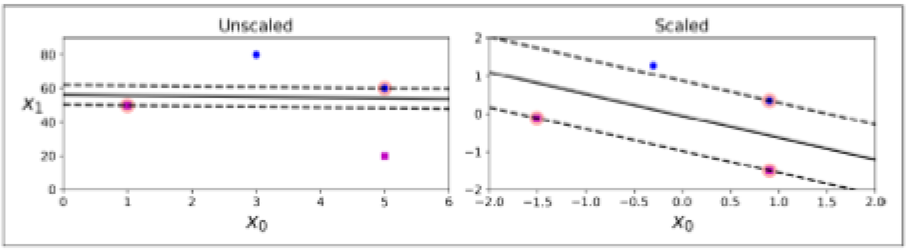
- Please type the solutions using a word processor such as MS Word, Latex, or write by hand neatly and upload the scanned copy of it.
- **I, Andrick (sign your name here), guarantee that this homework is my independent work and I have never copied any part from other resources. Also, I acknowledge and agree with the plagiarism penalty specified in the course syllabus.**
- Turn in your assignment before the deadline. Penalty will be applied to late submission.

A. <u>Exercise</u>:

1. (10pt) What is the fundamental idea behind Support Vector Machines?

2. (10pt) What is a support vector?

3. (10pt) Why is it important to scale the inputs when using SVMs? To better help you understand the question, here are two different results when we scale the data or not. Try to figure out which one is better and why?



4. (10pt) Can an SVM classifier output a confidence score when it classifies an instance? What about a probability?

5. (10pt) For the soft-margin linear SVM, we use the hyperparameter C to tune how much we want to penalize misclassifications. As C goes to infinite, does the soft-margin SVM become more similar or less similar to the hard-margin SVM? As C decrease to 0, what happens to the solution of the soft-margin SVM? Why?

6. (10pt) You had a dataset which does not seem to be linear separable at the first glance. Therefore, you trained an SVM classifier with an RBF kernel. Unfortunately, it seems to underfit the training set. Now, should you increase or decrease gamma? What about C? Explain your thoughts.

1. The fundamental idea of support vector machines is to find correlations between multiple data, for instance, let's say we can classify if a person is likely to get a car or not, given their age, credit score, and income, this case we would have 3 dimensions but it would still be able to separate the data in a way(the way im referring to is raising the current dimension one higher) where there are various lines depending on the data distribution, that indicate classifier, in other words, if a testing data point is behind boundary the corresponding hyperplane/margin if it is it classifies

it as such, and it misclassifies data in order to produce the best margins. Overall it separates the data (x and y where y is the target class) with the best margin calculated with the support vectors.

2. Support vectors are essentially what makes support vector machines work, as it is the observations of the data on the edge within the soft margin. In other words, support vectors are the data points that yield the best margin that makes up support vector machines. Finally, if not scaled SVM will most likely neglect small values in the dataset, it essentially separates the data (x and y where y is the target class) and allows for classification of data to happen.
3. If you do not scale all features to comparable ranges, the features with the largest range will completely dominate in the computation of the kernel matrix, and thus small values could be neglected. In other words, we scale in order to keep all data in a similar range to others in the same space.
4. It can output the distance between the test instance and the decision boundary, which can be used as the confidence score. This score can't be directly converted into a probability. Hence why we don’t have a probability score for this algorithm, because we don’t really have a way of creating a probability since everything to the right or left of a margin is a different class no probability at least for 2-dimensional data.
5. As C goes to infinity we have fewer misclassifications which means it would behave similar to a hard-margin SVM, and as C decreases to 0, the margin will be essentially at an optimal position where it separates the data with a margin that neglects the most misclassifications while keeping the style of a soft-margin SVM. Essentially even if your data is linear separable your still going to get misclassifications.

6. For this dataset based on the information given we would increase gamma and or C, because of the regularization, and by doing this we make the margin more curved and the smaller C wider margin which leaves us with misclassifications and hence make the data more spread and hopefully make fit.

Bonus Question:

- Gradient descent is 0 here because it found a value to converge to, in this case, it looks like 40, and what this number represents is the local minimum. This is consistent because we see the correlation between margin from support vector machines and gradient descent which in this case converges to a local minimum.