^2 + e^2} \\
&\text{and} \\
\theta_2 &= -\frac{e}{a^2 + e^2} \\
&\implies \theta^* = \begin{pmatrix} -\frac{a}{a^2+e^2} \\ -\frac{e}{a^2+e^2} \end{pmatrix}
\end{aligned}$$

(b) In this problem, we aim to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y_1\theta^T x_1 \geq 1$ and $y_2\theta^T x_2 \geq 1$. Thus we want to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that

$\theta_1 + \theta_2 \geq 1$ and $-\theta_1 \geq 1$. Thus we want to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $1-\theta_1-\theta_2 \leq 0$ and $1+\theta_1 \leq 0$. We again use Lagrange multipliers.

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

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

and

$$\frac{\partial \mathcal{L}}{\partial \theta_1} = \theta_2 - \alpha_1 \implies \theta_2 = \alpha_1$$

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

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

$$\frac{\partial \mathcal{L}}{\partial \alpha_1} = (\alpha_1 - \alpha_2) + 1 - 4\alpha_1 + \alpha_2 + \alpha_2 =$$

$$-2\alpha_1 + \alpha_2 + 1 = 0$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_2} = -(\alpha_1 - \alpha_2) + \alpha_1 + 1 + \alpha_1 - 2\alpha_2$$

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

We now have a system of equations as follows:

$$\begin{aligned} -2\alpha_1 + \alpha_2 + 1 &= 0 \\ \alpha_1 - \alpha_2 + 1 &= 0 \end{aligned}$$

This system gives rise to a solution where $\alpha_1 = 2$ and $\alpha_2 = 3$. This yields

$$\theta^* = \begin{pmatrix} \alpha_1 - \alpha_2 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Furthermore, we get

$$\gamma = \frac{1}{||\theta^*||} = \frac{1}{\sqrt{(-1)^2 + 2^2}} = \frac{1}{\sqrt{5}}$$

2.2 (b) **3 / 3**

✓ **- 0 pts Correct**

- 1 pts Wrong gamma

- 1 pts wrong theta

- 3 pts not attempted / not found

(c) We want to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $y_1(\theta^T x_1 + b) \geq 1$ and $y_2(\theta^T x_2 + b) \geq 1$. Thus we want to minimize $\frac{1}{2}(\theta_1^2 + \theta_2^2)$ such that $\theta_1 + \theta_2 + b \geq q$ and $-\theta_1 - b \geq 1$. Thus the constraint is $1 - \theta_1 - \theta_2 - b \leq 0$ and $1 + \theta_1 + b \leq 0$. We again use Lagrange multipliers.

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

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

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

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

The above equations also imply that $\theta_1 = \alpha_1 - \alpha_1 = 0$. Thus,

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

We get

$$\frac{\partial \mathcal{L}}{\partial \alpha_1} = -\alpha_1 + 2 = \implies \alpha_1 = 2 \implies \alpha_2 = 2$$

$$\implies \theta^* = \begin{pmatrix} \alpha_1 - \alpha_2 \\ \alpha_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$$

From the constraints, we have

$$
\begin{aligned}
1 - \theta_1 - \theta_2 - b \leq 0 \text{ and } 1 + \theta_1 + b \leq 0 \\
\implies 1 - 2 - b \leq 0 \text{ and } 1 + b \leq 0 \\
\implies b \leq -1 \text{ and } b \geq -1 \\
\implies b^* = -1
\end{aligned}
$$

Furthermore,

$$\gamma = \frac{1}{||\theta^*||} = \frac{1}{\sqrt{0^2 + 2^2}} = \frac{1}{2}$$

The added nonzero offset made the margin larger ($\frac{1}{2} > \frac{1}{\sqrt{5}}$).

3 Twitter snalysis using SVMs [26pts]

3.1 Feature Extraction [2pts]

Solution:

(a) Implemented in `twitter.py`

(b) Implemented in `twitter.py`

(c) Implemented in `twitter.py`

3.2 Hyperparameter Selection for a Linear-Kernel SVM [10 pts]

Solution:

(a) Implemented in `twitter.py`

(b) It might be beneficial to maintain class proportions across folds because we are trying to optimize hyperparameters for the original training set, and we should have the folds be representative of the orignial training data in order for those hyperparameters to be representative of the original training data.

(c) Implemented in `twitter.py`

2.3 (c) **3 / 3**

✓ **- 0 pts Correct**

- 1 pts gamma incorrect

- 1 pts theta incorrect

- 1 pts b incorrect

- 3 pts not attempted / not found

$$
\begin{aligned}
1 - \theta_1 - \theta_2 - b \leq 0 \text{ and } 1 + \theta_1 + b \leq 0 \\
\implies 1 - 2 - b \leq 0 \text{ and } 1 + b \leq 0 \\
\implies b \leq -1 \text{ and } b \geq -1 \\
\implies b^* = -1
\end{aligned}
$$

Furthermore,

$$\gamma = \frac{1}{||\theta^*||} = \frac{1}{\sqrt{0^2 + 2^2}} = \frac{1}{2}$$

The added nonzero offset made the margin larger ($\frac{1}{2} > \frac{1}{\sqrt{5}}$).

3 Twitter snalysis using SVMs [26pts]

3.1 Feature Extraction [2pts]

Solution:

(a) Implemented in `twitter.py`

(b) Implemented in `twitter.py`

(c) Implemented in `twitter.py`

3.2 Hyperparameter Selection for a Linear-Kernel SVM [10 pts]

Solution:

(a) Implemented in `twitter.py`

(b) It might be beneficial to maintain class proportions across folds because we are trying to optimize hyperparameters for the original training set, and we should have the folds be representative of the orignial training data in order for those hyperparameters to be representative of the original training data.

(c) Implemented in `twitter.py`

3.1 2 / 2

✓ - 0 pts Correct

$$
\begin{aligned}
1 - \theta_1 - \theta_2 - b \leq 0 \text{ and } 1 + \theta_1 + b \leq 0 \\
\implies 1 - 2 - b \leq 0 \text{ and } 1 + b \leq 0 \\
\implies b \leq -1 \text{ and } b \geq -1 \\
\implies b^* = -1
\end{aligned}
$$

Furthermore,

$$\gamma = \frac{1}{||\theta^*||} = \frac{1}{\sqrt{0^2 + 2^2}} = \frac{1}{2}$$

The added nonzero offset made the margin larger ($\frac{1}{2} > \frac{1}{\sqrt{5}}$).

3 Twitter snalysis using SVMs [26pts]

3.1 Feature Extraction [2pts]

Solution:

(a) Implemented in `twitter.py`

(b) Implemented in `twitter.py`

(c) Implemented in `twitter.py`

3.2 Hyperparameter Selection for a Linear-Kernel SVM [10 pts]

Solution:

(a) Implemented in `twitter.py`

(b) It might be beneficial to maintain class proportions across folds because we are trying to optimize hyperparameters for the original training set, and we should have the folds be representative of the orignial training data in order for those hyperparameters to be representative of the original training data.

(c) Implemented in `twitter.py`

(d) The table below shows the C scores for each metric

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}	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^{1}	10^{1}	10^{1}	10^{-3}	10^{1}

The 5-fold CV performance increases as C increases for every metric, except sensitivity, where it decreases. Sensitivity generally reports the highest scores across all C.

3.3 Hyperparameter Selection for an RBF-kernel SVM [8pts]

Solution:

(a) The hyperparameter γ regulates how much nearby support vectors influence each other. Large γ will cause the "radius of influence" of data points to be very small, while a small γ will cause it to be larger. Using this, we can use γ to manipulate the influence each data point has on its neighbors.

(b) I was aiming to use a large range of possible values for both C and γ, so I used a square grid with the same ranges in C and γ as the range for C given in the starter code, as $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$. This would allow me to choose an appropriate pair for (C,γ).

(c) The metrics with corresponding optimal C and γ are below.

metric	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	0.001	0.01
specificity	0.6047	100	0.01

3.2 10 / 10

✓ **- 0 pts Correct**

- 1 pts 3.2b) briefly describe why it might be beneficial to maintain class proportions across folds

- 1 pts How does the 5-fold CV performance vary with C and the performance metric?

- 5 pts table of results

- 10 pts no solution

(d) The table below shows the C scores for each metric

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}	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^{1}	10^{1}	10^{1}	10^{-3}	10^{1}

The 5-fold CV performance increases as C increases for every metric, except sensitivity, where it decreases. Sensitivity generally reports the highest scores across all C.

3.3 Hyperparameter Selection for an RBF-kernel SVM [8pts]

Solution:

(a) The hyperparameter γ regulates how much nearby support vectors influence each other. Large γ will cause the "radius of influence" of data points to be very small, while a small γ will cause it to be larger. Using this, we can use γ to manipulate the influence each data point has on its neighbors.

(b) I was aiming to use a large range of possible values for both C and γ, so I used a square grid with the same ranges in C and γ as the range for C given in the starter code, as $[10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$. This would allow me to choose an appropriate pair for (C,γ).

(c) The metrics with corresponding optimal C and γ are below.

metric	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	0.001	0.01
specificity	0.6047	100	0.01

3.3 **8 / 8**

✓ **- 0 pts** **Correct**

- 8 pts no answer/wrong

3.4 Test Set Performance [6 pts]

Solution:

(a) Based on the performances in 3.2, the best performance for linear kernel SVM was achieved at C=100 (except sensitivity). Based on the performances in 3.3, the best performance for RBF-kernel SVM was achieved at C=100 and $\gamma = 0.01$ (except sensitivity). Clearly, sensitivity is not a good measure for performance in this instance. These values for C and γ were chosen because they consistently showed the best performance across all metrics except sensitivity.

(b) Implemented in `twitter.py`

(c) The metrics with corresponding SVM performances are shown below.

metric	Linear-kernel SVM	RBF-kernel SVM
accuracy	0.7429	0.6143
F1-score	0.4375	0.5091
AUROC	0.6259	0.6293
precision	0.6364	0.4118
sensitivity	0.3333	0.6667
specificity	0.9184	0.5918

The 5-fold CV performance increases as C increases in every metric except sensitivity, where it decreases. Linear-kernel performed better in accuracy, precision, and specificity. RBF-kernel performed better in F1-score, slightly better in AUROC, and better in sensitivity.

3.4 5.5 / 6

- **0 pts** Correct/minor mistakes
- **0 pts** Sensitivity should be 1 for both cases
- **3 pts** Very wrong values perhaps your implementation and settings are not right but hard to say from the results what was done wrong exactly in your code won't be able to explain.
- **0 pts** Did not report your best Cs. Or even what values you tried.
- **1.5 pts** Wrong RBF results
- **1.5 pts** Wrong linear results
- **1 pts** Totally wrong f1 score for both linear and RBF
- **0 pts** Looks like your RBF is wrong. But it's in the ball park.
- **6 pts** No Ans

✓ - **0.5 pts** **RBF accuracy/specificity looks very wrong.**

- **3 pts** No idea, what is the results for each metric of linear model what is the results for RBF.
- **0.5 pts** Some linear/RBF results are also wrong.
- **1 pts** Sensitivity n Specifity looks wrong for both RBF+Linear