Classification Problems, Support Vector Machines & Duality

Goal: Use Lagrangian Duality to help us understand an applied problem.

Classification: Somebody gives you accurately classified data:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

$x_i \in \mathbb{R}^d$ e.g. $y_i \in \{+1, -1\}$

So x_i are the data points & y_i are their "labels".

Objective: Given $\{(x_i, y_i)\}_{i=1}^{N}$ ($N \leftarrow$ Number of data pts)

Learn a Function $f : \mathbb{R}^d \to \mathbb{R}$

so that when we get a new point x (with true label y), $f(x) > 0$ when $y = +1$ & $f(x) < 0$ when $y = -1$

Linear Discrimination

To make this problem easier, we assume that f is an affine function:

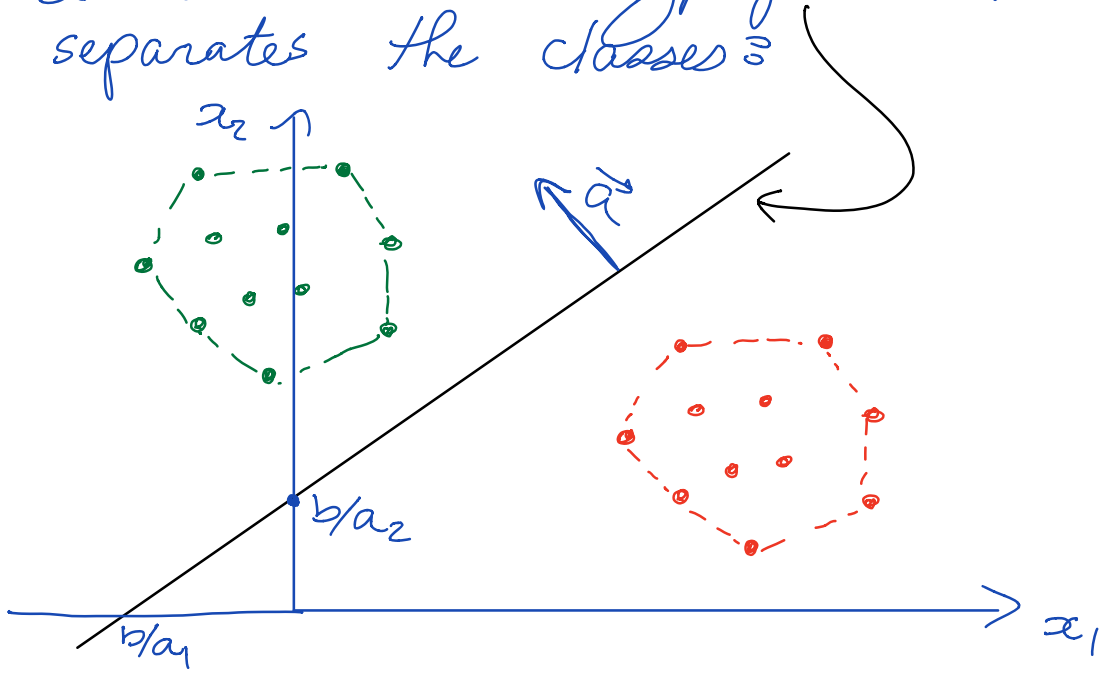
$$f(x) = a^T x + b$$

(a, b unknowns; $a \in \mathbb{R}^d$, $b \in \mathbb{R}$)

so we want

$$\begin{cases} a^T x_i + b > 0 & \text{when } y_i > 0 \\ a^T x_i + b < 0 & \text{when } y_i < 0 \end{cases} \quad (*)$$

So we want a hyperplane that separates the classes:



It will be convenient to work with (**) instead of (*), where

$$\begin{cases} a^T x_i + b > +1 & \text{when } y_i > 0 \\ a^T x_i + b < -1 & \text{when } y_i < 0 \end{cases} \quad \text{---(**)}$$

Support Vector Machines

Playing with (**) further, we see it is equivalent to

$$\begin{cases} a^T x_i + b > +1 & \text{when } y_i > 0 \\ a^T(-x_i) + (-b) > +1 & \text{when } y_i < 0 \end{cases}$$

Let $I := \{ i : y_i = +1 \}$

$J := \{ i : y_i = -1 \}$

$\rightarrow$ all things with label 1

$\rightarrow$ all things with label -1.

Proposition: (**) is feasible if & only if the convex hulls of

$A := \{x_i : i \in I\}$

$B := \{x_j : j \in J\}$

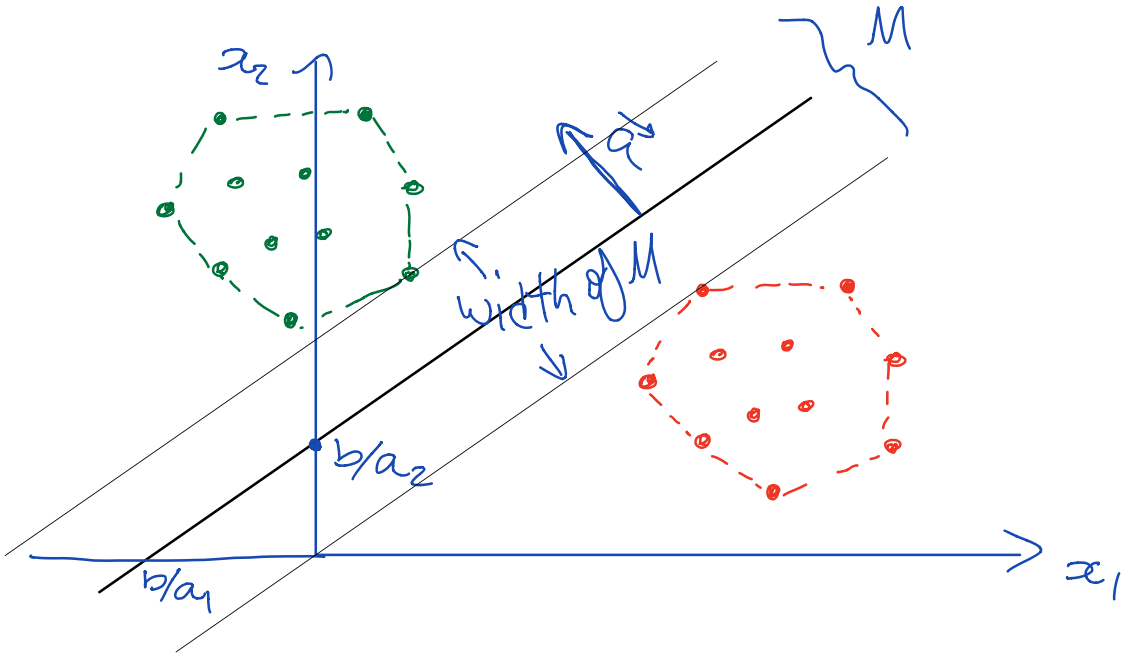
do not intersect.

Robust Separation:

We assume that the convex hulls in proposition above indeed don't intersect so there are choices of $a \in \mathbb{R}^d$ & $b \in \mathbb{R}$ that make (**) true.

Our goal is to find the "best" choice of a & b.

Fact: The set $M := \{x \in \mathbb{R}^d : -1 \leq a^T x + b \leq 1\}$ is called the classification margin & contains no data pts.



The width of M is the distance between the hyperplanes

$\{x : \vec{a}^T x + b = 1\}$ & $\{x : \vec{a}^T x + b = -1\}$

Fact: Width of $M = \frac{2}{\|\vec{a}\|}$

→ why? (Can you prove it?)

So, one reasonable thing to do is to choose a & b to maximize the width of M (so that the classes are maximally separated)

$\Rightarrow$ want to solve:

$$\max_{\substack{a \in \mathbb{R}^d \\ b \in \mathbb{R}}} \frac{2}{\|a\|} \quad \text{s.t.} \begin{cases} \bullet\ a^T x_i + b_i \geq 1 \\ \quad \forall i \in I \\ \bullet\ a_i^T x_i + b_i \leq -1 \\ \quad \forall i \in J. \end{cases}$$

$$\Updownarrow$$

$$\min_{\substack{a \in \mathbb{R}^d \\ b \in \mathbb{R}}} \frac{\|a\|^2}{4} \quad \text{s.t.} \begin{cases} \bullet\ a^T x_i + b_i \geq 1 \\ \quad \forall i \in I \\ \bullet\ a_i^T x_i + b_i \leq -1 \\ \quad \forall i \in J. \end{cases} \quad -(P)$$

The Dual of (P)

Recall that to compute the dual opt. problem, we must

(1) compute the Lagrangian L
(Here $L(a,b,\lambda,\mu)$, more on this in a bit)

(2) Compute the dual function

$$F(\lambda,\mu) = \min_{(a,b)} L(a,b,\lambda,\mu)$$

(3) Write the dual opt. prob.

$$\max_{(\lambda,\mu)} F(\lambda,\mu) \quad \text{s.t.} \quad \lambda \geq 0, \ \mu \geq 0$$

Let's start with (1)

$$L(a,b,\lambda,\mu) = \frac{\|a\|^2}{4} + \sum_{i\in I}\lambda_i\underbrace{(1 - a^T x_i - b)}_{\text{dec } 1-a^Tx_i+b\geq 0} \quad \forall i\in I$$

$$+ \sum_{j\in J}\mu_j(b + a^T x_j - 1)$$

where λ, μ are the dual variables

$\lambda \in \mathbb{R}^{|I|}$, $\mu \in \mathbb{R}^{|J|}$ (Here $|I| = \#$ elements in I)

and $\lambda \geq 0$, $\mu \geq 0$.

Rearranging the expression for L, we get

$$L(a,b,\lambda,\mu) = \frac{\|a\|^2}{4} + \left\langle a, \sum_{j\in J}\mu_j x_j - \sum_{i\in I}\lambda_i x_i \right\rangle$$

$$+ \left\langle b, \sum_{j\in J}\mu_j - \sum_{i\in I}\lambda_i \right\rangle + \sum_{i\in I}\lambda_i + \sum_{j\in J}\mu_j$$

Let's now calculate $F(\lambda,\mu)$:

$$F(\lambda,\mu) = \min_{a,b} L(a,b,\lambda,\mu)$$

Since $\mathcal{L}$ is convex in a & b, taking derivatives with respect to a, b and setting $\nabla_a \mathcal{L} = 0$, $\nabla_b \mathcal{L} = 0$, we get that at the minimizer

$$\begin{cases} \frac{\vec{a}}{2} + \sum_{j \in J} \mu_j \vec{x}_j - \sum_{i \in I} \lambda_i \vec{x}_i = 0 \quad \text{(+)} \\ \sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i = 0 \quad \text{(++)} \end{cases}$$

Plugging (+) & (++) into the exp. for $\mathcal{L}(a, b, \lambda, \mu)$ gives

$$F(\lambda, \mu) = -\left\| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \right\|^2 + \underbrace{\sum_{i \in I} \lambda_i}_{\text{Should be equal by (++)}} + \underbrace{\sum_{j \in J} \mu_j}_{\text{Should be equal by (++)}}$$

So now we can write the dual opt. problem

$$\max_{\lambda,\mu} \; -\Big\|\sum_{i\in I}\lambda_i x_i - \sum_{j\in J}\mu_j x_j\Big\|^2 + \sum_{i\in I}\lambda_i + \sum_{j\in J}\mu_j$$

$$\text{subject to } \begin{cases} \sum_{i\in I}\lambda_i = \sum_{j\in J}\mu_j & (\text{from } (++)) \\ \lambda \geq 0 \\ \mu \geq 0 \end{cases}$$

{ let's play with this by setting $\sum_{i\in I}\lambda_i = s$ (which is ≥ 0 bec $\lambda_i \geq 0$) and dividing by s

$\Rightarrow$

$$\max_{\lambda',\mu',s} \; -s^2\Big\|\sum_{i\in I}\lambda_i' x_i - \sum_{j\in J}\mu_j' x_j\Big\|^2 + 2s$$

$$\text{subject to } \begin{cases} \sum_{i\in I}\lambda_i' = \sum_{j\in J}\mu_j' = 1 \\ \lambda' \geq 0 \\ \mu' \geq 0 \end{cases}, \; s \geq 0$$

Now, notice that the maximum in s is obtained when

$$s^* = \frac{2}{\left\| \sum_{i \in I} \lambda_i' x_i - \sum_{j \in J} \mu_j' x_j \right\|^2}$$

so that the dual now becomes

$$\max_{\lambda', \mu'} \frac{1}{\left\| \sum_{i \in I} \lambda_i' x_i - \sum_{j \in J} \mu_j' x_j \right\|^2}$$

$$\text{subject to} \begin{cases} \sum_{i \in I} \lambda_i' = \sum_{j \in J} \mu_j' = 1 \\ \lambda' \geq 0 \ , \ s \geq 0 \end{cases}$$

(D)

Ok, but do we have strong duality?

Let's check slater's conditions

- f is convex ↘ ($\|a\|^2$ is cvx)
- g_i are convex ↘ ($\langle a, x_i \rangle + b - 1$ are convex)
- h_j are linear ↘ (there are no h_j here)
- $\exists \bar{x} : g_i(\bar{x}) < 0 \quad \forall i$

$\Updownarrow$

$\exists \bar{x}$: all inequalities are strict

but we assumed the convex hulls of the points in I & J don't intersect, so we're good here too. (why?) ↘

$$\Rightarrow \max_{a,b} \frac{2}{\|a\|} = \min_{\lambda,\mu} \left\| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \right\|$$

s.t. $\langle a, x_i \rangle + b \geq 1 \quad \forall i \in I$

$\langle a, x_j \rangle + b \leq 1 \quad \forall j \in J$

maximizes the width of the classification margin

s.t. $\sum_{i \in I} \lambda_i = \sum_{j \in J} \mu_j = 1$

$\lambda \geq 0, \mu \geq 0$

what does this mean?

Note that $\sum_{i \in I} \lambda_i x_i$ where $\sum_{i \in I} \lambda_i = 1$

is a pt. in the convex hull of the set A, and

$$\sum_{j \in J} \mu_j x_i \quad \text{where} \quad \sum_{j \in J} \mu_j = 1$$

is a pt. in the convex hull of the set B.

So the dual problem is finding the two pts in the convex hulls of A & B respectively that are closest to each other!

Strong duality $\Rightarrow$ this is the same as the margin of the best hyperplane

