

Homework #5

RELEASE DATE: 11/15/2023

DUE DATE: 12/06/2023 (**THREE WEEKS, YEAH!!**), BEFORE 13:00 on GRADESCOPE

QUESTIONS ARE WELCOMED ON DISCORD (INFORMALLY) OR NTU COOL (FORMALLY).

You will use Gradescope to upload your scanned/printed solutions. For problems marked with (), please follow the guidelines on the course website and upload your source code to Gradescope as well. Any programming language/platform is allowed.*

Any form of cheating, lying, or plagiarism will not be tolerated. Students can get zero scores and/or fail the class and/or be kicked out of school and/or receive other punishments for those kinds of misconducts.

Discussions on course materials and homework solutions are encouraged. But you should write the final solutions alone and understand them fully. Books, notes, and Internet resources can be consulted, but not copied from.

Since everyone needs to write the final solutions alone, there is absolutely no need to lend your homework solutions and/or source codes to your classmates at any time. In order to maximize the level of fairness in this class, lending and borrowing homework solutions are both regarded as dishonest behaviors and will be punished according to the honesty policy.

You should write your solutions in English or Chinese with the common math notations introduced in class or in the problems. We do not accept solutions written in any other languages.

This homework set comes with 240 points and 20 bonus points. In general, every homework set would come with a full credit of 240 points, with some possible bonus points.

Hard-Margin Support Vector Machine

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? :-))

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.)

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.)

4. (20 points) Let's call the dual solution to the uneven margin SVM above α_ρ^* . 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.)

Operation of Kernels

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.)

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.)

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 $\tilde{\phi}(x)$.)

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 $\mathcal{X} = \mathbb{R}^d - \{\mathbf{0}\}$.

(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.)

Experiments with Soft-Margin Support Vector Machine

For Problems 9 to 12, we are going to experiment with a real-world data set. Download the processed satimage data sets from LIBSVM Tools.

Training: <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/satimage.scale>

Testing: <https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/satimage.scale.t>

We will consider binary classification problems of the form “one of the classes” (as the positive class) versus “the other classes” (as the negative class).

The data set contains thousands of examples, and some quadratic programming packages cannot handle this size. We recommend that you consider the LIBSVM package

<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

Regardless of the package that you choose to use, please read the manual of the package carefully to make sure that you are indeed solving the soft-margin support vector machine taught in class like the dual formulation below:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N \alpha_n \alpha_m y_n y_m K(\mathbf{x}_n, \mathbf{x}_m) - \sum_{n=1}^N \alpha_n \\ \text{subject to} \quad & \sum_{n=1}^N y_n \alpha_n = 0 \\ & 0 \leq \alpha_n \leq C \quad n = 1, \dots, N. \end{aligned}$$

In the following problems, please use the 0/1 error for evaluating E_{val} and E_{out} (through the test set). Some practical remarks include

- (i) Please tell your chosen package to **not** automatically scale the data for you, lest you should change the effective kernel and get different results.
 - (ii) It is your responsibility to check whether your chosen package solves the designated formulation with enough numerical precision. Please read the manual of your chosen package for software parameters whose values affect the outcome—any ML practitioner needs to deal with this kind of added uncertainty.
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.
 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.
 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.
 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.

Bonus: Dual of Dual

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.