

hw5-Solution

● Graded

Student

Chiang Yi Jie

Total Points

216 / 260 pts

Question 1

Problem 1

20 / 20 pts

✓ + 20 pts Correct answer with reasonable solution.

+ 10 pts Correct answer but lack some description of how you derive it.

+ 0 pts Wrong Answer or blank.

- 2 pts Wrong page select or answer typo.

Question 2

Problem 2

12 / 20 pts

✓ + 5 pts Correct formulation of the Lagrangian.

✓ + 7 pts Clear derivation steps for solving the Lagrangian.

+ 8 pts Correct formulation of the dual problem.

+ 4 pts Wrong formulation of the dual problem but clear derivation steps.

+ 5 pts Effort points

- 3 pts Minor mistake.

+ 0 pts Totally wrong or blank.

💬 The derived final dual problem is not correct.

Question 3

Problem 3

14 / 20 pts

Primal solution to the given uneven margin SVM

✓ + 7 pts Both w_{1126}^* and b_{1126}^* are correct.

+ 4 pts w_{1126}^* is correct.

+ 4 pts b_{1126}^* is correct.

+ 0 pts Both w_{1126}^* and b_{1126}^* are wrong or not provided.

Systematic Process for Solution Conversion

✓ + 7 pts The systematic process is correct and clear.

+ 5 pts The systematic process is generally correct but unclear or has some minor mistakes.

+ 3 pts The systematic process has caught some important points (like scaling or moving the line on the plane) but is generally wrong.

+ 0 pts The systematic process is wrong or not provided.

Explanation / Proof of Correctness

+ 6 pts The proof or explanation is reasonable and clear.

+ 4 pts The proof or explanation is generally correct but lack of details or has some mistakes.

+ 3 pts The explanation of the correctness has only been vaguely mentioned in the systematic process.

✓ + 0 pts The proof or explanation is wrong or not provided.

Question 4

Problem 4

5 / 20 pts

+ 20 pts Correct.

+ 15 pts Generally correct, but has some minor logical errors, or lacks clear description.

+ 10 pts On the right path, but has some major logical problems.

✓ + 5 pts Wrong answer but deserve some effort points.

- 2 pts Minor mistakes.

+ 0 pts Totally wrong or blank.

Question 5

Problem 5

20 / 20 pts

✓ + 20 pts Correct proof with complete proof.

+ 15 pts Correct proof but there's something missing in the proof.

+ 0 pts Wrong answer.

– 2 pts Wrong page selection.

Question 6

Problem 6

20 / 20 pts

✓ + 20 pts Correct answer with complete derivation.

+ 5 pts Some reasonable efforts but fundamentally wrong.

+ 0 pts No solution or wrong answer.

– 2 pts Wrong page selection.

Question 7

Problem 7

20 / 20 pts

✓ + 20 pts Correct answer

+ 18 pts Answer with typo or minor mistake

+ 10 pts Answer with mistake or lack of details

+ 0 pts No or Wrong Answer

Question 8

Problem 8

20 / 20 pts

✓ + 20 pts Correct.

+ 15 pts There exist some errors.

+ 5 pts Wrong answer but with some reasonable efforts.

+ 0 pts Wrong answer

Question 9

Problem 9

20 / 20 pts

✓ + 20 pts Correct answer

+ 19 pts Answer with typo

+ 15 pts Answer with minor mistake (e.g. number of SV)

+ 0 pts Wrong or no answer

Question 10

Problem 10

20 / 20 pts

✓ + 20 pts Correct C

+ 0 pts Wrong or no answer

Question 11

Problem 11

20 / 20 pts

✓ + 20 pts Correct.

+ 17 pts Showing that the frequency of $C=1$ is the highest, even surpassing that of $C=10$.

+ 12 pts Incorrect at multiple C 's frequencies, but you still manifest similar values to the answer at other C s.

+ 8 pts With some reasonable efforts.

+ 0 pts Wrong answer.

Question 12

Problem 12

20 / 20 pts

✓ + 20 pts Correct answer

+ 15 pts line chart's axis meaning is ambiguous.

+ 12 pts There are some errors in your line chart.

+ 5 pts With reasonable efforts,

+ 0 pts Click here to replace this description.

+ 0 pts Wrong answer

Question 13

Problem 13

5 / 20 pts

- 20 pts Wrong answer

- 5 pts A minor logical flaw exists.

- 10 pts A logical flaw exists.

- 15 pts A major logical flaw exists.

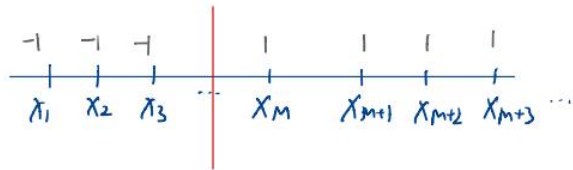
✓ - 15 pts Without rigorous explanation.

- 0 pts Totally correct.

Question assigned to the following page: [1](#)

1. (20 points) Consider N “linearly separable” 1D examples $\{(x_n, y_n)\}_{n=1}^N$. That is, $x_n \in \mathbb{R}$. Without loss of generality, assume that $x_1 \leq x_2 \leq \dots \leq x_M < x_{M+1} \leq x_{M+2} \leq \dots \leq x_N$, $y_n = -1$ for $n = 1, 2, \dots, M$, and $y_n = +1$ for $n = M+1, M+2, \dots, N$. Apply hard-margin SVM without transform on this data set. What is the largest margin achieved in terms of $\{(x_n, y_n)\}_{n=1}^N$ and M ? Prove your answer.

(Note: This is a “hard-margin” decision stump. Have we mentioned that a decision stump is just a 1D perceptron, and the hard-margin SVM is an extension of the perceptron model? :-))



$$\text{margin}(w) = \min_{n=1 \sim N} \text{distance}(x_n, w)$$

$$\frac{x_{M+1} - x_M}{2} \quad \#$$

Question assigned to the following page: [2](#)

2. (20 points) In some situations, we expect to achieve a smaller margin for the positive examples and a larger margin for the negative examples. For instance, when there are very few negative examples and a lot more positive examples, giving the negative examples a bigger margin could be more robust. Consider an *uneven-margin* SVM that solves

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{subject to} \quad & (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } y_n = +1 \\ & -(\mathbf{w}^T \mathbf{x}_n + b) \geq \rho \text{ for } y_n = -1. \end{aligned}$$

Derive the (Lagrange) dual problem for the uneven margin SVM above. List your derivation steps.

(Note: This is a very useful extension of SVM, and more importantly, "forces" you to go through the derivation steps of the class slides to gain a deeper understanding of how to derive a dual problem.)

putting two constraints into one

$$\begin{aligned} y_n (\mathbf{w}^T \mathbf{x}_n + b) &\geq 1 \\ y_n (\mathbf{w}^T \mathbf{x}_n + b) &\geq \rho \end{aligned} \rightarrow \begin{aligned} y_n (\mathbf{w}^T \mathbf{x}_n + b) + \mathbb{I}_{y=-1} (1-\rho) &\geq 1 \\ y_n [\mathbf{w}^T \mathbf{x}_n + b + \mathbb{I}_{y=-1} (\rho-1)] &\geq 1 \end{aligned}$$

$$\mathcal{L}(b, \mathbf{w}, \alpha) = \frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{n=1}^N \alpha_n (1 - y_n (\mathbf{w}^T \mathbf{x}_n + b + \mathbb{I}_{y=-1} (\rho-1)))$$

- $\max_{\alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \alpha) = \infty$, if $y_n (\mathbf{w}^T \mathbf{x}_n + b + \mathbb{I}_{y=-1} (\rho-1)) \leq 1$
- $\max_{\alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \alpha) = \text{constant}$ if $y_n (\mathbf{w}^T \mathbf{x}_n + b + \mathbb{I}_{y=-1} (\rho-1)) \geq 1$

$$\min_{b, \mathbf{w}} (\max_{\alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \alpha)) \geq \max_{\alpha_n \geq 0} (\min_{b, \mathbf{w}} \mathcal{L}(b, \mathbf{w}, \alpha))$$

$$\min_{b, \mathbf{w}} (\max_{\alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \alpha)) \geq \max_{\alpha_n \geq 0} \left(\min_{b, \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{w} + \sum_{n=1}^N \alpha_n (1 - y_n (\mathbf{w}^T \mathbf{x}_n + b + \mathbb{I}_{y=-1} (\rho-1))) \right) \right)$$

$\min_{b, \mathbf{w}} (\mathcal{L}(b, \mathbf{w}, \alpha))$ happens when $\frac{\partial \mathcal{L}}{\partial b} = 0$ and $\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0$

- $\frac{\partial \mathcal{L}}{\partial b} = 0 = \sum_{n=1}^N -y_n \alpha_n$

- $\frac{\partial \mathcal{L}}{\partial \mathbf{w}_i} = 0 = \mathbf{w}_i - \sum_{n=1}^N \alpha_n y_n \mathbf{x}_{n,i}$, $\mathbf{w}_i = \sum_{n=1}^N \alpha_n y_n \mathbf{x}_{n,i}$

$$\begin{aligned} \max_{\substack{\alpha_n \geq 0 \\ \sum y_n \alpha_n = 0 \\ \mathbf{w} = \sum \alpha_n y_n \mathbf{x}_n}} & -\frac{1}{2} \left\| \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n \right\|^2 + \sum_{n=1}^N (\alpha_n - y_n b + \mathbb{I}_{y=-1} (\rho-1)) \end{aligned}$$

Question assigned to the following page: [3](#)

3. (20 points) Let's call the primal solution to the uneven margin SVM above $(b_\rho^*, \mathbf{w}_\rho^*)$. Assume that Dr. Even has solved the original even-margin SVM

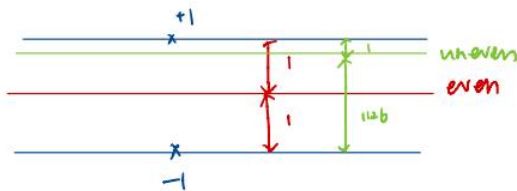
$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{subject to} \quad & (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } y_n = +1 \\ & -(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } y_n = -1. \end{aligned}$$

That is, Dr. Even obtained (b_1^*, \mathbf{w}_1^*) . Somehow zir boss likes the magic number of 1126 better and wants zir to return $(b_{1126}^*, \mathbf{w}_{1126}^*)$ instead. Describe a systematic process to convert (b_1^*, \mathbf{w}_1^*) to $(b_{1126}^*, \mathbf{w}_{1126}^*)$ *without* solving the QP problem again. Prove the correctness of your process.

(Note: The result should reinforce your geometric sense of what SVM is doing.)

$$\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad \text{subject to } y_n (\mathbf{w}^T \mathbf{x}_n + b) \geq 1$$

$$\min_{\mathbf{w}, b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \quad \text{subject to } \begin{cases} (\mathbf{w}^T \mathbf{x}_n + b) \geq 1 & , y = +1 \\ -(\mathbf{w}^T \mathbf{x}_n + b) \geq 1126 & , y = -1 \end{cases}$$



$$w_{\text{uneven}} = (x^+ - x^-) \cdot \frac{w_{1126}}{\|w_{1126}\|}$$

$$w_{\text{even}} = (x^+ - x^-) \cdot \frac{w}{\|w\|}$$

$$= \frac{1}{\|w\|} (x^+ w - x^- w)$$

$$= \frac{1}{\|w\|} ((1-b) - (-1-b))$$

$$= \frac{2}{\|w\|}$$

$$= \frac{1}{\|w_{1126}\|} (w_{1126} x^+ - w_{1126} x^-)$$

$$= \frac{1}{\|w_{1126}\|} ((1-b) - (-1126-b))$$

$$= \frac{1127}{\|w_{1126}\|}$$

$$\text{Since } \frac{2}{\|w_1\|} = \frac{1127}{\|w_{1126}\|}, \quad \|w_{1126}\| = \frac{1127}{2} \|w_1\| \quad \#$$

$$w_1 x^+ + b_1 = w_{1126} x^+ + b_{1126} = 1$$

$$\Rightarrow b_{1126} = b_1 + (w_1 - w_{1126}) x^+$$

$$= b_1 + -\frac{1125}{2} w_1 x^+ = b_1 - \frac{1125}{2} (1-b_1) = \frac{1127}{2} b_1 - \frac{1125}{2} \quad \#$$

Question assigned to the following page: [4](#)

4. (20 points) Let's call the dual solution to the uneven margin SVM above α_p^* . Assume that Dr. Even has solved the dual problem of the even-margin SVM and obtained α_1^* . Prove or disprove that α_1^* is also an optimal solution of the uneven margin SVM above for any $\rho > 0$.

(Note: Think about how your geometric sense from the previous problem can be extended to algebraic results.)

$$\rho > 0$$

$$* \min_{w, b} \left(\max_{\alpha_n > 0} \mathcal{L}(w, b, \alpha_n) \right) \geq \max_{\alpha_n > 0} \left(\min_{b, w} \left(\frac{1}{2} w^T w + \sum_{n=1}^N \alpha_n (1 - y_n (w^T x_n + b)) + \mathbb{I}[y = -1] (\rho - 1) \right) \right)$$

$$\min(\mathcal{L}_\rho(b, w, \alpha)) \text{ happens when } \frac{\partial \mathcal{L}_\rho}{\partial b} = 0 \text{ and } \frac{\partial \mathcal{L}_\rho}{\partial w} = 0$$

$$\bullet \frac{\partial \mathcal{L}_\rho}{\partial b} = 0 = \sum_{n=1}^N -y_n \alpha_n$$

$$\bullet \frac{\partial \mathcal{L}_\rho}{\partial w_i} = 0 = w_{\rho, i} - \sum_{n=1}^N \alpha_n y_n x_{n, i}, \quad w_{\rho, i} = \sum_{n=1}^N \alpha_n y_n x_{n, i}$$

$$\rho = 1$$

$$* \mathcal{L}(b, w, \alpha) = \frac{1}{2} w^T w + \sum_{n=1}^N \alpha_n (1 - y_n (w^T x_n + b))$$

$$\frac{\partial \mathcal{L}_1}{\partial b_1} = 0 = \sum_{n=1}^N -y_n \alpha_n$$

$$\frac{\partial \mathcal{L}_1}{\partial w_{1, i}} = 0, \quad w_{1, i} = \sum_{n=1}^N \alpha_n y_n x_{n, i}$$

→ Both $\frac{\partial \mathcal{L}}{\partial b}$, $\frac{\partial \mathcal{L}}{\partial w_i}$ are dependent of ρ , The optimal solution α_p^* is also the optimal solution for α_1^*

Question assigned to the following page: [5](#)

5. (20 points) For two valid kernels $K_1(\mathbf{x}, \mathbf{x}') = \phi_1(\mathbf{x})^T \phi_1(\mathbf{x}')$ and $K_2(\mathbf{x}, \mathbf{x}') = \phi_2(\mathbf{x})^T \phi_2(\mathbf{x}')$, consider a kernel function $K(\mathbf{x}, \mathbf{x}') = K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}')$. Prove that K is a valid kernel by deriving a transform function $\phi(\mathbf{x})$ such that

$$K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}').$$

(Note: The result shows that a multiplication of valid kernels is still a valid kernel. So, for instance, a multi-variate Gaussian kernel is a valid kernel; a polynomial kernel—repeated multiplications of the identity ϕ —is a valid kernel.)

$$\begin{aligned}
 K(\mathbf{x}, \mathbf{x}') &= K_1(\mathbf{x}, \mathbf{x}') \cdot K_2(\mathbf{x}, \mathbf{x}') \\
 &= \phi_1(\mathbf{x})^T \phi_1(\mathbf{x}') \cdot \phi_2(\mathbf{x})^T \phi_2(\mathbf{x}') \\
 &= \sum_{i=1}^N \phi_{1,i}(\mathbf{x}) \phi_{1,i}(\mathbf{x}') \cdot \sum_{j=1}^M \phi_{2,j}(\mathbf{x}) \phi_{2,j}(\mathbf{x}') \\
 &= \sum_{i=1}^N \sum_{j=1}^M (\phi_{1,i}(\mathbf{x}) \phi_{2,j}(\mathbf{x})) \cdot (\phi_{1,i}(\mathbf{x}') \phi_{2,j}(\mathbf{x}')) \\
 \phi_{3,k} &= \phi_{1,i} \cdot \phi_{2,j} \quad \left[\begin{aligned} &= \sum_{i=1}^N \sum_{j=1}^M \phi_{3,k}(\mathbf{x}) \cdot \phi_{3,k}(\mathbf{x}') \\ &= \phi_3(\mathbf{x})^T \phi_3(\mathbf{x}') \\ &= K_3(\mathbf{x}, \mathbf{x}') \end{aligned} \right.
 \end{aligned}$$

Question assigned to the following page: [6](#)

6. (20 points) For any feature transform ϕ from \mathcal{X} to \mathcal{Z} , the \mathcal{Z} -space distance between two examples \mathbf{x} and \mathbf{x}' is $\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|$, which can be computed with the kernel trick. Consider the degree-2 quadratic kernel $K_2(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T \mathbf{x}')^2$. For two unit vectors \mathbf{x} and \mathbf{x}' , what are their largest and smallest possible distances in the \mathcal{Z} space? Prove your claim.

(Note: This shall sharpen your understanding of the kernel trick.)

$$K_2(\mathbf{x}, \mathbf{x}') = 1 + \sqrt{2}x_1 + \sqrt{2}x_2 \dots \sqrt{2}x_d + x_1^2 + x_2^2 \dots x_d^2$$

$$\|\mathbf{x}\| = 1$$

$$\|\phi(\mathbf{x})\| = \sqrt{\phi(\mathbf{x})^T \phi(\mathbf{x})} = \sqrt{K_2(\mathbf{x}, \mathbf{x})} = \sqrt{(1 + \mathbf{x}^T \mathbf{x})^2} = 2$$

max $K_2(\mathbf{x}, \mathbf{x}')$ happens when \mathbf{x}, \mathbf{x}' in same direction $K_2(\mathbf{x}, \mathbf{x}') = (1 + 1)^2 = 4$
 min $K_2(\mathbf{x}, \mathbf{x}')$ happens when \mathbf{x}, \mathbf{x}' in opposite direction $K_2(\mathbf{x}, \mathbf{x}') = (1 - 1)^2 = 0$

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\| = \sqrt{\phi(\mathbf{x})^T \phi(\mathbf{x}') - 2\phi(\mathbf{x})^T \phi(\mathbf{x}')}$$

$$\max \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\| = \sqrt{4 + 4 - 2 \times 0} = 2\sqrt{2} \quad \#$$

$$\min \|\phi(\mathbf{x}) - \phi(\mathbf{x}')\| = \sqrt{4 + 4 - 2 \times 4} = 0 \quad \#$$

Question assigned to the following page: [Z](#)

7. (20 points) Recall that in the derivation of (one-dimensional) Gaussian kernel, we derived its associated transform

$$\phi(x) = \exp(-x^2) \cdot \left(1, \sqrt{\frac{2}{1!}}x, \sqrt{\frac{2^2}{2!}}x^2, \dots\right).$$

Let

$$\tilde{\phi}(x) = \left(1, \sqrt{\frac{2}{1!}}x, \sqrt{\frac{2^2}{2!}}x^2, \dots\right).$$

Prove that $\exp(-x^2) = \frac{1}{\|\tilde{\phi}(x)\|}$.

(Note: The result means $\phi(x)$ can be viewed as a normalized version of $\tilde{\phi}(x)$.)

$$\|\tilde{\phi}(x)\|^2 = \sum_{n=0}^{\infty} \frac{2^n}{n!} x^{2n} = \sum_{n=0}^{\infty} \frac{(\sqrt{2}x)^{2n}}{n!} = \sum_{n=0}^{\infty} \frac{(2x^2)^n}{n!}$$

$$\|\tilde{\phi}(x)\| = \sum_{n=0}^{\infty} \frac{(x^2)^n}{n!}$$

$$\downarrow \text{ Taylor series expansion of } e^{x^2} = \sum_{n=0}^{\infty} \frac{(x^2)^n}{n!}$$

$$\exp(x^2) = \frac{1}{\|\tilde{\phi}(x)\|} \quad \#$$

Question assigned to the following page: [8](#)

8. (20 points) Let $\cos(\mathbf{x}, \mathbf{x}')$ measure the cosine of the angle between two non-zero vectors \mathbf{x} and \mathbf{x}' in \mathbb{R}^d . The function \cos is typically called the cosine similarity between two vectors. Prove or disprove that $\cos(\mathbf{x}, \mathbf{x}')$ is a valid kernel on $\mathcal{X} = \mathbb{R}^d - \{0\}$.

(Hint: to prove, you'd better construct its associated transform; to disprove, you may use Mercer's condition to construct a counter-example on positive semi-definiteness.)

$$K_{i,j} = \cos(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \cdot \frac{\mathbf{x}_j}{\|\mathbf{x}_j\|} = \phi_i(\mathbf{x}) \cdot \phi_j(\mathbf{x})$$

By considering a vector $C = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \end{bmatrix}$

$$\cos(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^T}{\|\mathbf{x}\|} \cdot \frac{\mathbf{x}'}{\|\mathbf{x}'\|} = \frac{\mathbf{x}'}{\|\mathbf{x}'\|} \cdot \frac{\mathbf{x}^T}{\|\mathbf{x}\|} = \cos(\mathbf{x}', \mathbf{x}) \rightarrow \text{symmetric}$$

$$C^T K C = \sum_{i=1}^M \sum_{j=1}^N c_i c_j \cos(\mathbf{x}_i, \mathbf{x}_j)$$

$$= \sum_{i=1}^M \sum_{j=1}^N \frac{c_i \cdot c_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} \mathbf{x}_i \cdot \mathbf{x}_j$$

$$= \sum_{i=1}^M \sum_{j=1}^N \left(\frac{c_i}{\|\mathbf{x}_i\|} \cdot \mathbf{x}_i \cdot \frac{c_j}{\|\mathbf{x}_j\|} \mathbf{x}_j \right)$$

$$= \left| \sum_{i=1}^M \frac{c_i}{\|\mathbf{x}_i\|} \mathbf{x}_i \right|^2 \geq 0 \rightarrow K \text{ is positive semi-definite}$$

$\rightarrow \cos(\mathbf{x}, \mathbf{x}')$ is valid kernel.

Question assigned to the following page: [9](#)

9. (20 points, *) Consider the polynomial kernel $K(\mathbf{x}_n, \mathbf{x}_m) = (1 + \mathbf{x}_n^T \mathbf{x}_m)^Q$, where Q is the degree of the polynomial. Consider a binary classifier of “4” versus “not 4”. Among $C \in \{0.1, 1, 10\}$ and $Q \in \{2, 3, 4\}$, count the number of support vectors of the resulting SVM classifier. Which of the (C, Q) combination(s) result in the smallest number of support vectors? Provide your command/code.

最小支持向量数量: 629

对应的 (C, Q) 组合: (10, 4)

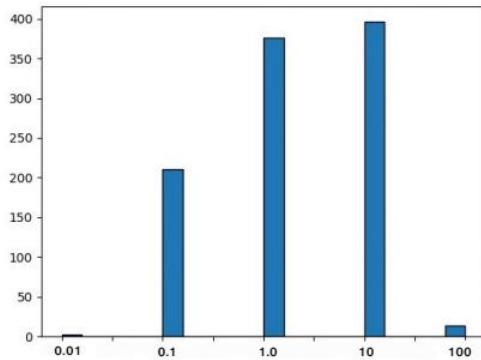
Question assigned to the following page: [10](#)

10. (20 points, *) Consider the Gaussian kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2)$. For the binary classification problem of “1” versus “not 1”, when fixing $\gamma = 1$, among $C \in \{0.01, 0.1, 1, 10, 100\}$, which value(s) of C result in the lowest E_{out} ? Provide your command/code.

```
最小 Eout: 0.5%  
对应的 C 值: 1
```


Question assigned to the following page: [11](#)

11. (*) Following Problem 10 and consider a validation procedure that randomly samples 200 examples from the training set for validation and leaves the other examples for training g_{SVM}^- . Fix $\gamma = 1$ and use the validation procedure to choose the best C among $\{0.01, 0.1, 1, 10, 100\}$ according to E_{val} . If there is a tie of E_{val} , choose the smallest C . Repeat the procedure 1000 times and count the number of times that each C is selected. Plot a bar chart of C versus its selection frequency. Compare the result to that of Problem 10. Describe your findings. Provide your command/code.

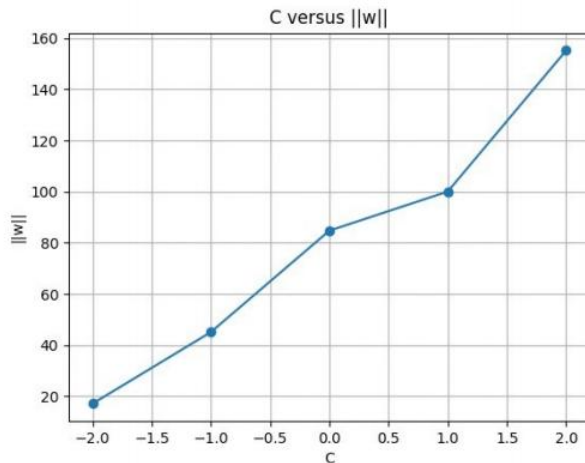


We repeat the process 1000 times and calculate every best c for Eval, the most frequent best c would be 10, while 0.01 appears less than 10 times. Since every time x_{train} and x_{val} are different set, we get different solutions.

In comparison in 5. 10, x_{train} and x_{test} is fixed, the best c would never change, best $c=1$.

Question assigned to the following page: [12](#)

12. (20 points, *) Consider the Gaussian kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2)$. For the binary classification problem of “3” versus “not 3”, when fixing $\gamma = 1$, vary $C \in \{0.01, 0.1, 1, 10, 100\}$ and calculate $\|\mathbf{w}\|$ of the resulting SVM classifier with the kernel trick. Plot a line chart of C versus $\|\mathbf{w}\|$. Describe your findings. Provide your command/code.



In the plot, when c gets larger, $\|\mathbf{w}\|$ gets larger.
when c gets smaller, $\|\mathbf{w}\|$ gets smaller.

$\|\mathbf{w}\|$ increase, since c indicates the penalty when wrong data occur, $\|\mathbf{w}\|$ indicates the dimension of \mathbf{w} , if we want less wrong data, \mathbf{w} must be in higher dimension, and might cause overfitting.

When C gets lower, the penalty on wrong data decrease, and algorithm would prefer shorter \mathbf{w} .

Question assigned to the following page: [13](#)

13. (Bonus 20 points) Derive a simplified Lagrange dual problem of the hard-margin SVM dual. Is your Lagrange dual problem of the hard-margin SVM dual the same as the hard-margin SVM primal? Are they "similar" in any sense? Describe your findings.

For your information, this is what chatGPT says.



For quadratic programming, is the dual problem of the dual problem the same as the primal problem?



In quadratic programming, the relationship between the primal and dual problems is more intricate than in linear programming. The dual problem of the dual problem may not necessarily be the same as the primal problem.

In quadratic programming, the primal problem involves minimizing a quadratic objective function subject to linear constraints. The dual problem is derived from this primal problem, and it involves maximizing a concave quadratic function subject to certain constraints.

When you take the dual of the dual problem in quadratic programming, it may not directly lead you back to the primal problem. The relationship is more complex due to the quadratic nature of the objective function.

In summary, for quadratic programming, the dual of the dual is not always the same as the primal problem.

$$\begin{aligned} \text{SVM: } \min_{w, b} (\max_{\alpha_n} \mathcal{L}(w, b, \alpha_n)) &\geq \max_{\alpha_n > 0} (\min_{w, b} \mathcal{L}(w, b, \alpha_n)) \\ &= \max_{\alpha_n > 0} \left(\min_{b, w} \frac{1}{2} w^T w + \sum_{n=1}^N \alpha_n (1 - y_n (w^T x_n + b)) \right) \end{aligned}$$

$$\begin{aligned} \text{Strong duality ensure that } \min_{\alpha_n} (\max_{b, w} \mathcal{L}(b, w, \alpha_n)) \\ &= \max_{\alpha_n} (\min_{b, w} \mathcal{L}(b, w, \alpha_n)) \end{aligned}$$

(convex primal.
feasible primal.
linear constraints)

The dual of the dual problem can be reversed.
back to the primal.

