TD - Support Vector Machines: solution

Exercise 1 (preliminaries). *Set $\mathcal{X} = \mathbb{R}^d$. Define the hyperplane*

$$\mathcal{H}_{w,b} = \{x \in \mathcal{X} : \langle w, x\rangle + b = 0\}$$

for some fixed $w \in \mathcal{X}$ ($w \neq 0$) and $b \in \mathbb{R}$. For a fixed $z \in \mathcal{X}$, consider the problem

$$\min_{x \in \mathcal{H}_{w,b}} \frac{1}{2}\|x - z\|^2 .$$

1. *Write the Lagragian function $L(x; \nu)$ associated with this problem.*

   ▶ $\forall (x, \nu) \in \mathcal{X} \times \mathbb{R} : L(x; \nu) = \frac{1}{2}\|x - z\|^2 + \nu\langle w, x\rangle + \nu b.$

2. *Solve the KKT conditions and characterize the solution.*

   ▶ Let $(x^\star; \nu^\star)$ be a solution of the saddle point problem:

   $$\min_{x \in \mathcal{H}_{w,b}} \max_{\nu \in \mathbb{R}} L(x; \nu).$$

   Then:

   (1) Primal feasibility: $\langle w, x^\star\rangle + b = 0$.
   (2) Dual feasibility: $\emptyset$.
   (3) First order primal optimality: $\nabla_x L(x^\star; \nu^\star) = 0$.
   (4) Complementary slackness: $\emptyset$.

   From (3) we obtain: $0 = \nabla_x L(x^\star; \nu^\star) = x^\star - z + \nu^\star w$, *i.e.* $x^\star = z - \nu^\star w$. By substituting the previous equality in (1), we get: $\nu^\star = \frac{\langle w, z\rangle + b}{\|w\|^2}$. This characterizes the solution $(x^\star, \nu^\star)$.

3. *Prove that the distance of a point z to $\mathcal{H}$ is equal to*

   $$d(z, \mathcal{H}_{w,b}) = \frac{|\langle w, z\rangle + b|}{\|w\|} .$$

   ▶ $d(z, \mathcal{H}_{w,b}) = \|x^\star - z\| = \|\nu^\star w\| = |\nu^\star| \, \|w\| = \frac{|\langle w, z\rangle + b|}{\|w\|}.$

Exercise 2 (linearly separable case). *Consider a training set formed by couples (x_i, y_i) for $i \in \{1, \dots, n\}$ where x_i is a feature vector in $\mathcal{X}$ and $y_i \in \{-1, +1\}$ for all i. The hyperplane $\mathcal{H}_{w,b}$ is called* separating *if*

$$\forall i, \quad y_i(\langle w, x_i\rangle + b) > 0 .$$

In the sequel, we assume that a separating hyperplane exists. Among all separating hyperplanes, we seek to find the one which maximizes the minimum distance

$$f(w, b) = \min_{i=1,\dots,n} d(x_i, \mathcal{H}_{w,b}) .$$

1. *Show that if (w, b) defines a separating hyperplane, then $f(w, b) = c(w, b)/\|w\|$ where $c(w, b) = \min_i y_i(\langle w, x_i\rangle + b)$.*

   ▶ We have:

$$\forall (w, b) \in \mathcal{X} \times \mathbb{R}\colon f(w, b) = \min_{i=1,\ldots,n} \frac{|\langle w, x_i\rangle + b|}{\|w\|} = \min_{i=1,\ldots,n} \frac{y_i(\langle w, x_i\rangle + b)}{\|w\|} = \frac{c(w, b)}{\|w\|}.$$

Thus, we are interested in solving the problem

$$\max_{w,b} \frac{c(w, b)}{\|w\|} \text{ such that } \forall i,\ y_i(\langle w, x_i\rangle + b) \geq 0\,.$$

Let $(w^\star, b^\star)$ be a solution and define

$$v^\star = \frac{w^\star}{c(w^\star, b^\star)} \text{ and } a^\star = \frac{b^\star}{c(w^\star, b^\star)}$$

2. *Justify that $(w^\star, b^\star)$ and $(v^\star, a^\star)$ define the same separating hyperplane.*

   ▶ We have assumed that $\exists (w_0, b_0) \in \mathcal{X} \times \mathbb{R}$ such that $\forall i = 1, \ldots, n\colon y_i(\langle w_0, x_i\rangle + b_0) > 0$. Thus (w_0, b_0) is a feasible point to the optimization problem at hand and $\frac{c(w_0,b_0)}{\|w_0\|} > 0$. As a consequence, $\frac{c(w^\star,b^\star)}{\|w^\star\|} \geq \frac{c(w_0,b_0)}{\|w_0\|} > 0$ and $c(w^\star, b^\star) > 0$.
   Thus

$$\begin{aligned}\forall (x, y) \in \mathcal{X} \times \{-1, +1\}\colon\ & y(\langle w^\star, x\rangle + b^\star) > 0\\ &\Leftrightarrow y(\langle w^\star, x\rangle + b^\star)/c(w^\star, b^\star) > 0\\ &\Leftrightarrow y(\langle v^\star, x\rangle + a^\star) > 0,\end{aligned}$$

   i.e. $(w^\star, b^\star)$ and $(v^\star, a^\star)$ define the same separating hyperplane.

3. *Prove that $(v^\star, a^\star)$ solves the optimization problem*

$$\max_{v,a} \frac{1}{\|v\|} \text{ such that } \forall i,\ y_i(\langle v, x_i\rangle + a) \geq 1\,.$$

   ▶ The solution is in two steps: first prove that $(v^\star, a^\star)$ is a feasible point, then that for any other feasible point (v, a), the objective value of (v, a) is less than the one of $(v^\star, a^\star)$ (zero-th order optimality condition).

   Primal feasibility:

$$\begin{aligned}\forall i = 1, \ldots, n\colon y_i(\langle v^\star, x_i\rangle + a^\star) &= \frac{y_i(\langle w^\star, x_i\rangle + b^\star)}{c(w^\star, b^\star)}\\ &= \frac{y_i(\langle w^\star, x_i\rangle + b^\star)}{\min_{i'=1,\ldots,n} y_{i'}(\langle w^\star, x_{i'}\rangle + b^\star)}\\ &\geq 1.\end{aligned}$$

Zero-th order optimality condition: $\forall (v,a) \in \mathcal{X} \times \mathbb{R}$ such that $\forall i = 1,\dots,n$: $y_i(\langle v, x_i \rangle + a) \geq 1$ we have that

- $c(v,a) \geq 1$, so $c(v,a) \neq 0$ and $\frac{1}{c(v,a)} \leq 1$;
- $\forall i = 1,\dots,n$: $y_i(\langle v, x_i \rangle + a) \geq 0$ so (v,a) is feasible to the first optimization problem and by optimality: $\frac{c(v,a)}{\|v\|} \leq \frac{c(w^\star,b^\star)}{\|w^\star\|}$. Moreover, by definition $\frac{c(w^\star,b^\star)}{\|w^\star\|} = \frac{1}{\|v^\star\|}$.

As a consequence:

$$\begin{aligned}
\frac{1}{\|v\|} &= \frac{c(v,a)}{\|v\|}\frac{1}{c(v,a)} && (c(v,a) \neq 0) \\
&\leq \frac{c(v,a)}{\|v\|} && (\frac{1}{c(v,a)} \leq 1) \\
&\leq \frac{c(w^\star,b^\star)}{\|w^\star\|} && \text{(optimality)} \\
&= \frac{1}{\|v^\star\|} && \text{(definition).}
\end{aligned}$$

So for all (v,a) feasible, $\frac{1}{\|v\|} \leq \frac{1}{\|v^\star\|}$. This second point concludes the solution.

4. *Deduce that $(v^\star, a^\star)$ solves the optimization problem*

$$\min_{v,a} \frac{\|v\|^2}{2} \textit{ such that } \forall i,\ 1 - y_i(\langle v, x_i \rangle + a) \leq 0\,. \tag{1}$$

► $\forall (v,v') \in \mathcal{X} \times \mathcal{X}$, $\frac{1}{\|v\|} \leq \frac{1}{\|v'\|} \Leftrightarrow \|v'\| \leq \|v\| \Leftrightarrow \frac{\|v'\|^2}{2} \leq \frac{\|v\|^2}{2}$. The constraints are the same.

5. *Write the Lagrangian $L(v,a;\phi)$.*

► We have:

$\forall (v,a,\phi) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^n$:

$$L(v,a;\phi) = \begin{cases} \frac{1}{2}\|v\|^2 + \sum_{i=1}^n \phi_i - \langle \sum_{i=1}^n \phi_i y_i x_i, v \rangle - a \sum_{i=1}^n \phi_i y_i \text{ if } \forall i = 1,\dots,n\colon \phi_i \geq 0 \\ -\infty \text{ otherwise.} \end{cases}$$

6. *Write the KKT conditions.*

► Let $(v^\star, a^\star; \phi^\star)$ be a saddle point. Then:

(1) Primal feasibility: $\forall i = 1,\dots,n$: $y_i(\langle v^\star, x_i \rangle + a^\star) \geq 1$.

(2) Dual feasibility: $\forall i = 1,\dots,n$: $\phi_i^\star \geq 0$.

(3) First order primal optimality:

$$\nabla_v L(v^\star, a^\star; \phi^\star) = v^\star - \sum_{i=1}^n \phi_i^\star y_i x_i = 0.$$
$$\nabla_a L(v^\star, a^\star; \phi^\star) = - \sum_{i=1}^n \phi_i^\star y_i = 0.$$

(4) Complementary slackness: $\forall i = 1, \ldots, n \colon 1 - y_i(\langle v^\star, x_i \rangle + a^\star) = 0$ or $\phi_i^\star = 0$.

7. *Let $(v, a; \phi)$ be a saddle point of the Lagrangian. Show that ϕ_i is non-zero only if $y_i(\langle v, x_i \rangle + a) = 1$.*

   ▶ We have to show that : $\phi_i \neq 0 \Rightarrow y_i(\langle v, x_i \rangle + a) = 1$. This is true from (4).

The training points (x_i, y_i) satisfying the above property are the closest to the hyperplane $\mathcal{H}_{v,a}$. The corresponding x_i's are often called support vectors.

8. *If one is given a dual solution $\phi^\star$, how to recover a primal solution $(v^\star, a^\star)$ from $\phi^\star$?*

   ▶ For $v^\star$, KKT conditions indicate that $v^\star = \sum_{i=1}^n \phi_i^\star y_i x_i$. On the other hand, for $a^\star$, let $\mathcal{I} = \{i = 1, \ldots, n \colon \phi_i^\star > 0\}$. If $\mathcal{I}$ is empty, then $v^\star = 0$ and $y_i a^\star \geq 1$ for all $i = 1, \ldots, n$. This is impossible as soon as there is at least one point in each class. So, $\mathcal{I} \neq \emptyset$. Let $j \in \mathcal{I}$, then $\phi_j^\star \neq 0$, so $y_j(\langle v^\star, x_j \rangle + a^\star) = 1$, from which we deduce that $a^\star = y_j - \langle v^\star, x_j \rangle$.

Define the $n \times n$ matrices $K = (\langle x_i, x_j \rangle)_{i,j=1\ldots n}$, $D = \mathrm{diag}(y_1 \ldots y_n)$ and $\mathbf{1}^T = (1, \ldots, 1)$.

9. *Prove that the dual problem reduces to*

$$\min_{\substack{\phi \geq 0 \\ y^T\phi = 0}} \frac{1}{2} \phi^T DKD\phi - \mathbf{1}^T \phi\,.$$

   ▶ The dual problem comes from keeping the dual variables feasible ($\forall i = 1, \ldots, n\colon$ $\phi_i \geq 0$) and minimizing the Lagrangian with respect to the primal variables:

$$\nabla_v L(v^\star, a^\star; \phi) = v^\star - \sum_{i=1}^n \phi_i y_i x_i = 0 \textit{ i.e. } v^\star = \sum_{i=1}^n \phi_i y_i x_i \textit{ and}$$
$$\nabla_a L(v^\star, a^\star; \phi) = - \sum_{i=1}^n \phi_i y_i = 0.$$

   By substitution, we obtain:

$$\begin{aligned} L(v^\star, a^\star; \phi) &= \frac{1}{2} \sum_{1 \leq i,j \leq n} K_{i,j} y_i y_j \phi_i \phi_j + \sum_{i=1}^n \phi_i - \frac{1}{2} \sum_{1 \leq i,j \leq n} K_{i,j} y_i y_j \phi_i \phi_j - a \times 0 \\ &= -\frac{1}{2} \phi^T DKD\phi + \mathbf{1}^T \phi. \end{aligned}$$

Thus, the dual problem consists in maximizing $-\frac{1}{2}\phi^T DKD\phi + \mathbf{1}^T\phi$ subject to $\phi_i \geq 0$ and $\sum_{i=1}^n \phi_i y_i = 0$. This concludes the solution.

10. *Assume that this algorithm has identified a dual solution $\phi^\star$. Write explicitely the classifier as a function of $\phi^\star$.*

    ▶ Classifier $C\colon x \in \mathcal{X} \mapsto \operatorname{sign}(\langle v^\star, x\rangle + a^\star) = \operatorname{sign}(\sum_{\substack{1\leq i\leq n\\ \phi_i>0}} \phi_i y_i \langle x_i, x\rangle + a^\star)$, where $a^\star$ has been defined above.

11. *What part of the training data do you need in order to implement the above classifier?*

    ▶ We only need the support vectors, *i.e.* the points x_i such that $\phi_i > 0$.

Exercise 3 (non separable case). *Consider the case when a separable hyperplane might not exist. The constraints $1 - y_i(\langle v, x_i\rangle + a) \leq 0$ in Problem (1) may not be jointly feasible. For a fixed $c > 0$, we consider the relaxed problem*

$$\min_{v,a,\xi} \frac{\|v\|^2}{2} + c\sum_i \xi_i \quad \textit{such that } \forall i,\ 1 - y_i(\langle v, x_i\rangle + a) \leq \xi_i \textit{ and } \xi_i \geq 0\,. \tag{2}$$

1. *How many constraints has this problem?*

   ▶ This problem has $2n$ constraints.

2. *Write the Lagrangian function.*

   ▶ We have:

   $\forall (v, a, \xi, \phi, \alpha) \in \mathcal{X} \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$:

   $$\begin{aligned} L(v, a, \xi; \phi, \alpha) &= \frac{1}{2}\|v\|^2 + c\sum_{i=1}^n \xi_i \\ &\quad + \sum_{i=1}^n \phi_i - \langle \sum_{i=1}^n \phi_i y_i x_i, v\rangle - a\sum_{i=1}^n \phi_i y_i \\ &\quad - \sum_{i=1}^n \phi_i \xi_i - \sum_{i=1}^n \alpha_i \xi_i, \end{aligned}$$

   if $\forall i = 1, \dots, n\colon \phi_i \geq 0$ and $\alpha_i \geq 0$ and $-\infty$ otherwise.

3. *Show that the dual problem reduces to*

   $$\min_{\substack{c\geq\phi\geq 0\\ y^T\phi=0}} \frac{1}{2}\phi^T DKD\phi - \mathbf{1}^T\phi\,.$$

▶ First order conditions for the primal variables give:

$$\begin{aligned}
\nabla_v L(v^\star, a^\star, \xi^\star; \phi, \alpha) &= v^\star - \sum_{i=1}^n \phi_i y_i x_i = 0 \ \textit{i.e.}\ v^\star = \sum_{i=1}^n \phi_i y_i x_i;\\
\nabla_a L(v^\star, a^\star, \xi^\star; \phi, \alpha) &= -\sum_{i=1}^n \phi_i y_i = 0;\\
\nabla_\xi L(v^\star, a^\star, \xi^\star; \phi, \alpha) &= c\mathbf{1} - \phi - \alpha = 0 \ \textit{i.e.}\ \alpha = c\mathbf{1} - \phi.
\end{aligned}$$

By substitution, we obtain:

$$\begin{aligned}
L(v^\star, a^\star, \xi^\star; \phi, \alpha) &= \frac{1}{2}\sum_{1\le i,j\le n} K_{i,j} y_i y_j \phi_i \phi_j + \sum_{i=1}^n \phi_i - \frac{1}{2}\sum_{1\le i,j\le n} K_{i,j} y_i y_j \phi_i \phi_j\\
&\quad - a \times 0 + \langle \xi, \underbrace{c\mathbf{1} - \phi - \alpha}_{0} \rangle\\
&= -\frac{1}{2}\phi^T DKD\phi + \mathbf{1}^T \phi.
\end{aligned}$$

To conclude, dual feasibility indicates that $\phi_i \ge 0$ and $\alpha_i = c - \phi_i \ge 0$. Thus, the dual problem consists in maximizing $-\frac{1}{2}\phi^T DKD\phi + \mathbf{1}^T\phi$ subject to $c \ge \phi_i \ge 0$ and $\sum_{i=1}^n \phi_i y_i = 0$. This concludes the solution.