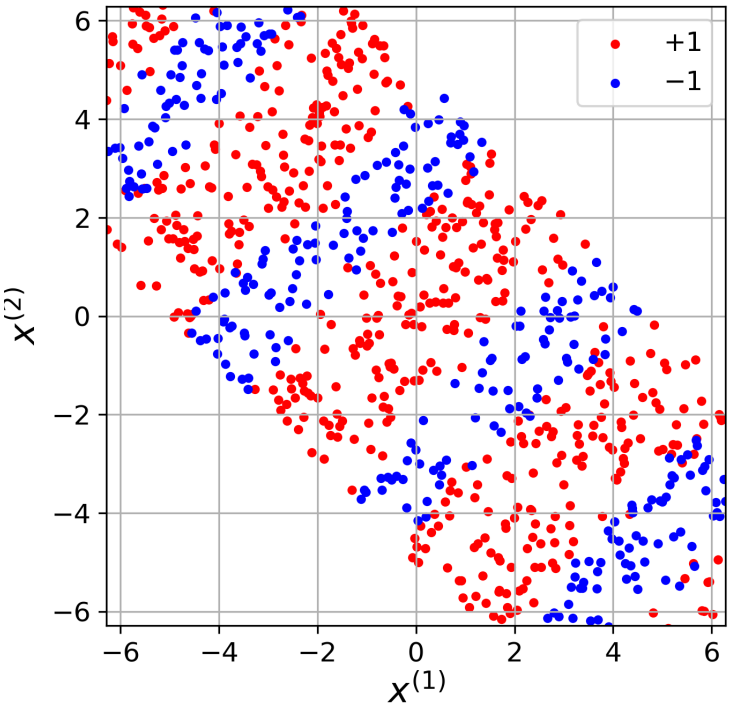
Fifth assignment in Machine learning 1 – 2023 – Paper 1

1 Maximum marginal classifier (Deadline: 28-th October)



Assume a dataset $\mathcal{D} = \{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$ where $\boldsymbol{x}_n \in \mathbb{R}^2$ and $t_n \in \{-1, +1\}$. Upon inspection of the data, we note a periodic pattern in the data (up to some exceptions). See the Figure above for an illustration.

To account for this periodicity, it seems reasonable to use some *trigonometric* functions to classify the data. In particular, we build classifier which will pick the class $t_n = +1$ or -1 depending on the sign of the following function:

$$y(\boldsymbol{x}_n) = w_0 \cos(x_n^{(1)} - x_n^{(2)}) + w_1 \sin(x_n^{(1)} - x_n^{(2)}) - \beta\,,$$

with $\boldsymbol{w} = (w_0, w_1)^T \in \mathbb{R}^2$, $\beta \in \mathbb{R}$ and $\boldsymbol{x}_n = (x_n^{(1)}, x_n^{(2)})^T \in \mathbb{R}^2$. To simplify our notation, we define the function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ as

$$\phi(\boldsymbol{x}_n) = \begin{bmatrix} \cos(x_n^{(1)} - x_n^{(2)}) \\ \sin(x_n^{(1)} - x_n^{(2)}) \end{bmatrix},$$

such that $y(\boldsymbol{x}_n) = \phi(\boldsymbol{x}_n)^T \boldsymbol{w} - \beta$. Under the assumption that the two classes are perfectly separable, the distance between the decision boundary and any datapoint $\boldsymbol{x}_n$ is given by:

$$\forall n, \quad \frac{t_n\big(\phi(\boldsymbol{x}_n)^T \boldsymbol{w} - \beta\big)}{||\boldsymbol{w}||_2} \geq 0\,. \tag{1}$$

As usual, it is convenient to leverage the scale invariance of this distance. Indeed,

$$\frac{\alpha}{\alpha} \cdot \frac{t_n\big(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - \beta\big)}{||\boldsymbol{w}||_2} = \frac{t_n\big(\phi(\boldsymbol{x}_n)^T(\alpha\boldsymbol{w}) - \alpha\beta\big)}{||\alpha\boldsymbol{w}||_2} \,.$$

Because of the arbitrary choice of α, we are free to decide that we want to find a solution $\boldsymbol{w}$ such that $t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) = 1$ for those points closest to the decision boundary, with $b = \alpha\beta$. In this case, the following constraint holds for all datapoints:

$$\forall n, \quad t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) \geq 1 \,.$$

Unfortunately, based on the Figure above, it does not seem that the data is perfectly separable, hence, we will introduce slack variables. We will now state the primal program that will find the optimal decision boundary:

$$\begin{aligned} \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} & \frac{1}{2}||\boldsymbol{w}||_2^2 + C\sum_{n=1}^{N}\xi_n \,, \\ \text{s.t. } & \forall n : t_n(\phi(\boldsymbol{x}_n)^T\boldsymbol{w} - b) \geq 1 - \xi_n, & (i) \\ & \forall n : \xi_n \geq 0 \,. & (ii) \end{aligned}$$

Answer the following questions:

(a) What does the dataset $\mathcal{D}$ look like after applying transformation ϕ? You do not have to provide a graphical answer, a description with words suffices. *Hint:* Think of ϕ as a composition of two transformations $\phi = \phi_2 \circ \phi_1$ with $\phi_1(\boldsymbol{x}) = [x^{(1)} - x^{(2)},\, x^{(1)} - x^{(2)}]^T$ and $\phi_2(\boldsymbol{x}) = [\cos(x^{(1)},)\, \sin(x^{(2)})]^T$. [1 point]

(b) How does the decision boundary $y(\boldsymbol{x}) = 0$ look in the new, transformed space? How does the vector $\boldsymbol{w}$ affect the decision boundary? What about b? [1 point]

(c) Introduce Lagrange multipliers for the constraints and write down the primal Lagrangian. You should use the target function $\frac{1}{2}||\boldsymbol{w}||_2^2$ (note that you can answer this question even if you did not find a solution to the previous ones). Use the following notation: $\{\lambda_n\}$ are the Lagrange multipliers for the first constraint (i) and $\{\mu_n\}$ for the second (ii). [1 point]

(d) How many KKT conditions are there? Write down all KKT conditions (not the *stationary conditions*). [2 points]

(e) Derive the dual Lagrangian and specify the dual optimization problem. That is, derive the *stationary conditions* (by computing $\partial\ell/\partial\rho = 0$, where ℓ is the primal Lagrangian and ρ is a primal variable) and use the resulting identities to eliminate the primal variables $\{\boldsymbol{w}, b, \xi_1 \ldots, \xi_N\}$. Do not forget to specify the constraints on the remaining dual variables. [3 points]

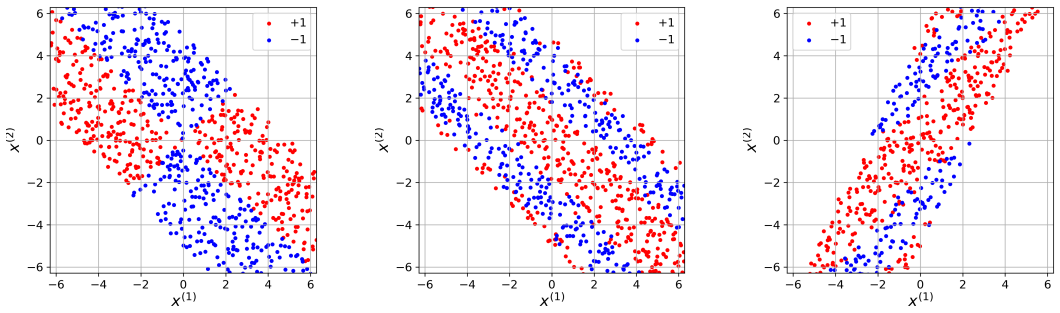
(f) Note that, because we have applied the transformation ϕ, we have already written down a kernelized dual Lagrangian. This results in a dual Lagrangian that no longer depends on $\boldsymbol{x}_n^T \boldsymbol{x}_m$ but on $\kappa(\boldsymbol{x}_n, \boldsymbol{x}_m)$. What is the explicit form of $\kappa(\boldsymbol{x}_n, \boldsymbol{x}_m)$ in your final solution to the dual lagrangian? [1 point]

(g) Assume we have solved the dual program and found optimal values for λ_n and b. Now we want to apply our maximum margin classifier, to a new test case $\boldsymbol{x}^\star$. Describe how to classify the new datapoint $\boldsymbol{x}^\star$ in dual space. [1 point]

(h) Use the KKT conditions to derive for which values of λ_n, μ_n and ξ_n, a data point $\boldsymbol{x}_n$ lies (i) outside the margin and is correctly classified, (ii) on the margin (i.e., is support vector), (iii) on the right side of the decision boundary within the margin, and (iv) on the wrong side of the decision boundary. [2 points]

(i) Find the optimal values for the other dual variables μ_n assuming we have found optimal λ_n. Then, solve for the primal variables $\{\boldsymbol{w}, b, \xi_1 \ldots, \xi_N\}$. Note that you do not need to know the dual optimization program to solve this question. You only need the KKT conditions. *Hint*: Consider stationary conditions from question (e) when solving for optimal μ_n. [3 points]

(j) Now assume you obtain datasets with the following pattern. For each dataset, provide the function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ such that two classes are linearly separable after the transformation.



[3 points]