Homework #5

RELEASE DATE: 12/04/2020

RED BUG FIX: 12/11/2020 17:00

BLUE BUG FIX: 12/16/2020 15:30

DUE DATE: 12/25/2020 (MERRY XMAS!!), BEFORE 13:00 on Gradescope

RANGE: MOOC LECTURES 201-204 (WITH BACKGROUND FROM ML FOUNDATIONS)

QUESTIONS ARE WELCOMED ON THE NTU COOL FORUM.

We will instruct you on how to use Gradescope to upload your choices and 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. You are encouraged to (but not required to) include a README to help the TAs check your source code. 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 400 points. For each problem, there is one correct choice. For most of the problems, if you choose the correct answer, you get 20 points; if you choose an incorrect answer, you get -10 points. That is, the expected value of random guessing is -20 per problem, and if you can eliminate two of the choices accurately, the expected value of random guessing on the remaining three choices would be 0 per problem. For other problems, the TAs will check your solution in terms of the written explanations and/or code. The solution will be given points between $[-20, 20]$ based on how logical your solution is.

Hard-Margin SVM and Large Margin

1. (Lecture 201) Consider a three-example data set in 1D: $\{(x_n, y_n)\}_{n=1}^3 = \{(-2, -1), (0, +1), (2, -1)\}$, and a polynomial transform $\boldsymbol{\phi}(x) = [1, x, x^2]^T$. Apply the hard-margin SVM on the transformed examples $\{(\boldsymbol{\phi}(x_n), y_n)\}_{n=1}^3$ to get the optimal $(b^*, \mathbf{w}^*)$ in the transformed space. What is the optimal w_1^* that corresponds to the "constant" feature transform? Choose the correct answer; **provide steps of your "human optimization" like page 17 of Lecture 201 slides**.

[a] $w_1^* = 4$

[b] $w_1^* = 2$

[c] $w_1^* = 1$

[d] $w_1^* = 0$

[e] $w_1^* = -1$

(*Hint: If you must, you can use the fact that all three examples are support vector candidates (i.e. on the fat boundary) for this problem and the next one. But you can also challenge itself by solving it without using this fact first.*)

2. (Lecture 201) Following Problem 1, what is the margin achieved by the optimal solution? Choose the correct answer; **provide steps of your "human optimization" like page 17 of Lecture 201 slides**.

[a] 1

[b] 2

[c] 4

[d] 8

[e] 16

(*Hint: You can use the same hint as the previous problem, and write your solution steps for both problems together if needed. Page 14 of Lecture 201 slides should remind you the relationship between* $(b^*, \mathbf{w}^*)$ *and the margin.*)

3. (Lecture 201) 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 $x_1 \leq x_2 \leq \ldots x_M < x_{M+1} \leq x_{M+2} \ldots \leq x_N$, $y_n = -1$ for $n = 1, 2, \ldots, M$, and $y_n = +1$ for $n = M+1, M+2, \ldots, N$. Apply hard-margin SVM without transform on this data set. What is the largest margin achieved? Choose the correct answer; explain your answer.

[a] $\frac{1}{2}(x_N - x_M)$

[b] $\frac{1}{2}(x_{M+1} - x_1)$

[c] $\frac{1}{2}\left(\frac{1}{N-M}\sum_{n=M+1}^{N} x_n - \frac{1}{M}\sum_{n=1}^{M} x_n\right)$

[d] $\frac{1}{2}(x_N - x_1)$

[e] $\frac{1}{2}(x_{M+1} - x_M)$

(*Hint: Have we mentioned that a decision stump is just a 1D perceptron, and the hard-margin SVM is an extension of the perceptron model? :-)*)

4. (Lecture 201) Two points x_1 and x_2 are sampled from a uniform distribution in $[0, 1]$. Consider a large-margin perceptron algorithm that either returns a 1D perceptron with margin at least ρ, or returns a default constant hypothesis of $h(x) = -1$. For $\rho \in [0, 0.5]$, what is the expected number of dichotomies that this algorithm can produce, where expectation is taken over the process that generated (x_1, x_2)? Choose the correct answer; explain your answer.

[a] $2 + 2 \cdot (1 - 2\rho)^2$

[b] $2 + 2 \cdot (2\rho)^2$

[c] $4 \cdot (1 - 2\rho)^2$

[d] $2 - 2 \cdot (1 - 2\rho)^2$

[e] $2 - 2 \cdot (2\rho)^2$

(*Hint: We are mimicking page 24 of Lecture 201 here, and you are encouraged to think about the distance between two points.*)

Dual Problem of Quadratic Programming

In the hard-margin SVM that we introduced in class, we hope to get a hyperplane such that the margin to the positive examples is the same as the margin to the negative examples. Sometimes we need to have different margins for different classes. The need can be written as the following uneven-margin SVM (in its linear form) with parameters $\rho_+ > 0$ and $\rho_- > 0$:

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

Our original hard-margin SVM is just a special case with $\rho_+ = \rho_- = 1$.

5. (Lecture 202) The dual problem of the uneven-margin SVM can be written as

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{x}_n^T\mathbf{x}_m + \square \\ \text{subject to} \quad & \sum_{n=1}^{N} y_n\alpha_n = 0 \\ & \alpha_n \geq 0 \text{ for } n = 1, 2, \ldots, N. \end{aligned}$$

What is $\square$? Choose the correct answer; explain your answer.

[a] $-\sum_{n=1}^{N}\rho_+^{-1}[\![y_n = +1]\!]\,\alpha_n - \sum_{n=1}^{N}\rho_-^{-1}[\![y_n = -1]\!]\,\alpha_n$

[b] $-\sum_{n=1}^{N}\rho_+^{0}[\![y_n = +1]\!]\,\alpha_n - \sum_{n=1}^{N}\rho_-^{0}[\![y_n = -1]\!]\,\alpha_n$

[c] $-\sum_{n=1}^{N}\rho_+[\![y_n = +1]\!]\,\alpha_n - \sum_{n=1}^{N}\rho_-[\![y_n = -1]\!]\,\alpha_n$

[d] $-\sum_{n=1}^{N}\rho_+^{2}[\![y_n = +1]\!]\,\alpha_n - \sum_{n=1}^{N}\rho_-^{2}[\![y_n = -1]\!]\,\alpha_n$

[e] none of the other choices

6. (Lecture 202) Let $\boldsymbol{\alpha}^*$ be an optimal solution of the original hard-margin SVM (i.e. even margin). Which of the following is an optimal solution of the uneven-margin SVM for a given pair of non-negative (ρ_-, ρ_+)? Choose the correct answer; explain your answer.

[a] $\boldsymbol{\alpha}^*$

[b] $\sqrt{\rho_+ \cdot \rho_-}\boldsymbol{\alpha}^*$

[c] $\frac{2}{\rho_+ + \rho_-}\boldsymbol{\alpha}^*$

[d] $\frac{\rho_+^2 + \rho_-^2}{2}\boldsymbol{\alpha}^*$

[e] $\frac{\rho_+ + \rho_-}{2}\boldsymbol{\alpha}^*$

Properties of Kernels

7. (Lecture 203) Let $K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$ be a valid kernel function with range $\subseteq [0, 2)$. Which of the following function is **not** always a valid kernel? Choose the correct answer; **explain your answer either by providing a counter-example of your choice (highly recommended), or by explaining why other choices are all valid kernels**.

[a] $2^{K(\mathbf{x},\mathbf{x}')}$

[b] $(2 - K(\mathbf{x}, \mathbf{x}'))^{-2}$

[c] $2 + K(\mathbf{x}, \mathbf{x}')$

[d] $\log_2 K(\mathbf{x}, \mathbf{x}')$

[e] $(K(\mathbf{x}, \mathbf{x}'))^2$

8. (Lecture 203) For any feature transform $\boldsymbol{\phi}$ from $\mathcal{X}$ to $\mathcal{Z}$, the squared distance between two examples $\mathbf{x}$ and $\mathbf{x}'$ is $\|\boldsymbol{\phi}(\mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}')\|^2$ in the $\mathcal{Z}$-space. For the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, compute the squared distance with the kernel trick. Then, for any two examples $\mathbf{x}$ and $\mathbf{x}'$, what is the tightest upper bound for their squared distance in the $\mathcal{Z}$-space? Choose the correct answer; explain your answer.

[a] 0

[b] 1

[c] 2

[d] 3

[e] 4

9. (Lecture 203) For a set of examples $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ and a kernel function K, consider a hypothesis set that contains

$$h_{\boldsymbol{\alpha},b}(\mathbf{x}) = \text{sign}\left(\sum_{n=1}^{N} y_n \alpha_n K(\mathbf{x}_n, \mathbf{x}) + b\right).$$

The classifier returned by SVM can be viewed as one such $h_{\boldsymbol{\alpha},b}$, where the values of $\boldsymbol{\alpha}$ is determined by the dual QP solver and b is calculated from the KKT conditions.

In this problem, we study a simpler form of $h_{\boldsymbol{\alpha},b}$ where $\boldsymbol{\alpha} = \mathbf{1}$ (the vector of all 1's) and $b = 0$. Let us name $h_{\mathbf{1},0}$ as $\hat{h}$ for simplicity. We will show that when using the Gaussian kernel $K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2)$, if γ is large enough, $E_{\text{in}}(\hat{h}) = 0$. That is, when using the Gaussian kernel, we can "easily" separate the given data set if γ is large enough.

Assume that the distance between any pair of different $(\mathbf{x}_n, \mathbf{x}_m)$ in the $\mathcal{X}$-space is no less than ϵ. That is,

$$\|\mathbf{x}_n - \mathbf{x}_m\| \geq \epsilon \quad \forall n \neq m.$$

What is the tightest lower bound of γ that ensures $E_{\text{in}}(\hat{h}) = 0$? Choose the correct answer; explain your answer.

[a] $\frac{\ln^2(N+1)}{\epsilon^2}$

[b] $\frac{\ln(N+1)}{\epsilon^2}$

[c] $\frac{\ln(N)}{\epsilon^2}$

[d] $\frac{\ln(N-1)}{\epsilon^2}$

[e] $\frac{\ln^2(N-1)}{\epsilon^2}$

Kernel Perceptron Learning Algorithm

10. (Lecture 203) In this problem, we are going to apply the kernel trick to the perceptron learning algorithm introduced in Machine Learning Foundations. If we run the perceptron learning algorithm on the transformed examples $\{(\boldsymbol{\phi}(\mathbf{x}_n), y_n)\}_{n=1}^N$, the algorithm updates $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$ when the current $\mathbf{w}_t$ makes a mistake on $(\boldsymbol{\phi}(\mathbf{x}_{n(t)}), y_{n(t)})$:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + y_{n(t)}\boldsymbol{\phi}(\mathbf{x}_{n(t)})$$

Because every update is based on one (transformed) example, if we take $\mathbf{w}_0 = \mathbf{0}$, we can represent every $\mathbf{w}_t$ as a linear combination of $\{\boldsymbol{\phi}(\mathbf{x}_n)\}_{n=1}^N$. We can then maintain the linear combination coefficients instead of the whole $\mathbf{w}$. Assume that we maintain an N-dimensional vector $\boldsymbol{\alpha}_t$ in the t-th iteration such that

$$\mathbf{w}_t = \sum_{n=1}^{N} \alpha_{t,n}\boldsymbol{\phi}(\mathbf{x}_n)$$

for $t = 0, 1, 2, \ldots$. Set $\boldsymbol{\alpha}_0 = \mathbf{0}$ (N zeros) to match $\mathbf{w}_0 = \mathbf{0}$ ($\tilde{d}+1$ zeros). How should $\boldsymbol{\alpha}_t$ be updated to $\boldsymbol{\alpha}_{t+1}$ when the current $\mathbf{w}_t$ (represented by $\boldsymbol{\alpha}_t$) makes a mistake on $(\boldsymbol{\phi}(\mathbf{x}_{n(t)}), y_{n(t)})$? Choose the correct answer; explain your answer.

[a] $\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t$ except $\alpha_{t+1,n(t)} \leftarrow \alpha_{t,n(t)} + 1$

[b] $\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t$ except $\alpha_{t+1,n(t)} \leftarrow \alpha_{t,n(t)} - 1$

[c] $\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t$ except $\alpha_{t+1,n(t)} \leftarrow \alpha_{t,n(t)} + y_{n(t)}$

[d] $\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t$ except $\alpha_{t+1,n(t)} \leftarrow \alpha_{t,n(t)} - y_{n(t)}$

[e] $\boldsymbol{\alpha}_{t+1} \leftarrow \boldsymbol{\alpha}_t + \mathbf{y}$

(*Hint: Although we did not teach Lecture 205, if you have watched it by yourself from YouTube, you will find its page 15 loosely related. You should be able to solve this problem without watching Lecture 205, though.*)

11. (Lecture 203) Following Problem 10, the update rule takes care of the training iterations. In addition, we need to evaluate $\mathbf{w}_t^T\boldsymbol{\phi}(\mathbf{x})$ not only for predicting new $\mathbf{x}$ but also for checking whether $\mathbf{w}_t$ makes any mistake on some example $\mathbf{x}$ during training. Which of the following equation computes $\mathbf{w}_t^T\boldsymbol{\phi}(\mathbf{x})$ with the kernel trick $K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$? Choose the correct answer; explain your answer.

[a] $\sum_{n=1}^N \alpha_{t,n}K(\mathbf{x}_n, \mathbf{x})$

[b] $-\sum_{n=1}^N \alpha_{t,n}K(\mathbf{x}_n, \mathbf{x})$

[c] $\sum_{n=1}^N y_n\alpha_{t,n}K(\mathbf{x}_n, \mathbf{x})$

[d] $\sum_{n=1}^N \alpha_{t,n}^2(K(\mathbf{x}_n, \mathbf{x}))^2$

[e] $\sum_{n=1}^N \alpha_{t,n}^2K(\mathbf{x}_n, \mathbf{x})$

Soft-Margin SVM

12. (Lecture 204) Consider the soft-margin SVM taught in our class. Assume that after solving the dual problem, every example is a bounded support vector. That is, the optimal solution $\boldsymbol{\alpha}^*$ satisfies $\boldsymbol{\alpha}_n^* = C$ for every example. In this case, there may be multiple solutions for the optimal b^* for the primal SVM problem. What is the largest such b^*? Choose the correct answer; explain your answer.

[a] $\min\limits_{n=1,2,\dots,N}\Big(1 - \sum_{m=1}^{N} y_m \alpha_m K(x_n, x_m)\Big)$

[b] $\min\limits_{n:\ y_n>0}\Big(1 - \sum_{m=1}^{N} y_m \alpha_m K(x_n, x_m)\Big)$

[c] $\min\limits_{n:\ y_n<0}\Big(1 - \sum_{m=1}^{N} y_m \alpha_m K(x_n, x_m)\Big)$

[d] $\operatorname{average}\limits_{n:\ y_n>0}\Big(1 - \sum_{m=1}^{N} y_m \alpha_m K(x_n, x_m)\Big)$

[e] $\operatorname{average}\limits_{n:\ y_n<0}\Big(1 - \sum_{m=1}^{N} y_m \alpha_m K(x_n, x_m)\Big)$

13. (Lecture 204) In class, we taught the non-linear soft-margin SVM as follows.

$$
\begin{aligned}
(P_1) \quad \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n \\
\text{subject to} \quad & y_n\Big(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) + b\Big) \geq 1 - \xi_n, \text{ for } n = 1, 2, \dots, N, \\
& \xi_n \geq 0, \text{ for } n = 1, 2, \dots, N.
\end{aligned}
$$

The SVM penalizes the margin violation linearly. Another popular formulation penalizes the margin violation quadratically. In this problem, we derive the dual of such a formulation. The formulation as follows:

$$
\begin{aligned}
(P_2) \quad \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{n=1}^{N}\xi_n^2 \\
\text{subject to} \quad & y_n\Big(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n) + b\Big) \geq 1 - \xi_n, \text{ for } n = 1, 2, \dots, N.
\end{aligned}
$$

We do not have the $\xi_n \geq 0$ constraints as any negative ξ_n would never be an optimal solution of (P_2)—you are encouraged to think about why. Anyway, the dual problem of (P_2) will look like this:

$$
\begin{aligned}
(D_2) \quad \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \cdot \Diamond - \sum_{n=1}^{N}\alpha_n \\
\text{subject to} \quad & \sum_{n=1}^{N} y_n\alpha_n = 0 \\
& \alpha_n \geq 0, \text{ for } n = 1, 2, \dots, N.
\end{aligned}
$$

Let the kernel function $K(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^T\boldsymbol{\phi}(\mathbf{x}')$. What is $\Diamond$? Choose the correct answer; explain your answer.

[a] $(2C \cdot K(\mathbf{x}_n, \mathbf{x}_m))$

[b] $(K(\mathbf{x}_n, \mathbf{x}_m) + 2C\,[\![n = m]\!])$

[c] $(K(\mathbf{x}_n, \mathbf{x}_m) + C\,[\![n = m]\!])$

[d] $(K(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{C}\,[\![n = m]\!])$

[e] $(K(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{2C}\,[\![n = m]\!])$

14. (Lectures 202/204) After getting the optimal $\boldsymbol{\alpha}^*$ for (D_2), how can we calculate the optimal $\boldsymbol{\xi}^*$ for (P_2)? Choose the correct answer; explain your answer.

[a] $\boldsymbol{\xi}^* = \boldsymbol{\alpha}^*$

[b] $\boldsymbol{\xi}^* = 2\boldsymbol{\alpha}^*$

[c] $\boldsymbol{\xi}^* = C\boldsymbol{\alpha}^*$

[d] $\boldsymbol{\xi}^* = \frac{1}{C}\boldsymbol{\alpha}^*$

[e] $\boldsymbol{\xi}^* = \frac{1}{2C}\boldsymbol{\alpha}^*$

Experiments with Soft-Margin SVM

For Problems 15 to 20, 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:

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

In the following problems, please use the 0/1 error for evaluating E_{in}, 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.

15. (Lectures 201/204, *) Consider the linear soft-margin SVM. That is, either solve the primal formulation of soft-margin SVM with the given $\mathbf{x}_n$, or take the linear kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \mathbf{x}_n^T \mathbf{x}_m$ in the dual formulation. With $C = 10$, and the binary classification problem of "3" versus "not 3", which of the following numbers is closest to $\|\mathbf{w}\|$ after solving the linear soft-margin SVM? Choose the closest answer; provide your command/code.

[a] 7.0

[b] 7.5

[c] 8.0

[d] 8.5

[e] 9.0

16. (Lectures 203/204, *) 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. With $C = 10$, $Q = 2$, which of the following soft-margin SVM classifiers reaches the lowest E_{in}? Choose the correct answer; provide your command/code.

[a] "1" versus "not 1"

[b] "2" versus "not 2"

[c] "3" versus "not 3"

[d] "4" versus "not 4"

[e] "5" versus "not 5"

17. (Lectures 203/204, *) Following Problem 16, which of the following numbers is closest to the maximum number of support vectors within those five soft-margin SVM classifiers? Choose the closest answer; provide your command/code.

[a] 500

[b] 600

[c] 700

[d] 800

[e] 900

18. (Lectures 203/204, *) Consider the Gaussian kernel $K(\mathbf{x}_n, \mathbf{x}_m) = \exp\left(-\gamma ||\mathbf{x}_n - \mathbf{x}_m||^2\right)$. For the binary classification problem of "6" versus "not 6", when fixing $\gamma = 10$, which of the following values of C results in the lowest E_{out}? Choose the correct answer; provide your command/code.

[a] 0.01

[b] 0.1

[c] 1

[d] 10

[e] 100

19. (Lectures 203/204, *) Following Problem 18, when fixing $C = 0.1$, which of the following values of γ results in the lowest E_{out}? Choose the correct answer; provide your command/code.

[a] 0.1

[b] 1

[c] 10

[d] 100

[e] 1000

20. (Lectures 203/204, *) Following Problem 18 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 $C = 0.1$ and use the validation procedure to choose the best γ among $\{0.1, 1, 10, 100, 1000\}$ according to E_{val}. If there is a tie of E_{val}, choose the smallest γ. Repeat the procedure 1000 times. Which of the following values of γ is selected the most number of times? Choose the correct answer; provide your command/code.

[a] 0.1

[b] 1

[c] 10

[d] 100

[e] 1000