| 10-315 | Intro to Machine Learning | HW3 |
|---|---|---|

**INSTRUCTIONS**

- **Due: Monday, 8 November 2021 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 | Total |
|---|---|---|---|
| / 25 | / 20 | / 20 | / 65 |

# Q1. [25pts] Kernels

**(a)** Kernel Computation Cost

**(i)** [4pts] 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 $\phi(\mathbf{x}) = [x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1]^T$. What is the corresponding kernel function, i.e. $k(\mathbf{x}, \mathbf{z})$? Do not leave $\phi(\cdot)$ in your final answer. Simplify your answer to write it using input vectors $\mathbf{x}, \mathbf{z}$ and show your work.

**(ii)** [4pts] 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.

**Num:**

**Work:**

**(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.

**Num:**

**Work:**

**(b)** [5pts] Sum of Kernels

Assume $k_1(\cdot,\cdot)$ is a kernel with corresponding feature mapping $\phi_1 : \mathbb{R}^M \to \mathbb{R}^{M_1}$, and $k_2(\cdot,\cdot)$ is a kernel with corresponding feature mapping $\phi_2 : \mathbb{R}^M \to \mathbb{R}^{M_2}$, 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 $\phi'(\cdot)$.

**(c)** [10pts] Which of the following are valid kernels for SVM and why (explain your reasoninge.g. if proposed kernel doesn't satisfy a property, explain why)? Data points $x, z$ are scalars and $\mathbf{x}, \mathbf{z}$ are vectors.

(i) $K(x,z) = -xz$

(ii) $K(\mathbf{x},\mathbf{z}) = 10\mathbf{x}\cdot\mathbf{z} + (\mathbf{x}\cdot\mathbf{z}+1)^8$

(iii) $K(x,z) = x^2 z$

(iv) $K(\mathbf{x},\mathbf{z}) = -\exp(\|\mathbf{x}-\mathbf{z}\|^2)$

(v) $K(\mathbf{x},\mathbf{z}) = \exp(-(\|\mathbf{x}\|^2 + \|\mathbf{z}\|^2))$

# Q2. [20pts] Linear and Kernel SVMs

**(a)** [12pts] Recall that the soft-margin primal SVM problem is

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}\cdot\mathbf{w} + C\sum_{i=1}^{n}\xi_i$$
$$\text{s.t.} \quad \xi_i \geq 0 \quad \forall i = 1,\ldots,n$$
$$(\mathbf{w}\cdot\mathbf{x}_i + b)y_i \geq (1-\xi_i) \quad \forall i = 1,\ldots,n.$$

For hard-margin primal SVM, $\xi_i = 0, \forall i$. We can get the kernel SVM by taking the dual of the primal problem and then replace the product of $\mathbf{x}_i \cdot \mathbf{x}_j$ by $k(\mathbf{x}_i, \mathbf{x}_j)$, where $k(.,.)$ can be any kernel function:

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j) + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.t.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0, \forall i = 1,2,...,n$$
$$\alpha_i \geq 0, \forall i = 1,2,...,n$$

Figure 1 plots SVM decision boundaries resulting from using different kernels and/or different slack penalties. In Figure 1, there are two classes of training data, with labels $y_i \in \{-1,1\}$, represented by circles and squares respectively. The SOLID circles and squares represent the support vectors. Match each plot in Figure 1 with the letter of the optimization problem below and explain WHY you pick the figure for a given kernel.

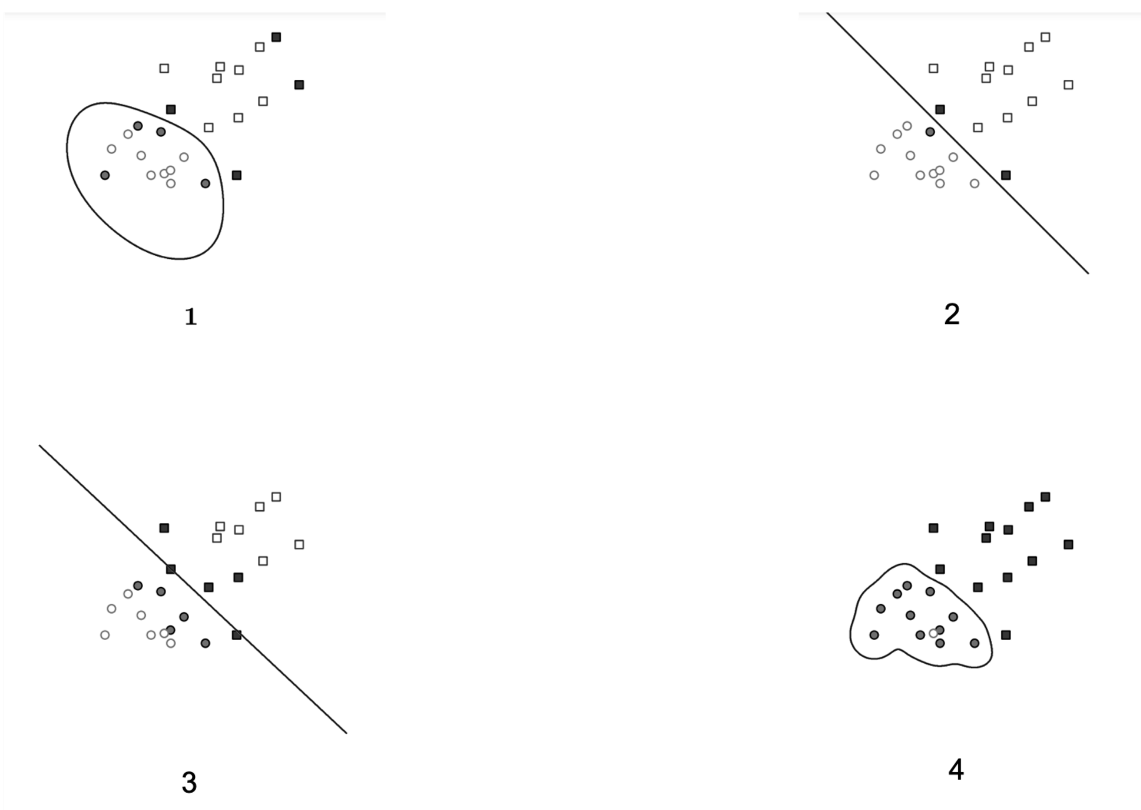


Figure 1: Induced Decision Boundaries

(a) A soft-margin linear SVM with C = 0.1.

(b) A soft-margin linear SVM with C = 10.

(c) A hard-margin kernel SVM with $K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{1}{4}\|\mathbf{u} - \mathbf{v}\|^2\right)$

(d) A hard-margin kernel SVM with $K(\mathbf{u}, \mathbf{v}) = \exp\left(-4\|\mathbf{u} - \mathbf{v}\|^2\right)$

Hint: It may help to think about the decision boundary for kernel SVM based on derivation in last question.

**(b)** [8pts] You are given a training dataset, as shown in Fig 2. Note that the training data comes from sensors which can be error-prone, so you should avoid trusting any specific point too much. For this problem, assume that we are training an SVM with a quadratic kernel.

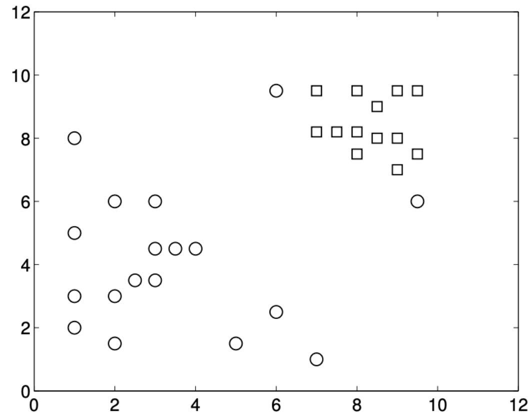


Figure 2: Training dataset

(a) Where would the decision boundary be for very large values of C (i.e., C→ $\infty$)? Draw on figure and justify your answer.

(b) For C close to 0, indicate in the figure where you would expect the decision boundary to be? Justify your answer.

(c) Which of the two cases above would you expect to work better in the classification task? Why?

# Q3. [20pts] 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 polynomial kernel with d=3, and the RBF kernel with gamma = 0.1.

| **Plot polynomial, d=3:** | **Plot RBF, gamma=0.1:** |
|---|---|

**(b)** Kernel Regression

**(i)** [4pts] Include surface plots for the kernel regression with N=2 training points with the polynomial kernel with d=3, and the RBF kernel with gamma = 0.1.

| **Plot N=2 polynomial, d=3:** | **Plot N=2 RBF, gamma=0.1:** |
|---|---|

**(ii)** [4pts] Include surface plots for the kernel regression with N=200 training points with the polynomial kernel with d=3, and the RBF kernel with gamma = 0.1.

| **Plot N=200 polynomial, d=3:** | **Plot N=200 RBF, gamma=0.1:** |
|---|---|
| | |

**(iii)** [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.**

**(iv)** [3pts] 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:**