

## INSTRUCTIONS

- **Due:** Thursday, 9 April 2020 at 11:59 PM EDT.
- **Format:** Complete this pdf with your work and answers. Whether you edit the latex source, use a pdf annotator, or hand write / scan, make sure that your answers (tex'ed, typed, or handwritten) are within the dedicated regions for each question/part. If you do not follow this format, we may deduct points.
- **How to submit:** Submit a pdf with your answers on Gradescope. Log in and click on our class 10-315, click on the appropriate *Written* assignment, and upload your pdf containing your answers. Don't forget to submit the associated *Programming* component on Gradescope if there is any programming required.
- **Policy:** See the course website for homework policies and Academic Integrity.

Name	
Andrew ID	
Hours to complete (both written and programming)?	

For staff use only

Q1	Q2	Q3	Q4	Total
/16	/12	/20	/22	/ 70

# Q1. [16pts] Kernels

## (a) Kernel Computation Cost

- (i) [2pts] Suppose we have a two-dimensional input space such that the input vector is  $\mathbf{x} = [x_1, x_2]^T$ . Define the feature mapping  $\phi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1x_2, x_2^2]^T$ . What is the corresponding kernel function, i.e.  $k(\mathbf{x}, \mathbf{z})$ ? Do not leave  $\phi(\mathbf{x})$  in your final answer. Simplify your answer and show your work.

- (ii) [2pts] Suppose we want to compute the value of the kernel function  $k(\mathbf{x}, \mathbf{z})$  from the previous question, on two vectors  $\mathbf{x}, \mathbf{z} \in \mathbb{R}^2$ . How many operations (additions, multiplications, powers) are needed if you map the input vector to the feature space and then perform the dot product on the mapped features? Show your work.

<b>Num:</b>
-------------

<b>Work:</b>
--------------

- (iii) [2pts] How many operations (additions, multiplications, powers) are needed if you compute through the kernel function you derived in question 1? Show your work.

<b>Num:</b>
-------------

<b>Work:</b>
--------------

**(b)** [10pts] Sum of Kernels

Assume  $k_1(\cdot, \cdot)$  is a kernel with corresponding feature mapping  $\phi_1 : \mathbb{R}^M \rightarrow \mathbb{R}^{M_1}$ , and  $k_2(\cdot, \cdot)$  is a kernel with corresponding feature mapping  $\phi_2 : \mathbb{R}^M \rightarrow \mathbb{R}^{M_2}$ , both acting on the same space. Prove that,  $k'(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$  is also a valid kernel by constructing its corresponding feature mapping  $\phi'(\cdot)$ .

## Q2. [12pts] Matrix Properties

Consider the following matrices:

$$\mathbf{Z}_1 = \begin{bmatrix} 2 & 0 & 0 & 0 \\ -6 & 5 & -1 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & -5 & -1 \end{bmatrix}$$

$$\mathbf{Z}_3 = \begin{bmatrix} 1 & 0.39 & 0.42 & 0.54 \\ 0.39 & 1 & 0.94 & 0.50 \\ 0.42 & 0.94 & 1 & 0.38 \\ 0.54 & 0.50 & 0.38 & 1 \end{bmatrix}$$

$$\mathbf{Z}_2 = \begin{bmatrix} 8 & -6 & 3 & 0 \\ 0 & 3 & 0 & 0 \\ -10 & 12 & -3 & 0 \\ -10 & -10 & 5 & 8 \end{bmatrix}$$

$$\mathbf{Z}_4 = \begin{bmatrix} 39 & -7 & 6 & -8 \\ -7 & 41 & 46 & -10 \\ 6 & 46 & 58 & -15 \\ -8 & -10 & -15 & 5 \end{bmatrix}$$

- (a) [4pts] What are the eigenvalues for each of these matrices (to 2sf)? You don't have to calculate these by hand. For example, you can code it up really quick with `numpy.linalg.eigvals`.

<b>Z<sub>1</sub>:</b>
<b>Z<sub>2</sub>:</b>
<b>Z<sub>3</sub>:</b>
<b>Z<sub>4</sub>:</b>

- (b) [4pts] Valid kernel functions should satisfy conditions in Mercer's theorem. Which of the matrices above could be valid Mercer kernels? Select all that apply.

- ☐ **Z<sub>1</sub>**  
☐ **Z<sub>2</sub>**  
☐ **Z<sub>3</sub>**  
☐ **Z<sub>4</sub>**

- (c) [4pts] Which of the matrices would cause the following optimization problem to be convex? Select all that apply.

$$\min_{\mathbf{x}} \mathbf{x}^T \mathbf{Z}_i \mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \preceq \mathbf{b}$$

- ☐ **Z<sub>1</sub>**  
☐ **Z<sub>2</sub>**  
☐ **Z<sub>3</sub>**  
☐ **Z<sub>4</sub>**

### Q3. [20pts] SVM and Duality

In this question, we are considering the kernelized version of the hard-margin SVM. The primal form is given by:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y^{(i)} \left( \mathbf{w}^T \phi(\mathbf{x}^{(i)}) + b \right) \geq 1 - \xi_i, \quad \forall i \in \{1, \dots, N\} \\ & \xi_i \geq 0, \quad \forall i \in \{1, \dots, N\} \end{aligned}$$

- (a) [4pts] Write the Lagrangian for this SMV. Please use  $\alpha_i$  as the dual variables on the first set of constraints and use  $\eta_i$  as the dual variables on the second set of constraints.

$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \eta)$ :

To find the dual form of this SVM, we need to find a closed form solution to the following optimization of the Lagrangian:

$$J(\alpha) = \min_{\mathbf{w}, b, \xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \eta)$$

- (b) Give the partial derivative of the Lagrangian with respect to each primal variable and set each partial derivative equal to zero.

- (i) [2pts]

$\partial \mathcal{L} / \partial \mathbf{w}$ :

- (ii) [2pts]

$\partial \mathcal{L} / \partial b$ :

- (iii) [2pts]

$\partial \mathcal{L} / \partial \xi_i$ :

(c) [10pts]

Utilizing the expressions derived in the previous part, convert the Lagrangian into an expression for  $J(\alpha)$  in terms of just the  $\alpha_i$  dual variables, the data  $y^{(i)}$  and  $\mathbf{x}^{(i)}$ , and the kernel function  $k(\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})^T \phi(\mathbf{z})$ . Do not include  $\phi(\cdot)$  in your final answer.

Hint: Plug an expression for  $\mathbf{w}$  from the previous part into the Lagrangian.

$\mathcal{L}(\alpha)$ :

(continued if needed)

## Q4. [22pts] Programming

The following questions should be completed after you work through the programming portion of this assignment.

(a) [4pts] Kernel Functions

Include surface plots for the boxcar kernel with width=2, the RBF kernel with  $\gamma = 0.1$ , the linear kernel, and the polynomial kernel with  $d=2$ .

**Plot boxcar, width=2:**

**Plot RBF, gamma=0.1:**

**Plot linear:**

**Plot polynomial, d=2:**



**(b)** Kernel Regression

- (i) [5pts] Include surface plots for the kernel regression with  $N=2$  training points with the boxcar kernel with width=2, the RBF kernel with  $\gamma = 0.1$ , the linear kernel, the polynomial kernel with  $d=2$ , and the polynomial kernel with  $d=3$ .

**Plot N=2 boxcar, width=2:**

**Plot N=2 RBF,  $\gamma=0.1$ :**

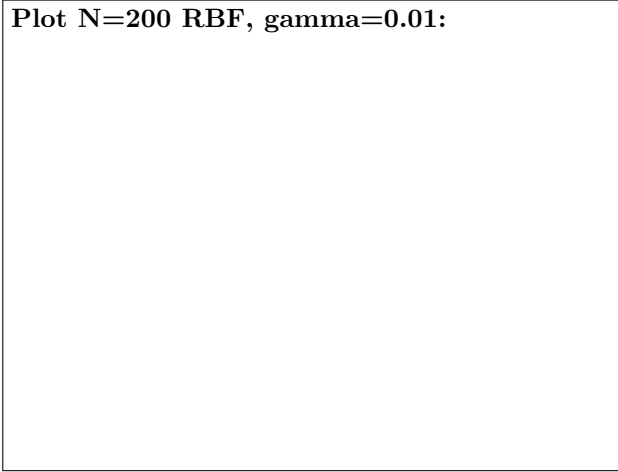
**Plot N=2 linear:**

**Plot N=2 polynomial,  $d=2$ :**

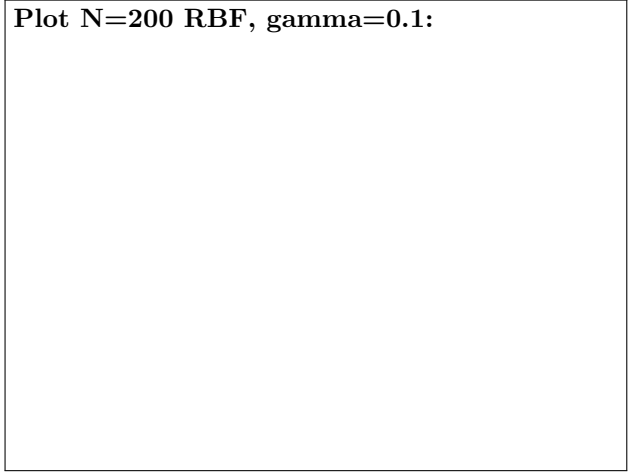
**Plot N=2 polynomial,  $d=3$ :**

- (ii) [5pts] Include surface plots for the kernel regression with  $N=200$  training points with the RBF kernel with  $\gamma = 0.01, 0.1$ , and  $1$ .

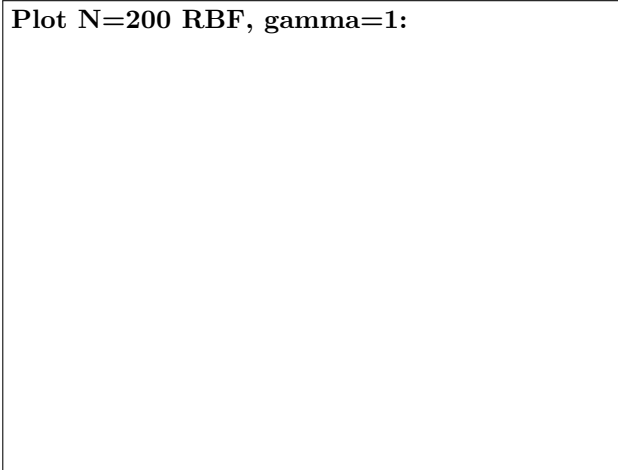
**Plot  $N=200$  RBF,  $\gamma=0.01$ :**



**Plot  $N=200$  RBF,  $\gamma=0.1$ :**



**Plot  $N=200$  RBF,  $\gamma=1$ :**

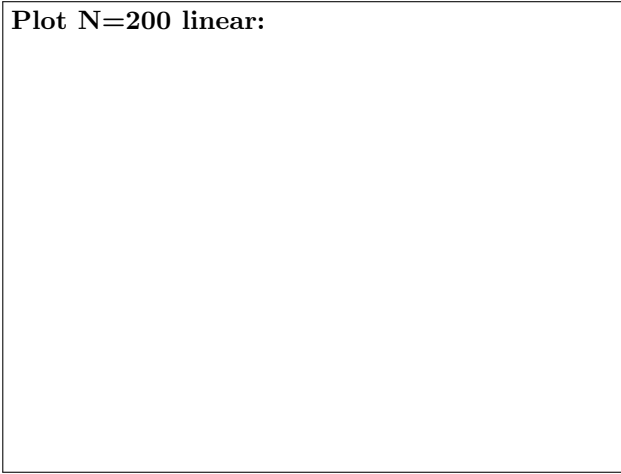


**Explain the relationship between settings of  $\gamma$  in the RBF filter and over/under fitting.**

A large empty rectangular box intended for the student's explanation of the relationship between the  $\gamma$  parameter in the RBF filter and the concepts of overfitting and underfitting.

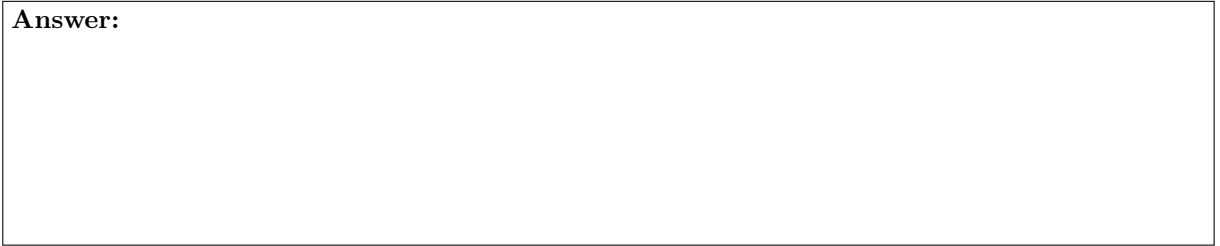
- (iii) [4pts] Include surface plots for the kernel regression with  $N=200$  training points with the linear kernel.

**Plot  $N=200$  linear:**



For the linear kernel with  $N=200$  training points, why is the prediction surface significantly below the training points?

**Answer:**



- (iv) [4pts] Among all of the kernels and hyperparameter settings that the autograder test cases ran through, which kernel and hyperparameter combination should you choose? Why?

**Answer:**

