For your answers pdf file: Please include all answers organized in the order of the problems as given below. For example, computer numerical outputs and plots for Problem 2(a)-(b) should be clearly labeled and in the same place as the text "observe and explain" answers, all in the section where you have Pr. 2(a)-(b) answers. Thank you.

1. In a support vector machine classifier, for the primal Lagrangian that includes slack variables (Lecture 14 Eq. (3') on page 12), note that there are 4 terms on the right-hand side (RHS) of the equation for L. Please answer, by inspection and thought (answers can be short):

   (a) Which term relates most directly to margin width? For that term, what variable change will increase the margin width?

   (b) Consider the second term on the RHS (the term that includes the coefficient C). If the data is linearly separable in $\underline{u}$-space, is it possible for this term to be 0 at the SVM solution ($\xi_i^* = 0 \ \forall i$)? If yes, then what does the 0 value mean in terms of data placement relative to margin boundaries and decision boundary? If no, then why not?

   (c) Suppose you have a dataset that is linearly separable in $\underline{u}$-space, and you have tried an SVM classifier with $C = 1$ and you got good results. What do you think would happen if you made C very small ($C \ll 1$)? What do you think would happen if you made C very large ($C \gg 1$)? You will try this numerically below in Problem 2 to find out.

2. In this problem you will code and use a 2-class SVM classifier on datasets from Homework 1.
   For the SVM classifier, use `sklearn.svm.SVC`. This implementation includes slack variables with parameter C (and extensions to multiclass problems using OvO, which we won't use in this problem); it is based on LIB-SVM, a well-known SVM classification package.
   You may also use `plotSVMBoundaries()` to plot decision boundaries and regions; it is posted along with this homework.

   This problem uses **un**augmented notation.

   For each part below, and for each stated dataset, report the following:

   (i) Classification accuracy on training set and on testing set

   (ii) For linear kernel, the weight vector $\underline{w}$ and bias w_0.

   (iii) For each given value of parameters C and γ, give 2 plots:
   Plot of training data, showing decision boundary and regions, and indicating which vectors are support vectors*;

Plot of testing data, also showing decision boundary and regions learned from the training data.

***Tips:** (1) support vectors are defined and described at the end of this problem; (2) `sklearn.svm.SVC` can output support vectors with the attribute `support_vectors`.

(iv) Also answer anything requested specifically in each part.

(a) dataset1 and dataset2. Using a linear kernel, run SVM for $C = 0.01$, $C = 1$, $C = 100$. Observe and explain if you can, the effect of linear separability of the data on the results (decision boundaries and regions, accuracies). Also observe and explain if you can, the effect of C on the results and on the support vectors. (12 plots total.)

(b) dataset1 and dataset2. Use RBF kernel, and run SVM for $C = 1$, $\gamma = 1, 10$. Observe and explain if you can, the effect of linear separability or nonseparability of the data on the results (decision boundaries and regions, accuracies, and overfitting/underfitting). Also observe and explain if you can, the effect of γ on the results (decision boundaries and regions, accuracies, and overfitting/underfitting). (8 plots total.)

(c) dataset3. Using a linear kernel, run SVM for $C = 1$. Observe and explain the effect of linear separability or nonseparability of the data on the results (decision boundaries and regions, accuracies) (e.g., by comparing with (a)). Do the support vectors indicated in the plot make sense? Briefly justify your answer. (2 plots total.)

(d) dataset3. Using an RBF kernel and $C = 1$, try $\gamma = 0.1, 10, 1000$. For this part only, no need to indicate which vectors are support vectors; such indicators might obscure the visibility of the decision regions. Observe and explain the difference relative to a linear kernel, and the effect of γ on the results. (6 plots total.)

Support vectors are the data points (from the training set) that determine the position and orientation of the max-margin decision-boundary hyperplane. The support vectors are mathematically defined as the training data points that have Lagrange multiplier value (at the solution) of $\lambda_i \neq 0$. For linearly separable data solved using the linearly separable version of the SVM Lagrangian, the support vectors are the vectors that lie on the margin boundary (for nonlinear kernels, this applies in the expanded dimensional feature space). For datasets that are not linearly separable in the expanded dimensional feature space, and solved using the not-linearly-separable version of the SVM Lagrangian, the support vectors are the vectors that lie on the margin boundary or on the wrong side of the margin boundary. (You can infer this behavior from the KKT conditions in the SVM solutions (c.f. lecture notes).)

3. The optimization problem for SVM learning, linearly separable (in $\underline{u}$-space) case, can be stated as:

$$\text{Minimize } J(\underline{w}) = \tfrac{1}{2}\|\underline{w}\|^2$$
$$\text{s.t. } z_i\left(\underline{w}^T\underline{u}_i + w_0\right) - 1 \geq 0 \quad \forall i$$

Use **un**augmented notation in this problem.

(a) What does the set of constraints (second line of equations above) mean?

(b) Start from the primal Lagrangian function $L(\underline{w}, w_0, \underline{\lambda})$ for the minimization problem stated above. (You can get it from Lecture 14 notes, linearly separable case.)

Derive the dual Lagrangian L_D, by proceeding as follows:

(i) Minimize L with respect to (w.r.t.) the weights.
Hint: solve $\nabla_{\underline{w}} L = \underline{0}$ for the optimal weight $\underline{w}^*$ (in terms of λ_i and other variables); and set $\dfrac{\partial L}{\partial w_0} = 0$ and simplify.

(ii) Substitute your expressions from part (i) into L, and use your expression from $\dfrac{\partial L}{\partial w_0} = 0$ as a new constraint, to derive L_D as:

$$L_D(\underline{\lambda}) = -\tfrac{1}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j z_i z_j \underline{u}_i^T \underline{u}_j\right] + \sum_{i=1}^{N}\lambda_i$$

subject to the (new) constraint: $\sum_{i=1}^{N}\lambda_i z_i = 0$. Also give all the other KKT conditions (include the conditions that carry over from the primal form). **Hint:** you should have 5 sets of KKT conditions total.

4. [**Extra credit.**] Now consider SVM for data that is not linearly separable in $\underline{u}$-space. We stated the optimization problem as:

$$\text{Minimize } J(\underline{w}) = \tfrac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } z_i\left(\underline{w}^T\underline{u}_i + w_0\right) \geq 1 - \xi_i \quad \forall i$$
$$\xi_i \geq 0 \quad \forall i$$

Tip: Write clearly so that μ_i and $\underline{u}_i$ are always distinguishable.

(a) Write the Lagrangian function $L\left(\underline{w}, w_0, \underline{\xi}, \underline{\lambda}, \underline{\mu}\right)$ for the minimization problem stated above. Use $\lambda_i, i = 1, 2, \cdots, N$ for the Lagrange multipliers relating to the first set of constraints, and use $\mu_i, i = 1, 2, \cdots, N$ for the Lagrange multipliers relating to the second set of constraints State the KKT conditions. **Hint:** there are 6 KKT conditions: 4 sets of restrictions on the λ_i and $\underline{u}_i$, and the 2 given sets of constraints.

(b) Derive the dual Lagrangian L_D, by minimizing L w.r.t. $\underline{w}$, w_0 , and $\underline{\xi}$, and give the formula for $\underline{w}^*$. Also list the restrictions and constraints on the Lagrange multipliers ($\underline{\lambda}$ and $\underline{\mu}$): there are 5 sets of them.

Hint: Each ξ_i can be treated as another independent variable, so that $\dfrac{\partial \xi_i}{\partial \underline{w}} = 0$.