1 Computational Aspects of SVM

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 (SVM) classifier with an RBF Kernel that results in S support vectors.

A) How many parameters are required to store the trained model?

B) Assume that each parameter has a float data type that takes 8 bytes in memory. What is the required memory to store the model if dataset is 100-dimensional (i.e., $N = 100$) and the number of support vectors is $S = 10,000$?

C) Assume that, in the previous question, only 1% of all datapoints became support vectors, meaning that total number of datapoints is $M = 1,000,000$. The training time complexity for SVM is $O(MN^2)$. Furthermore, assume that training in a smaller problem with $M = 1000$ and $N = 10$ takes 0.1 second. How much time do we approximately need to train the classifier for the initial problem?

D) Average modern laptop CPU requires 50W of power under full load. Using the time from the previous question, how much energy (in Wh) does one need to train such SVM? For comparison, note that a regular kettle draws 1500W, and it takes 5 minutes to boil approximately 2 liters of water. Training this SVM model is equivalent to boiling how many liters of water?

2 Classification with SVM

A) Consider a 2-dimensional classification problem with only 2 datapoints, including $\boldsymbol{x}^1 = [0.5, 0.5]^\top$ and $\boldsymbol{x}^2 = [-0.5, -0.5]^\top$ with $+1$ and -1 class labels, respectively (see fig. 1). Compute the coefficients α_i and the bias term b for a SVM classifier run on this problem with an RBF kernel where $k(\boldsymbol{x}^1, \boldsymbol{x}^2) = 0.5$ ($\phi(\cdot)$ is the corresponding feature map). Moreover, draw the isolines of the classifier function and the classifier hyperplane.

Hint: Recall that the SVM classifier function is given by

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

Furthermore, the necessary conditions for optimality are provided below.

$$\begin{cases} \boldsymbol{w} = \sum_i \alpha_i y_i \phi(\boldsymbol{x}^i) \\ \sum_i \alpha_i y_i = 0 \quad \text{(appearing in the dual problem)} \\ y_i\left(\sum_j \alpha_j y_j k(\boldsymbol{x}^j, \boldsymbol{x}^i) + b\right) \geq 1, \quad \forall i = 1, \ldots, M \quad \text{(primal feasibility)} \\ \alpha_i \geq 0, \quad \forall i = 1, \ldots, M \quad \text{(dual feasibility)} \\ \alpha_i\left(y_i\left(\sum_j \alpha_j y_j k(\boldsymbol{x}^j, \boldsymbol{x}^i) + b\right) - 1\right) = 0, \quad \forall i = 1, \ldots, M \quad \text{(KKT condition)} \end{cases}. \tag{2}$$

B) Two more points are added to this dataset in different ways as illustrated in figs. 2 and 3. How would the α_i and b parameters change in each case? Draw the support vectors and the classification boundary for each case.

C) Consider the binary classification problem among red and white classes shown in fig. 4. For the case of SVM with an RBF kernel, draw the separating line in each case. Do not compute it nor run

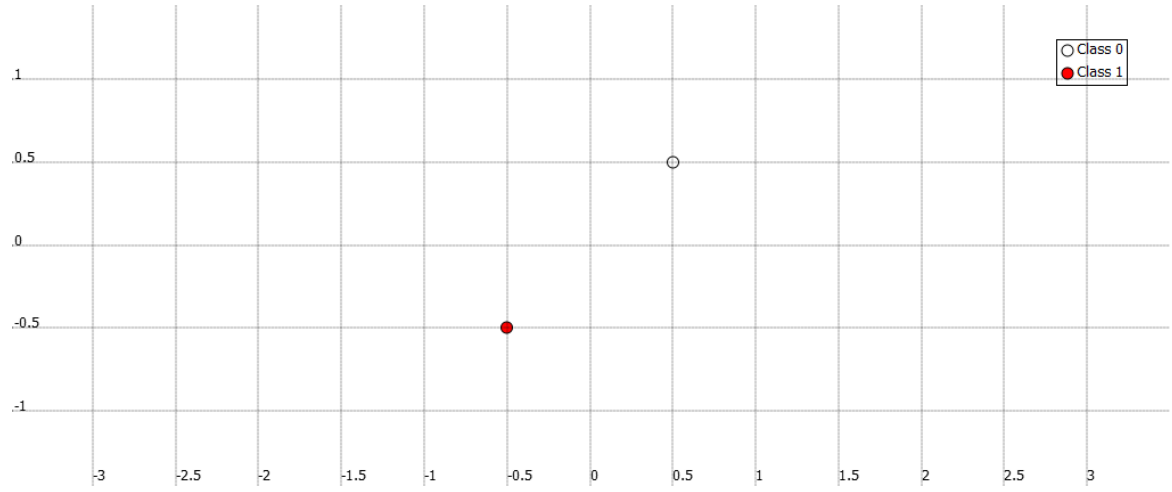


Figure 1: Question 2.A

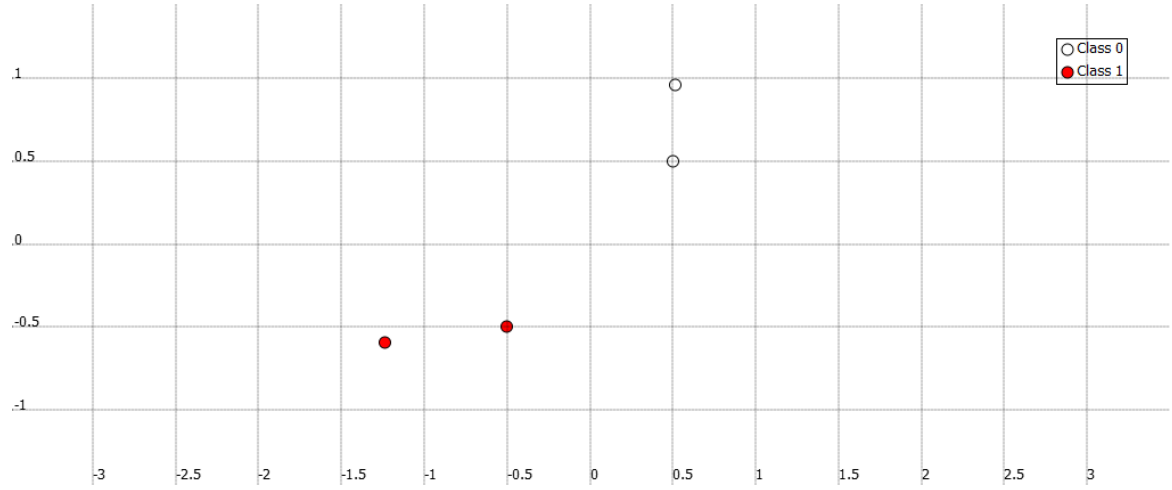


Figure 2: Question 2.B - Case (1)

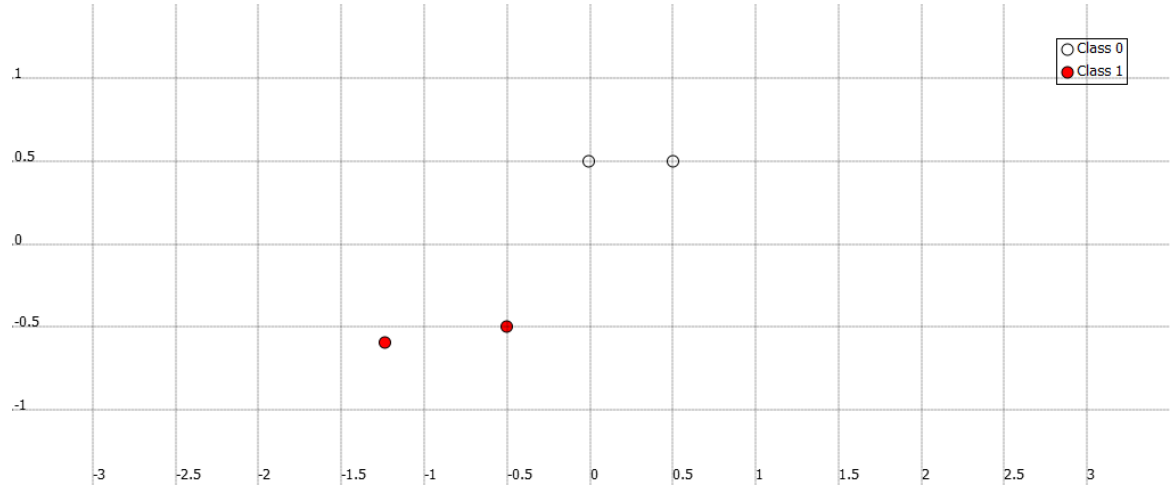
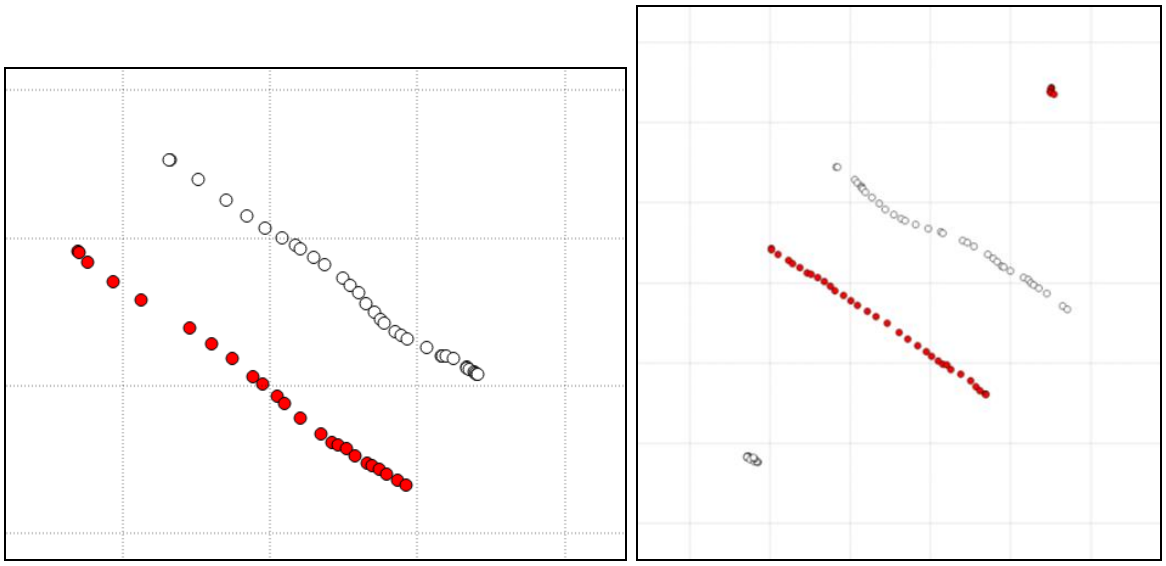


Figure 3: Question 2.B - Case (2)

MLDemos; instead, infer what the line would look like from your intuition. Discuss how this line changes as a function of the penalty factor C and the kernel width σ.



(i) **(ii)**

Figure 4: Question 2.C

3 Optimization of SVM

A: Convex Optimization - Multiplicity of Solutions in SVM) SVM is based on solving a convex optimization problem, where the objective function $\|\boldsymbol{w}\|^2$ is strictly convex. As discussed in the lecture, while the convex problem admits a single global optimum and hence leads to a unique vector $\boldsymbol{w} \in \mathbb{R}^N$, there can be multiple ways in which $\boldsymbol{w}$ is constructed. Indeed, $\boldsymbol{w}$ is constructed as a linear combination of support vectors. If one has at disposal a set of K support vectors with $K > N$, these vectors are linearly dependent. Therefore, there exists more than one combination of scalars $\alpha_i,\ \forall i = 1, \ldots, K$, yielding the same $\boldsymbol{w}$ constructed as $\boldsymbol{w} = \sum_{i=1}^{K} \alpha_i y_i \boldsymbol{x}^i$.

Convince yourself that this is the case when considering linear SVM for binary classification, assuming that $N = 2$ and that you have at your disposal 3 non-zero and non-collinear datapoints $\boldsymbol{x}^i,\ i = 1, 2, 3$ that satisfy the constraint $y_i(\boldsymbol{w}^\top \boldsymbol{x}^i + b) = 1,\ \forall i$. Show that there exists another combination of points that can construct the same $\boldsymbol{w}$.

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 $\boldsymbol{w}^\top \boldsymbol{x} + b = \pm 1$. Show that the constant 1 is arbitrary and does not affect the solution.

C: Convexity and Optimality of the Relaxed Problem) The introduction of slack variables in the SVM optimization problem allows to find a solution to the problem that would otherwise been deemed infeasible. The drawback is that the slack leads to solutions that are suboptimal in a sense that it allows violations of the strict constraints in the unrelaxed problem. Note that the problem remains convex, but the slacks shift the optimum to a value different from the true optimum.

Prove first that the relaxed problem remains convex. Recall the conditions for convexity and strict convexity: a convex function f is such that $f(\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}) \leq f(\lambda \boldsymbol{x}) + f((1 - \lambda)\boldsymbol{y})$ for $0 \leq \lambda \leq 1$. Strict convexity arises when the inequality is replaced by a strict inequality ($<$ in place of $\leq$) for $0 < \lambda < 1$ and $\boldsymbol{x} \neq \boldsymbol{y}$.

Secondly, explain under which conditions the optimum in the relaxed problem is identical to the original problem for linear SVM.