Classification Problems, Support Vector Machines & Duality

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

Classification: Somebody gives you accurately classified data:

\[ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \]

\( x \in \mathbb{R}^d \) e.g. \( x \in \{+1, -1\} \)

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

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

Learn a function \( f: \mathbb{R}^d \rightarrow \mathbb{R} \)

so that when we get a new point \( x \), \( f(x) > 0 \) when \( y = +1 \)

(with true label \( y \)) & \( f(x) < 0 \) when \( y = -1 \)
Linear Discrimination

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

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

so we want

$$\begin{cases} a^T x_i + b > 0 \text{ when } y_i > 0 \\ a^T x_i + b < 0 \text{ when } y_i < 0 \end{cases}$$

So we want a hyperplane that separates the classes.
It will be convenient to work with (**) instead of (*) , where
\[
\begin{cases}
  a^T x_i + b > +1 & \text{when } y_i > 0 \\
  a^T x_i + b < -1 & \text{when } y_i < 0
\end{cases}
\]

---

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 \( I := \{ i : y_i = +1 \} \)

\( J := \{ i : y_i = -1 \} \)

all things with label +1

all things with label -1.
**Proposition**: (**) is feasible if and only if the convex hulls of 
\[ \mathcal{A} := \{ x_i : i \in I \} \] do not intersect, 
\[ \mathcal{B} := \{ x_j : j \in J \} \]

**Robust Separation?**

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

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

**Fact**: The set \( \mathcal{M} := \{ x \in \mathbb{R}^d : -1 \leq ax + b \leq 1 \} \) is called the classification margin & contains no data pts.
The width of $M$ is the distance between the hyperplanes

$\{ x = a_1^T x + b = 1/2 \} \land \{ x = a_1^T x + b = -1/2 \}$

Fact 2: Width of $M = \frac{2}{\| a_1 \|}$

Why? (Can you prove it?)
So, one reasonable thing to do is to choose $a$ and $b$ to maximize the width of $M$ (so that the classes are maximally separated).

We want to solve:

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

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

   $\max F(\lambda, \mu)$  s.t.  $\lambda \geq 0$

   $(\lambda, \mu)$  $\mu \geq 0$

Let’s start with (1)
\[ L(a,b,\lambda,\mu) = \frac{1}{4} \| a \|^2 + \sum_{i \in I} \lambda_i (1 - a^T x_i - b) \]
\[ + \sum_{j \in J} \mu_j (b + a^T x_i - 1) \]

where \( \lambda, \mu \) are the dual variables.

\( \lambda \in \mathbb{R}^{|I|}, \quad \mu \in \mathbb{R}^{|J|} \)

(Here \( |I| \) = \# of elements in \( I \))

and \( \lambda \geq 0, \mu \geq 0 \).

Rearranging the expression for \( L \), we get:

\[ L(a,b,\lambda,\mu) = \frac{1}{4} \| a \|^2 + \langle a, \sum_{j \in J} \mu_j x_j - \sum_{i \in I} \lambda_i x_i \rangle \]
\[ + \langle b, \sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i \rangle + \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j \]

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

\[
\begin{align*}
\begin{cases}
\frac{\partial}{\partial a} + \sum_{j \in J} \mu_j x_j - \sum_{i \in I} \lambda_i x_i = 0 &\text{--- (t)} \\
\sum_{j \in J} \mu_j - \sum_{i \in I} \lambda_i = 0 &\text{--- (tt)}
\end{cases}
\end{align*}
\]

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

\[
F(\lambda, \mu) = - \| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \|^2
\]

\[
+ \sum_{i \in I} \lambda_i + \sum_{j \in J} \mu_j
\]

[Should be equal by (tt)]

So now we can write the dual opt. problem
\[
\max - \frac{1}{2} \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} m_j x_j + \sum_{i \in I} \lambda_i + \sum_{j \in J} m_j
\]
subject to \(\sum_{i \in I} \lambda_i = \sum_{j \in J} m_j\) (from (++)
\[
\begin{cases}
\lambda \geq 0 \\
M \geq 0
\end{cases}
\]

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

\[
\max - \frac{1}{2} \sum_{i \in I} \lambda_i' x_i - \sum_{j \in J} m'_j x_j + 2s
\]
subject to \(\sum_{i \in I} \lambda_i' = \sum_{j \in J} m'_j = 1\)
\[
\begin{cases}
\lambda' \geq 0 \\
S \geq 0 \\
M \geq 0
\end{cases}
\]
Now, notice that the maximum in $s$ is obtained when

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

so that the dual now becomes

$$\max_{\mathbf{\lambda} \geq 0} \frac{1}{\| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} m_j x_j \|^2}$$

subject to

$$\begin{cases} \sum_{i \in I} \lambda_i = \sum_{j \in J} m_j = 1 \\ \lambda_i \geq 0, \quad m_j \geq 0 \end{cases}$$

(D)

Ok, but do we have strong duality?
Let's check Slater's conditions

- \( f \) is convex \( \forall (11a11^2 \text{ is cvx}) \)
- \( g_i \) are convex \( \forall (\langle a, x_i \rangle + b - 1 \text{ are convex}) \)
- \( h_j \) are linear \( \forall (\text{there are no } h_j \text{ here}) \)
- \( \exists \overline{x}: g_i(\overline{x}) < 0 \ \forall i \)

\[ \exists \overline{x}: \text{all inequalities are strict} \]

but we assumed the convex hulls \( \mathcal{H} \) the points in \( I \) & \( J \) don't intersect, so we're good here too. (why?)

\[ \Rightarrow \max_{a, b} \frac{2}{11a11} = \min_{\lambda, \mu} \| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \| \]

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

\( \langle a, x_j \rangle + b \leq 1 \ \forall j \in J \)

maximizes the width of the classification margin

\[ \text{minimize } \| \sum_{i \in I} \lambda_i x_i - \sum_{j \in J} \mu_j x_j \| \]

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 $\sum_{j \in J} \mu_j x_i$ where $\sum_{j \in J} \mu_j = 1$ 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 $\implies$ this is the same as the margin of the best hyperplane.