Instructions: This assignment covers topics of classification.

Submission Requirements:

In *w3*, this part, you will be asked to walk through methods and calculations **manually** or display understanding of concepts and topics. In *a3-R* or *a3-python*, you will use packages and libraries to implement classification models, data reduction and text mining methods.

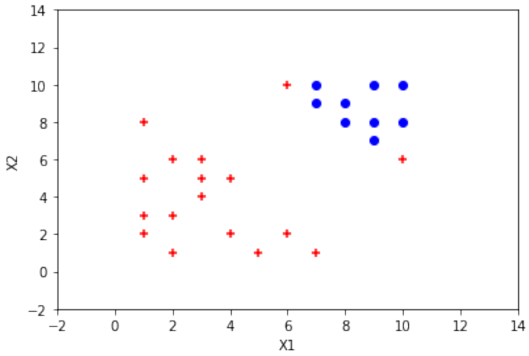
You must prepare your solution to this part as a document using LaTeX. I have provided a template for this assignment, where you can create your answers.

For this assignment, you are to work in your **groups**. I highly suggest using Overleaf to work on your submission together in your group.

Follow the submission template where work for each question must **start on a new page**. Do not put work for multiple problems on the same page.

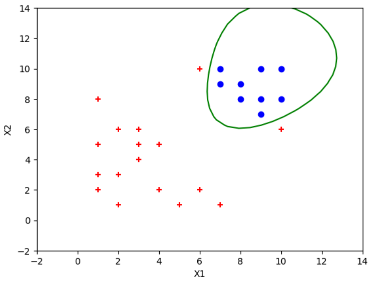
Questions:

1. (15 points) Support Vector Machines: Consider the following data set consisting of two variables $X1$ and $X2$ and two classes: plus (red) and circle (blue).

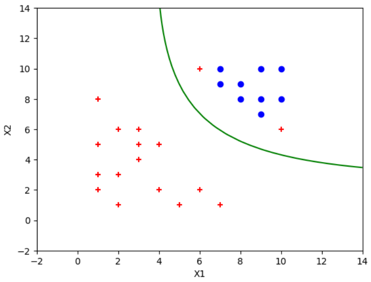


Each of the figures below show the data along with a SVM decision boundary. For each figure, select the parameters of the SVM and SVM kernel that was used to create the decision boundary. Note, for some figures, there may be more than one that are possible.

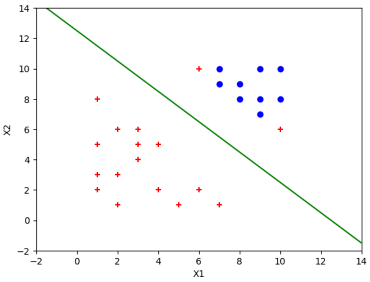
I. linear kernel, C=1
II. linear kernel, C=1000
III. polynomial kernel d=2, C=1
IV. polynomial kernel d=2, C=1000
V. rbf kernel, C=1
VI. rbf kernel, C=1000



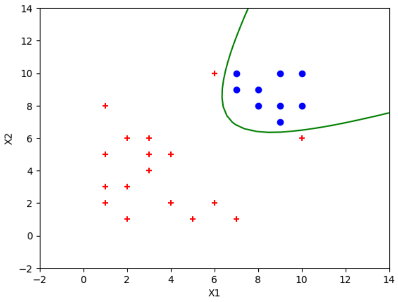
A. ______________



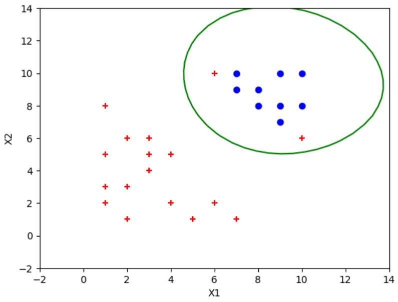
B. ______________



C. ______________



D. ______________



E. ______________

2. Random Forests and Cross-validation
   A random forest is created from a number of decision trees. A hyperparameter for random forests is the number of decision trees in the random forest.

   Let T be the number of decision trees in the random forest we create. We will consider three values for T: α, β, and γ. We want to use δ-fold cross-validation to tune the hyperparameter (find best value of T). The values for α, β, γ, and δ are integers.

   (a) (4 points) When we perform this cross-validation process, how many **random forests** will we train? Answer in an expression that can be a combination of terms integers, operators $(+, -, *, /)$, and α, β, γ, and/or δ.

   (b) (4 points) When we perform this cross-validation process, how many **decision trees** will we train? Answer in an expression that can be a combination of terms integers, operators, and α, β, γ and/or δ.

3. (9 points) Ensemble Methods

   Assume we have created 10 bootstrapped samples from a data set with two classes blue and green.

   A classification tree is learned for each bootstrap sample. A test sample, $\mathbf{x}_{test}$, is applied to each learned tree and the model produces an estimate of: $P(\text{Class is blue} \mid \mathbf{x}_{test})$

   Therefore, for the 10 bootstrap samples / models, you have the following 10 probabilities:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, 0.75$$

   The estimates can be combined into a single prediction for the ensemble approach using two ways:

   - **majority vote** - the most commonly occurring class among the predictions; for this problem assume a threshold of 0.5 is used to determine the class (blue / green).
   - **average probability** - average the probabilities for each sample/model. Assume a threshold of 0.5 on this probability to determine the class (blue / green)

   What is the final classification (blue / green) under the two approaches?

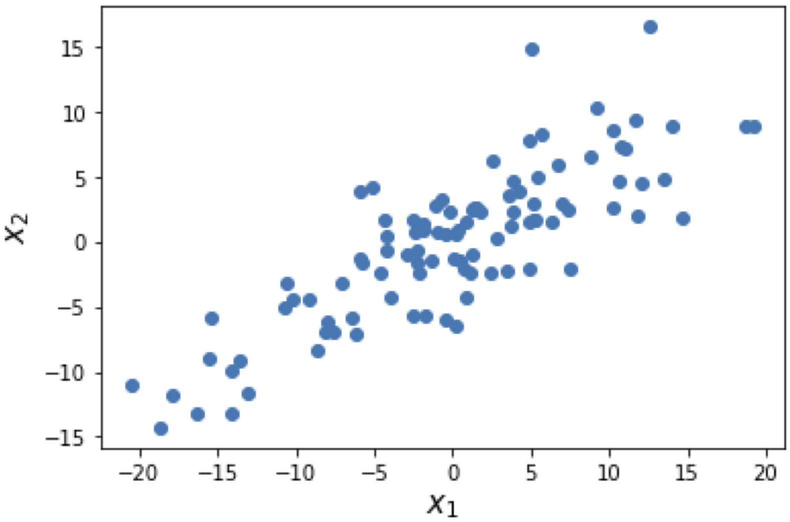
   Majority Vote (blue/green): ____________

   Ave. Prob. (prob.): ____________

   Ave. Prob. Prediction: (blue/green): ____________

4. (8 points) PCA

Suppose you were given the following 2-dimensional data set. We want to perform PCA on this data.



Given the following equations of lines. Write the letter of the line (A - I) that is most likely representing the direction of PC1 and PC2.

(A) $x_2 = \frac{11}{3}x_1 - 9$

(B) $x_2 = 3x_1$

(C) $x_2 = -\frac{20}{3}x_1 + 5$

(D) $x_2 = \frac{2}{3}x_1$

(E) $x_2 = -3x_1$

(F) $x_2 = -\frac{3}{2}x_1$

(G) $x_2 = -4x_1 + 10$

(H) $x_2 = \frac{1}{3}x_1 + 5$

(I) $x_2 = -\frac{2}{3}x_1$

PC1: ____________

PC2: ____________