1 SVM

A) Assume that we have a dataset containing M points in N dimensions (i.e. each datapoint is described by N floats). After investigating the data and finding optimal hyperparameters, we train a Support Vector Machine classifier with Gaussian RBF kernel, that results in S support vectors.
Question 1: How many floats do we need to store the trained model? Assume that all numbers in the model are stored in floats.
Question 2: Assume that each float takes 8 bytes in memory, the data is 100-dimensional (i.e. $N = 100$), and amount of support vectors $S = 10,000$. How much memory do we need to store the model?
Question 3: Let's assume that in previous question only 1% of all datapoints became support vectors, meaning that total amount of data $M = 1,000,000$. The training time complexity for SVM is $O(MN^2)$. You train a smaller problem with $M = 1000$ and $N = 10$, and training time amounts to 0.1 second. How much time do you approximately need to train the classifier for the initial problem?
Question 4: Average modern laptop CPU requires 50W of power under full load. Using the time from previous question, how much energy (in kWh) does one need to train such SVM? For comparison: regular kettle draws 1500 W, and it takes 5 minutes to boil water in it (approx 2 liters). How many kettles boiled is equivalent to training the SVM model in question?

B) Consider a 2-dimensional classification problem with only 2 data points at $\mathbf{x}^1 = [-0.5, -0.5]$ and $\mathbf{x}^2 = [0.5, 0.5]$, with class labels $+1$ and -1 respectively (See Figure 1). Compute the weights α_i and the bias b for a SVM classifier run on this problem with a Gaussian RBF kernel such that $k(\mathbf{x}^1, \mathbf{x}^2) = 0.5$. Also, draw the isolines of the classifier function and the classifier hyperplane. Recall that the classifier function is given by:

$$f(\mathbf{x}) = \text{sign}\left(\sum_i \alpha_i y_i k(\mathbf{x}, \mathbf{x}^i) + b\right) \tag{1}$$

and that we have the necessary conditions for optimality:

$$\begin{cases} w = \sum_i \alpha_i y_i x^i \\ \sum_i \alpha_i y_i = 0 \\ y_i(\sum_j \alpha_j y_j k(x^j, x^i) + b) \geq 1, \ \forall i = 1..M \text{ (primal feasibility)} \\ \alpha_i \geq 0, \ \forall i = 1..M \text{ (dual feasibility)} \\ \alpha_i(y_i(\sum_j \alpha_j y_j k(x^j, x^i) + b) - 1) = 0, \ \forall i = 1..M \text{ (KKT cond.)} \end{cases} \tag{2}$$

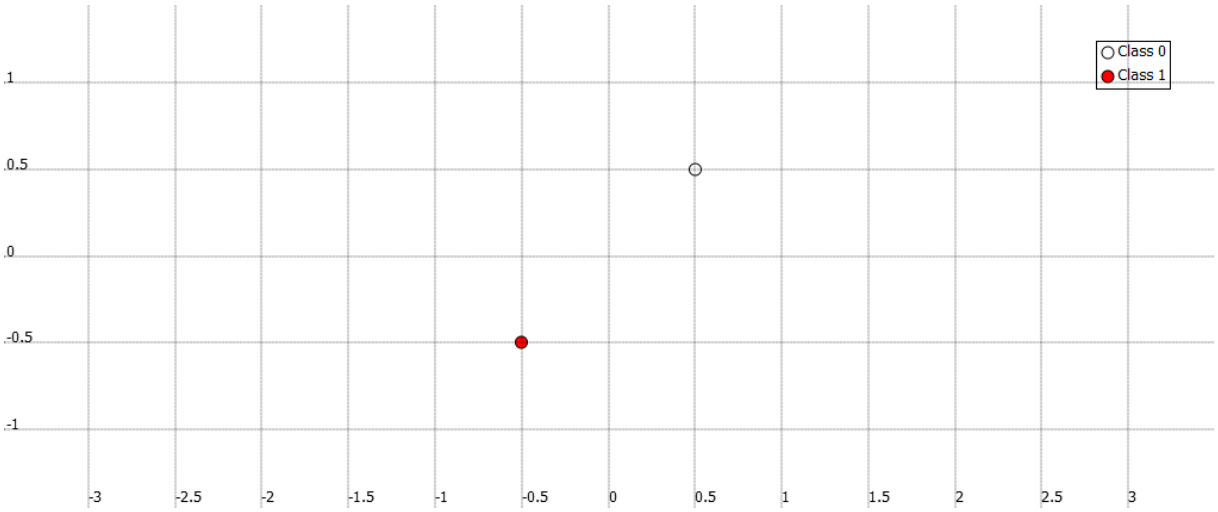


Figure 1

C) Let us add two more points to this problem in different ways as shown in Figures 2-3. How would the α_i and b change in each case? Draw the support vectors, the classifier function and the boundary in each case.

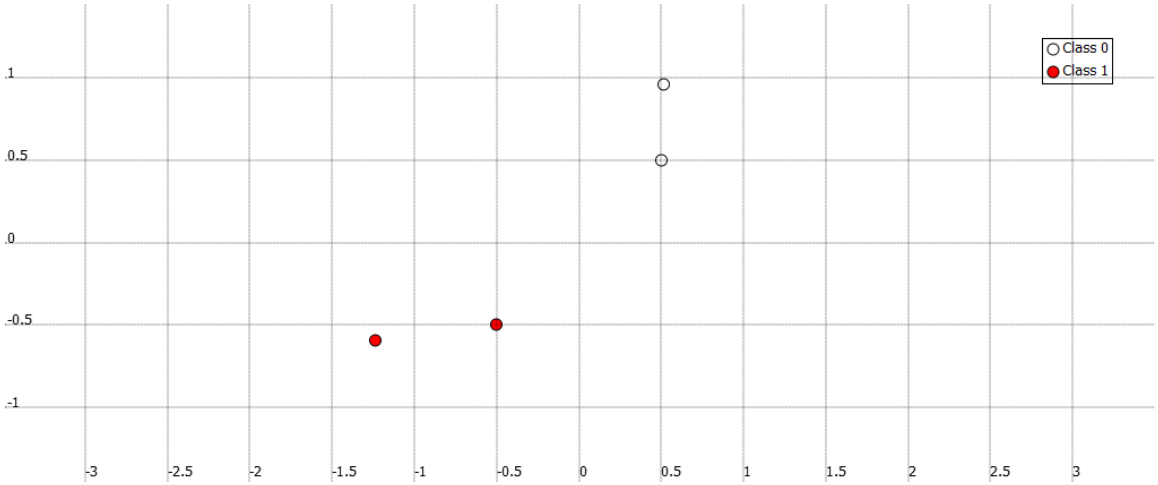


Figure 2

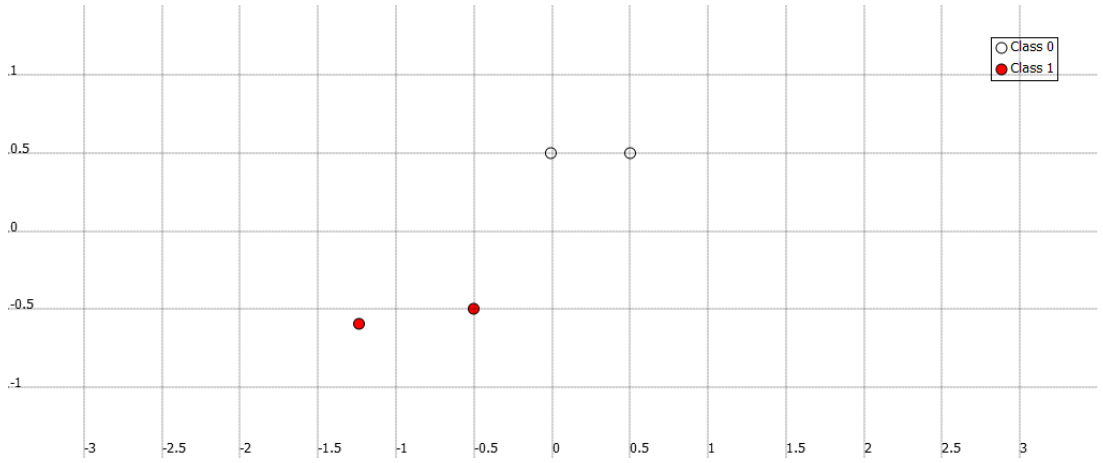
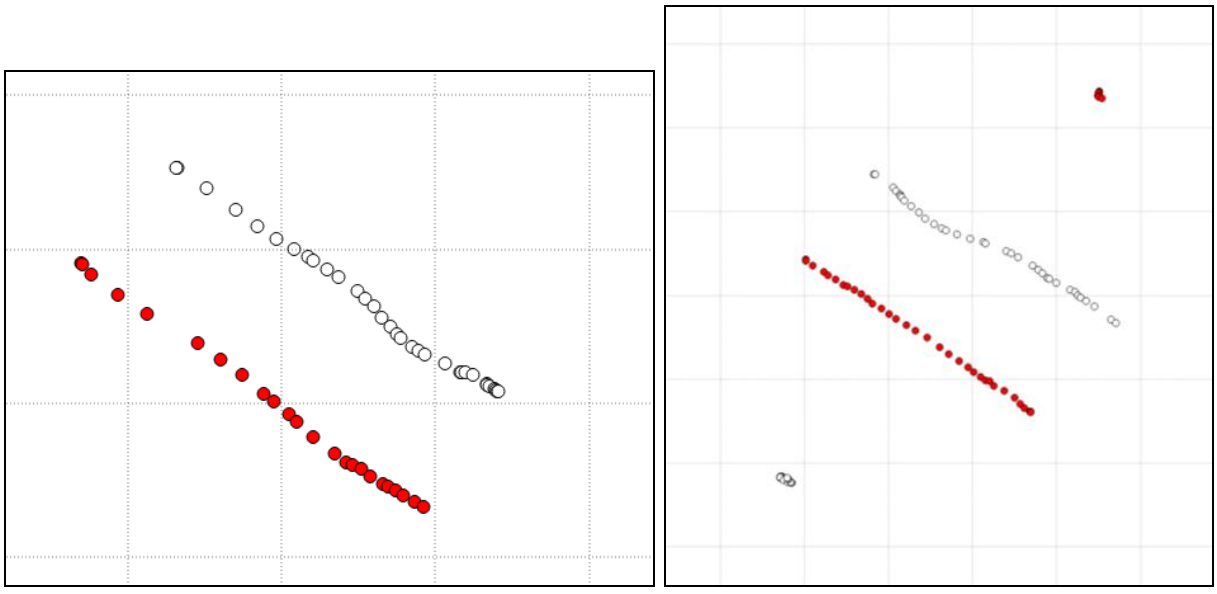


Figure 3

D) Consider the two-class classification problems (with dark and white classes) shown in Figure 4. Assume that you have run C-SVM with a RBF kernel. Draw what would be the separating line in each case (do not compute it nor run mldemos, but try to infer what it would look like from your intuition). Discuss how this line changes as a function of the value of the penalty factor C and of the kernel width (from very big to very small C and kernel width respectively):



(i) **(ii)**

Figure 4

2 Optimization of SVM

A: Convex Optimization: multiplicity of solution in SVM SVM is based on solving a convex optimization problem, where the objective function $\|w\|^2$ is strictly convex. As discussed in the lecture, while the convex problem admits a single global optimum and hence leads to a unique vector w $in\Re^N$, there can be multiple ways in which w is constructed. Indeed, w is constructed as a linear combination of support vectors. If one has at disposal a set of K linearly independent support vectors with $K > N$, then there exists more than one combination of scalars $\alpha_i, i = 1...K$, such that $w = \sum_i = 1^K \alpha_i x^i$ is not unique.

Convince yourself that this is the case when considering the linear SVM case, assuming that $N = 2$ and that you have at your disposal 3 non-zero and not-collinear vector point $x^i, i = 1, 2, 3$ that satisfy the constraint $y_i(w^T x^i + b) = 1, \forall i$. Show that there exist another combination of points that can construct w.

B: Margin The constraints of the SVM problem specify that all support vectors should lie on either of the two hyperplanes parallel to the separating hyperplane with equations $w^T x + b = \pm 1$. Show that the constant 1 is arbitrary and does not affect the solution.

C: Convexity of the relaxed problem: Is $f(w, \xi) = \|w\|^2 + \sum_i \xi$, $\xi > 0 \forall i$ convex?

D: Optimum of the relaxed problem: The introduction of slack variables in the SVM optimization allows to find a solution to a problem that would otherwise been deemed infeasible. The drawback is that the slacks lead to solutions that are "suboptimal". Note that the problem remains convex, but the slacks shift the optimum to a value different from the true optimum.

Prove first that the problem remains convex. Recall the conditions for convexity and strict convexity: a convex function f is such that $f(\lambda x + (1 - \lambda)y \leq f(\lambda x) + f((1 - \lambda)y)$. Strict convexity arise when the inequality is replaced by a strict inequality ($<$ in place of $\leq$).

Prove that the optimum in the relaxed problem is identical to the original problem only under certain conditions for the linear SVM problem.