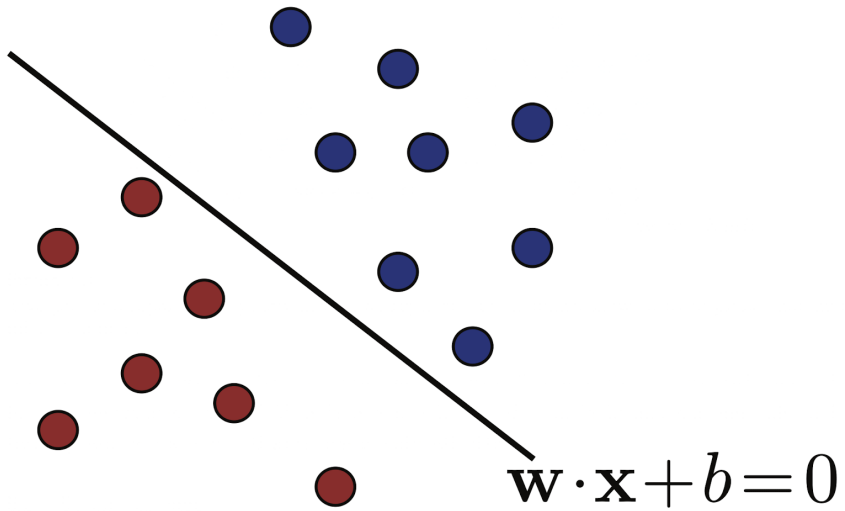
Support Vector Machine (SVM)

Linear Classifiers

- Given a set of labeled points $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \{-1, +1\}$, one of the easiest hypothesis class we can consider is the set

$$\mathcal{H} = \{x \mapsto \mathsf{sign}(w \cdot x + b) | w \in \mathbb{R}^n, b \in \mathbb{R}\}$$

- The expression $w \cdot x + b = 0$ defines an hyper-plane in $\mathbb{R}^n$ and the corresponding classifier assign a positive label to the points that belong to the half space $w \cdot x + b > 0$ and a negative label to the points that belong to the half space $w \cdot x + b < 0$



$\mathcal{X}$ perfectly separable

- Assume that the points in $\mathcal{X}$ are linearly separable, i.e., exists a pair $(w, b) \in \mathbb{R}^n \times \mathbb{R}$ such that

$$y_i(w \cdot x_i + b) \geq 0, \qquad \forall i \leq m$$

- If such a pair exists, then there are infinity hyper-plane that satisfy the same condition. In order to determine a unique solution we can use the following

Definition

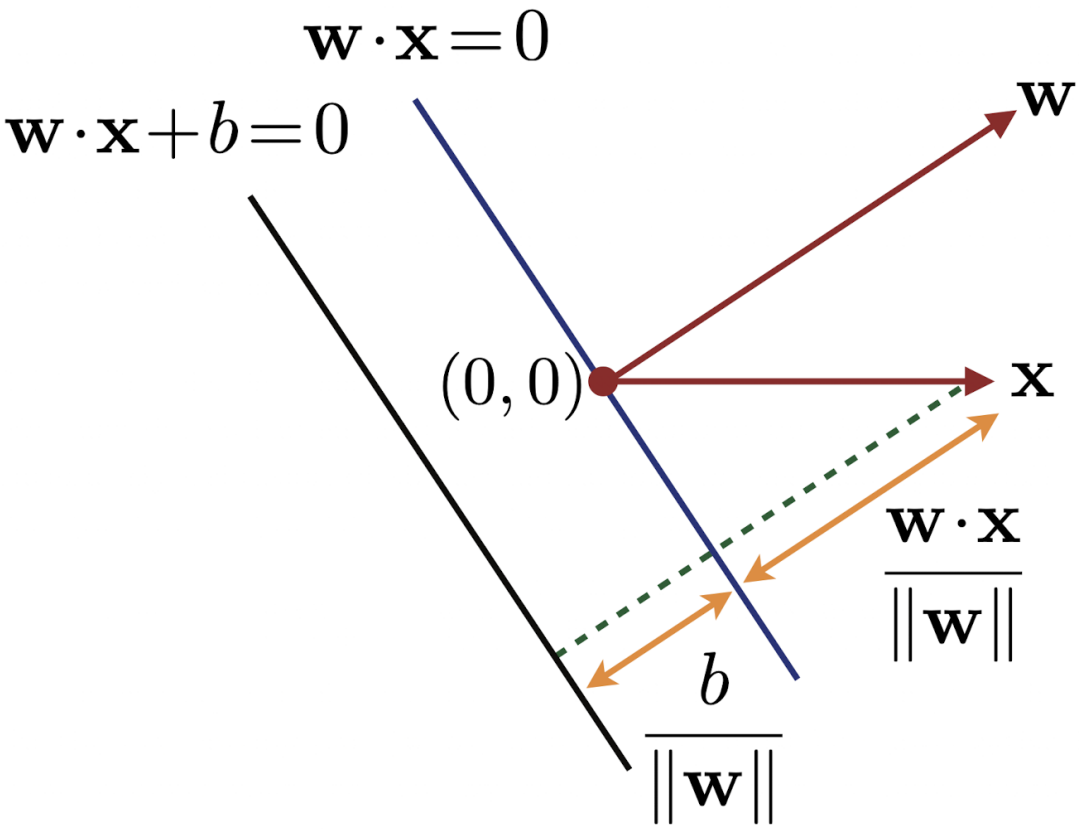
Given a linear classifier $f(x) = \mathsf{sign}(w \cdot x + b)$, the geometric margin of f in $x \in \mathbb{R}^n$ is the euclidean distance from x to the hyper-plane $w \cdot x + b = 0$:

$$\rho_f(x) = \frac{|w \cdot x + b|}{\|w\|_2}.$$

The geometric margin of the classifier f is the minimum value of the geometric margin of f along all the points in the input set $\mathcal{X}$: $\rho_f = \min_i \rho_f(x_i)$.

Maximum margin solution

- The Support Vector Machine (SVM) model aims at finding the hyper-plane with the maximum geometric margin among all the separating hyper-planes for $\mathcal{X}$



Primal Optimization problem for Hard-margin SVM

- The maximum geometric margin is given by

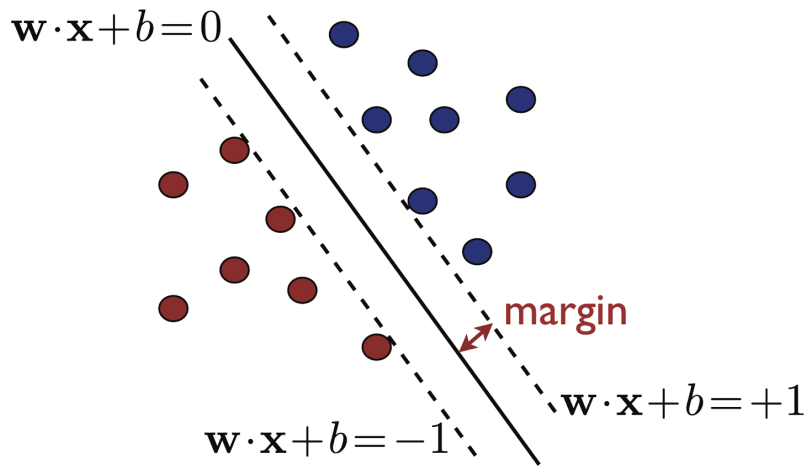
$$\rho = \max_{(w,b):y_i(w\cdot x_i+b)\geq 0} \left[\min_{i\leq m} \frac{|w\cdot x_i+b|}{\|w\|_2}\right] = \max_{(w,b)} \left[\min_{i\leq m} \frac{y_i(w\cdot x_i+b)}{\|w\|_2}\right]$$

  where the second equality follows from the fact that if the points are separable then $y_i(w\cdot x_i+b)\geq 0$

- The previous expression is invariant by scalar multiplication of (w,b), therefore we can consider hyper-planes satisfying $\min_i y_i(w\cdot x_i+b)=1$

$$\rho = \max_{(w,b):\min_i y_i(w\cdot x_i+b)=1} \left[\frac{1}{\|w\|_2}\right] = \max_{(w,b):y_i(w\cdot x_i+b)\geq 1,\forall i} \left[\frac{1}{\|w\|_2}\right]$$

Primal Optimization problem for Hard-margin SVM



- The marginal hyper-plan are defined by the expressions

$$w \cdot x + b = \pm 1$$

and they are parallel to the separating hyper-plane and passing by the closest points to it in the negative and positive half-spaces

Primal Optimization problem for Hard-margin SVM

- Maximizing $\frac{1}{\|w\|_2}$ is equivalent to minimizing $\frac{1}{2}\|w\|_2^2$, thus we can write the SVM optimization problem as

$$\begin{cases} \min\limits_{w,b} & \frac{1}{2}\|w\|_2^2 \\ s.t. & y_i(w \cdot x_i + b) \geq 1, \quad i \leq m \end{cases}$$

- Both the objective function and the constraints are convex, therefore the problem admits a unique optimal solution
- The objective function is a quadratic function in the variables and the problem is solvable by several commercial solvers (but not by using the simplex method)

Support Vectors

- Let $\alpha_i \geq 0$ be the Lagrangian multipliers associated to the constraints of the SVM problem and let $\alpha = (\alpha_1, \ldots, \alpha_m)$ the array of the multipliers.
- The corresponding Lagrangian function is given by

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{m} \alpha_i [y_i(w \cdot x_i + b) - 1]$$

- The KKT conditions are obtained by imposing the derivatives of $\mathcal{L}(w, b, \alpha)$ (w.r.t. w and b) equal to zero and by writing the complementarity conditions:

$$\begin{aligned}
\nabla_w \mathcal{L} = w - \textstyle\sum_i \alpha_i y_i x_i = 0 \quad &\Rightarrow \quad w = \textstyle\sum_i \alpha_i y_i x_i \\
\nabla_b \mathcal{L} = -\textstyle\sum_i \alpha_i y_i = 0 \quad &\Rightarrow \quad \textstyle\sum_i \alpha_i y_i = 0 \\
\forall i, \alpha_i [y_i(w \cdot x_i + b) - 1] = 0 \quad &\Rightarrow \quad \alpha_i = 0 \text{ o } y_i(w \cdot x_i + b) = 1
\end{aligned}$$

Support Vectors

- From the first KKT condition we have that w is a linear combination of the points x_i
- Every x_i with $\alpha_i \neq 0$ are called support vectors, and from the complementarity condition we obtain

  $y_i(w \cdot x_i + b) = 1 \Rightarrow$ the support vectors lay on the marginal hyper-planes

- Observe that even if the solution (w, b) is unique, the same is not necessarily true for the support vectors

Dual optimization problem for Hard-margin SVM

- In order to obtain the dual problem for SVM, we substitute the expression derived from the KKT conditions in the Lagrangian function:

$$\mathcal{L} = \frac{1}{2}\|\sum_i \alpha_i y_i x_i\|_2^2 - \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_i \alpha_i y_i b + \sum_i \alpha_i$$

$$= \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

- The dual optimization problem for SVM is

$$\begin{cases} \max\limits_{\alpha} & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ t.c. & \sum_i \alpha_i y_i = 0 \\ & \alpha_i \geq 0, \quad i \leq m \end{cases}$$

Observe that the dual problem depends only on the scalar product $x_i \cdot x_j$ and not on the single points x_i

Dual optimization problem for Hard-margin SVM

- Since we can use the Strong Duality Theorem for the SVM problem, we can determine the prediction function using α

$$f(x) = \mathsf{sign}(w \cdot x + b) = \mathsf{sign}\left(\sum_i \alpha_i y_i (x_i \cdot x) + b\right)$$

- If x_i is a support vector, then $w \cdot x_i + b = y_i$ and we can compute b as

$$b = y_i - \sum_j \alpha_j y_j (x_j \cdot x_i)$$

- It is also possible to determine the geometric margin of f using α:

$$\rho_f^2 = \frac{1}{\|w\|_2^2} = \frac{1}{\sum_i \alpha_i} = \frac{1}{\|\alpha\|_1}$$

$\mathcal{X}$ non separable

- Usually the points in $\mathcal{X}$ are not perfectly separable by an hyper-plane, i.e., exists at least one $x_i \in \mathcal{X}$ such that

$$y_i(w \cdot x_i + b) \ngeq 1$$

- We can introduce some slack variables $\xi = (\xi_1, \ldots, \xi_m) \geq 0$ and consider the relaxed constraints

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i$$

- ξ_i measures how much the point x_i violates the constraint $y_i(w \cdot x_i + b) \geq 1$
- We want to both maximize the geometric margin and minimize the empirical error represented by the ξ's

Primal optimization problem for Soft-Margin SVM

- The primal optimization problem is given by

$$\begin{cases} \min\limits_{w,b,\xi} & \frac{1}{2}\|w\|_2^2 + C\sum_i \xi_i \\ t.c. & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \qquad i \leq m \\ & \xi_i \geq 0, \qquad\qquad\qquad\quad i \leq m \end{cases}$$

- C is an hyper-parameter of the model and can be estimated via k-fold cross validation Grid-search
- Even in this case the problem is convex and we can use the KKT conditions to obtain the dual problem

Support Vectors

- Let α be the vector of Lagrangian multipliers for the first m constraints and β the one for the non negativity constraints of the slack variables:

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{\|w\|_2^2} + C \sum_i \xi_i - \sum_i \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] - \sum_i \beta_i \xi_i$$

- As before we can write the KKT conditions considering the partial derivatives of $\mathcal{L}$ and the complementarity conditions

$$\begin{aligned}
&\nabla_w \mathcal{L} = w - \textstyle\sum_i \alpha_i y_i x_i = 0 && \Rightarrow \quad w = \textstyle\sum_i \alpha_i y_i x_i \\
&\nabla_b \mathcal{L} = -\textstyle\sum_i \alpha_i y_i = 0 && \Rightarrow \quad \textstyle\sum_i \alpha_i y_i = 0 \\
&\nabla_{\xi_i} \mathcal{L} = C - \alpha_i - \beta_i = 0 && \Rightarrow \quad \alpha_i + \beta_i = C \\
&\forall i \quad \alpha_i [y_i (w \cdot x_i + b) - 1 + \xi_i] = 0 && \Rightarrow \quad \alpha_i = 0 \text{ o } y_i (w \cdot x_i + b) = 1 - \xi_i \\
&\forall i \quad \beta_i \xi_i = 0 && \Rightarrow \quad \beta_i = 0 \text{ o } \xi_i = 0
\end{aligned}$$

Support Vectors

- As in the Hard-margin case, the vector w is given by a linear combination of the points x_i
- A point x_i such that $\alpha_i \neq 0$ is called support vector, but in this case we have two different types of support vectors:
  - if $\alpha_i \neq 0$ and $\xi_i = 0$ then x_i belong to the marginal hyperplane: $y_i(w \cdot x_i + b) = 1$
  - if $\alpha_i \neq 0$ and $\xi_i \neq 0$ then x_i is an outlier and $\beta_i = 0$, thus $\alpha_i = C$

Dual optimization problem for Soft-Margin SVM

- We write the dual problem by replacing the expression for w given by the KKT condition in the Lagrangian function:

$$\mathcal{L} = \frac{1}{2}\|\sum_i \alpha_i y_i x_i\|_2^2 - \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_i \alpha_i y_i b + \sum_i \alpha_i$$

$$= \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j)$$

- In this case we have to consider also the constraints $\beta_i \geq 0$ which correspond to $\alpha_i \leq C$

$$\begin{cases} \max\limits_{\alpha} & \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \\ t.c. & \sum_i \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i \leq m \end{cases}$$

Homeworks

1) In Soft-margin SVM, the objective function term in the slack variables ξ has the form

$$\xi \mapsto \sum_{i=1}^{m} \xi_i.$$

Replace this function with

$$\xi \mapsto \sum_{i=1}^{m} \xi_i^p, \quad p > 1$$

and compute the dual formulation of the optimization problem.

2) If we want to maximize the sparsity of the vector w instead of the geometric margin, we can use the L_p norm instead of the euclidean norm:

$$\begin{cases} \min\limits_{w,b,\xi} & \sum_{i=1}^{m} w_i^p + C \sum_i \xi_i \\ t.c. & y_i(w \cdot x_i + b) \geq 1 - \xi_i, \qquad i \leq m \\ & \xi_i \geq 0, \qquad i \leq m \end{cases}$$

Derive the dual problem for $p = 1$.