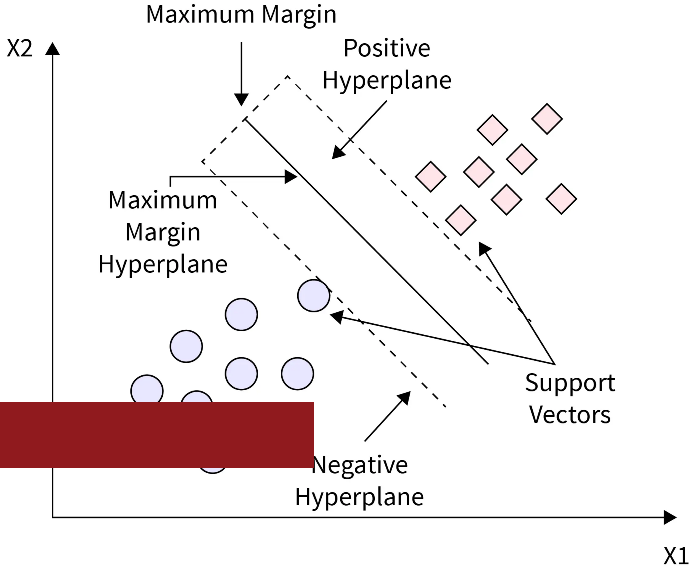
Support Vector Machine (SVM)

Nirupam Gupta

Department of Computer Science

UNIVERSITY OF COPENHAGEN

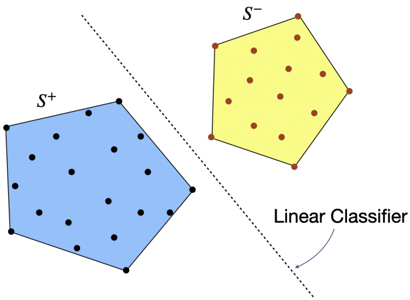
Linear Classifier



Linear separability

Consider a dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ such that $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, +1\}$ for all i.
Define $S^+ = (x \mid (x, y) \in S\ ,\ \ y = +1)$ and $S^- = (x \mid (x, y) \in S\ ,\ \ y = -1)$.

Suppose that $\mathsf{Co}(S^+) \cap \mathsf{Co}(S^-) = \emptyset$. Hence, $\exists$ a hyperplane $f(x) := w^\intercal x + b$ such that $f(x) > 0\ , \forall x \in S^+$ and $f(x) < 0\ , \forall x \in S^-$.



Determining a linear classifier

Determining $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ such that $\mathbf{sign}(w^\intercal x + b) > 0$ for all $x \in S^+$ and $\mathbf{sign}(w^\intercal x + b) < 0$ for all $x \in S^-$ reduces to the following LP:

$$\begin{array}{ll} \underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Minimize}} & 1 \\ \text{Subject to} & y_i(w^\intercal x_i + b) > 0 \quad,\ i = 1, \ldots, n \end{array}$$

Determining a linear classifier

Determining $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ such that $\mathsf{sign}(w^\intercal x + b) > 0$ for all $x \in S^+$ and $\mathsf{sign}(w^\intercal x + b) < 0$ for all $x \in S^-$ reduces to the following LP:

$$\begin{aligned} \underset{w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Minimize}} \quad & 1 \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) > 0 \quad , \ i = 1, \ldots, n \end{aligned}$$

There can be infinitely many solutions to the above optimization problem.

Determining a linear classifier

Determining $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ such that $\mathsf{sign}(w^\mathsf{T} x + b) > 0$ for all $x \in S^+$ and $\mathsf{sign}(w^\mathsf{T} x + b) < 0$ for all $x \in S^-$ reduces to the following LP:

$$\begin{array}{ll} \underset{w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Minimize}} & 1 \\ \text{Subject to} & y_i(w^\mathsf{T} x_i + b) > 0 \quad , \ i = 1, \ldots, n \end{array}$$

$$\frac{|w^\mathsf{T} x + b|}{\|w\|}$$

There can be infinitely many solutions to the above optimization problem.

SVM. Determine $(w, b) \in \mathbb{R}^m \times \mathbb{R}$ that maximizes the **separation margin**. That is, on top of satisfying the separation constraint, we would like to

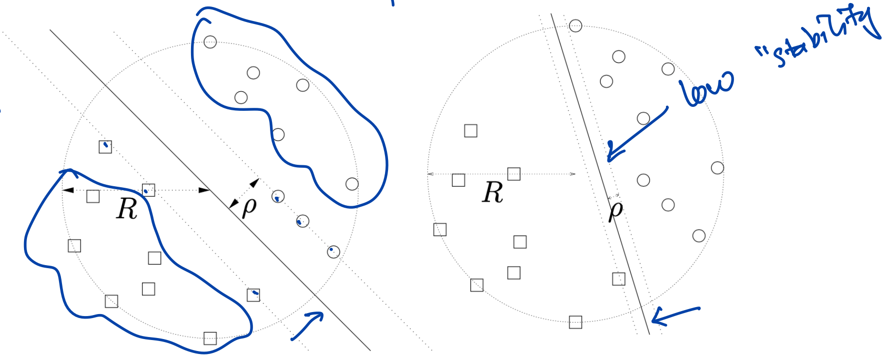
$$\text{Maximize} \min_{i=1,\ldots,n} \frac{|w^\mathsf{T} x_i + b|}{\|w\|}$$

Linear SVM

$$\left[\begin{array}{ll} \underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} & \left(\min_{i=1,\dots,n} \frac{|w^\intercal x_i + b|}{\|w\|}\right) \\ \text{Subject to} & \underline{y_i(w^\intercal x_i + b) > 0} \quad , \ i = 1, \dots, n \end{array}\right]$$

↳ separation constraint

high "algorithmic stability"



Linear SVM

Not convex

$$
\begin{aligned}
\underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \min_{i=1,\dots,n} \frac{|w^\intercal x_i + b|}{\|w\|} \\
\text{Subject to} \quad & y_i(w^\intercal x_i + b) > 0 \qquad , \; i = 1, \dots, n
\end{aligned}
$$

affine

Is the above a convex optimization problem?

Linear SVM

$$\begin{aligned} \underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \min_{i=1,\dots,n} \frac{|w^\intercal x_i + b|}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) > 0 \qquad , \ i = 1, \dots, n \end{aligned}$$

Is the above a convex optimization problem?

No.

Linear SVM

$$\begin{array}{ll} \underset{w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Maximize}} & \min_{i=1,\ldots,n} \frac{|w^\intercal x_i + b|}{\|w\|} \\ \text{Subject to} & y_i(w^\intercal x_i + b) > 0 \quad , \ i = 1, \ldots, n \end{array}$$

We can reduce the above to the following:

new variable.

$$\begin{array}{ll} \underset{\rho \in \mathbb{R},\ w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Maximize}} & \frac{\rho}{\|w\|} \\ \text{Subject to} & y_i(w^\intercal x_i + b) > 0 \quad , \ i = 1, \ldots, n \\ & |w^\intercal x_i + b| \underline{\underline{\geq \rho}} \quad , \ i = 1, \ldots, n \\ & \rho > 0 \end{array}$$

enforce.

Linear SVM

$$\begin{aligned} \underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \min_{i=1,\dots,n} \frac{|w^\intercal x_i + b|}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) > 0 \qquad , \ i = 1,\dots,n \end{aligned}$$

We can reduce the above to the following:

$$\begin{aligned} \underset{\rho\in\mathbb{R},\ w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \frac{\rho}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) > 0 \qquad , \ i = 1,\dots,n \\ & |w^\intercal x_i + b| \geq \rho \qquad , \ i = 1,\dots,n \\ & \rho > 0 \end{aligned}$$

dummy variable.

$|y_i| = 1$

This can be further reduced to (why?):

$$\begin{aligned} \underset{\rho\in\mathbb{R},\ w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \frac{\rho}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) \geq \rho \qquad , \ i = 1,\dots,n \\ & \rho > 0 \end{aligned}$$

Linear SVM as quadratic optimization problem

Recall that $\rho > 0$. Define $(w', b') = \frac{1}{\rho}(w, b)$. With this substitution, we obtain that

$$\begin{aligned} \underset{\rho \in \mathbb{R},\ w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Maximize}} \quad & \frac{\rho}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) \geq \rho \quad , \ i = 1, \dots, n \\ & \rho > 0 \end{aligned}$$

Linear SVM as quadratic optimization problem

Recall that $\rho > 0$. Define $(w', b') = \frac{1}{\rho}(w, b)$. With this substitution, we obtain that

$w = \rho w'$, $b = \rho b'$

$$
\begin{aligned}
\underset{\cancel{\rho\in\mathbb{R},}\ w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \frac{\rho}{\|w\|} \\
\text{Subject to} \quad & y_i(w^\intercal x_i + b) \geq \rho \quad , \ i = 1, \ldots, n \\
& \rho > 0
\end{aligned}
$$

is reducible to

min. $\|w\|$

$$
\begin{aligned}
\underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \frac{1}{\|w\|} \\
\text{Subject to} \quad & y_i(w^\intercal x_i + b) \geq 1 \quad , \ i = 1, \ldots, n
\end{aligned}
$$

Linear SVM as quadratic optimization problem

Recall that $\rho > 0$. Define $(w', b') = \frac{1}{\rho}(w, b)$. With this substitution, we obtain that

$$\begin{aligned} \underset{\rho\in\mathbb{R},\ w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \frac{\rho}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) \geq \rho \quad , \ i = 1, \ldots, n \\ & \rho > 0 \end{aligned}$$

is reducible to

$f_i(w) \leq 0$

$$\begin{aligned} \underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Maximize}} \quad & \frac{1}{\|w\|} \\ \text{Subject to} \quad & y_i(w^\intercal x_i + b) \geq 1 \quad , \ i = 1, \ldots, n \end{aligned}$$

Algorithmic acceleration

This can be solved using the following **quadratic programming** (QP):

$$\begin{aligned} \underset{w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Minimize}} \quad & \frac{1}{\cancel{2}}\|w\|^2 \\ \text{Subject to} \quad & 1 - y_i(w^\intercal x_i + b) \leq 0 \quad , \ i = 1, \ldots, n \end{aligned} \qquad \text{(Linear SVM)}$$

Lagrange dual of linear SVM

Lagrange dual function:

$$\mathcal{L}(w, b, \lambda) := \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \lambda_i \left(1 - y_i(w^\intercal x_i + b)\right).$$

Lagrange dual of linear SVM

Lagrange dual function:

$$\mathcal{L}(w, b, \lambda) := \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \lambda_i \left(1 - y_i(w^\top x_i + b)\right).$$

$\phi(\lambda) = \min_{w} \mathcal{L}(w, b)$

Dual optimization problem:

$$\begin{array}{ll} \underset{\lambda \in \mathbb{R}^n}{\text{Maximize}} & \phi(\lambda) := \sum_{i=1}^{n} \lambda_i - \frac{1}{2}\left\|\sum_{i=1}^{n} \lambda_i y_i x_i\right\|^2 \\ \text{Subject to} & \lambda \succeq \mathbf{0} \\ & \sum_{i=1}^{n} \lambda_i y_i = 0 \end{array}$$

(Dual of Linear SVM)

Lagrange dual of linear SVM

Lagrange dual function:

$$\mathcal{L}(w, b, \lambda) := \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \lambda_i \left(1 - y_i(w^\top x_i + b)\right).$$

Dual optimization problem:

$$\begin{aligned} \underset{\lambda \in \mathbb{R}^n}{\text{Maximize}} \quad & \phi(\lambda) := \textstyle\sum_{i=1}^n \lambda_i - \frac{1}{2}\left\|\sum_{i=1}^n \lambda_i y_i x_i\right\|^2 \\ \text{Subject to} \quad & \lambda \succeq \mathbf{0} \\ & \textstyle\sum_{i=1}^n \lambda_i y_i = 0 \end{aligned} \qquad \text{(Dual of Linear SVM)}$$

Let (w^*, b^*) and λ^* be solutions to the (Linear SVM) and (Dual of Linear SVM), respectively.

KKT optimality conditions of linear SVM

$$
\begin{aligned}
1 - y_i(\langle w^*, x_i\rangle + b^*) &\leq 0\ , \quad i = 1, \dots, n \\
\lambda^* &\succeq \mathbf{0} \\
\lambda_i^*\left(1 - y_i(\langle w^*, x_i\rangle + b^*)\right) &= 0\ , \quad i = 1, \dots, n \\
w^* - \textstyle\sum_{i=1}^n \lambda_i^* y_i x_i &= \mathbf{0} \\
\textstyle\sum_{i=1}^n \lambda_i^* y_i &= 0
\end{aligned}
\qquad \text{(KKT Linear SVM)}
$$

— constraint

— dual

complementary slackness

$\nabla_{(w,b)} L(w^*, b^*, \lambda^*) = 0$.

KKT optimality conditions of linear SVM

$$
\begin{aligned}
1 - y_i(\langle w^*, x_i \rangle + b^*) &\leq 0\ , \quad & i = 1, \ldots, n \\
\lambda^* &\succeq \mathbf{0} & \\
\lambda_i^*\,(1 - y_i(\langle w^*, x_i \rangle + b^*)) &= 0\ , \quad & i = 1, \ldots, n \\
w^* - \textstyle\sum_{i=1}^n \lambda_i^* y_i x_i &= \mathbf{0} & \\
\textstyle\sum_{i=1}^n \lambda_i^* y_i &= 0 &
\end{aligned}
\qquad \text{(KKT Linear SVM)}
$$

Support vectors: set of points (x_i, y_i) for which $\lambda_i^* > 0$. Due to complementary slackness, $y_i(\langle w^*, x_i \rangle + b) = 1$ (or $\langle w^*, x_i \rangle + b = y_i$) for all support vectors.

KKT optimality conditions of linear SVM

$$
\begin{aligned}
1 - y_i(\langle w^*, x_i\rangle + b^*) &\leq 0\ , \quad i = 1, \ldots, n \\
\lambda^* &\succeq \mathbf{0} \\
\lambda_i^* \left(1 - y_i(\langle w^*, x_i\rangle + b^*)\right) &= 0\ , \quad i = 1, \ldots, n \\
w^* - \textstyle\sum_{i=1}^{n} \lambda_i^* y_i x_i &= \mathbf{0} \\
\textstyle\sum_{i=1}^{n} \lambda_i^* y_i &= 0
\end{aligned}
\tag{KKT Linear SVM}
$$

Support vectors: set of points (x_i, y_i) for which $\lambda_i^* > 0$. Due to complementary slackness, $y_i(\langle w^*, x_i\rangle + b) = 1$ (or $\langle w^*, x_i\rangle + b = y_i$) for all support vectors.

$w^* = \sum_{i \in SV} \lambda_i^* y_i x_i$, where $SV \subseteq [n]$ denotes the index set of all the support vectors.

KKT optimality conditions of linear SVM

$$
\begin{aligned}
1 - y_i(\langle w^*, x_i\rangle + b^*) &\leq 0\ , \quad i = 1, \dots, n \\
\lambda^* &\succeq \mathbf{0} \\
\lambda_i^* \boxed{(1 - y_i(\langle w^*, x_i\rangle + b^*))} &= 0\ , \quad i = 1, \dots, n \\
w^* - \textstyle\sum_{i=1}^n \lambda_i^* y_i x_i &= \mathbf{0} \\
\textstyle\sum_{i=1}^n \lambda_i^* y_i &= 0
\end{aligned}
\qquad \text{(KKT Linear SVM)}
$$

Support vectors: set of points (x_i, y_i) for which $\lambda_i^* > 0$. Due to complementary slackness, $y_i(\langle w^*, x_i\rangle + b) = 1$ (or $\langle w^*, x_i\rangle + b = y_i$) for all support vectors.

- $w^* = \sum_{i\in SV} \lambda_i^* y_i x_i$, where $SV \subseteq [n]$ denotes the index set of all the support vectors.
- For any $i \in SV$, $\langle w^*, x_i\rangle + b = y_i$. Thus, $b = y_i - \langle w^*, x_i\rangle = y_i - \sum_{j=1}^n \lambda_i^* y_j \boxed{\langle x_j, x_i\rangle}$

KKT optimality conditions of linear SVM

$$
\begin{aligned}
1 - y_i(\langle w^*, x_i\rangle + b^*) &\leq 0\,, \quad i = 1, \dots, n \\
\lambda^* &\succeq \mathbf{0} \\
\lambda_i^* \left(1 - y_i(\langle w^*, x_i\rangle + b^*)\right) &= 0\,, \quad i = 1, \dots, n \\
w^* - \textstyle\sum_{i=1}^n \lambda_i^* y_i x_i &= \mathbf{0} \\
\textstyle\sum_{i=1}^n \lambda_i^* y_i &= 0
\end{aligned}
\tag{KKT Linear SVM}
$$

Support vectors: set of points (x_i, y_i) for which $\lambda_i^* > 0$. Due to complementary slackness, $y_i(\langle w^*, x_i\rangle + b) = 1$ (or $\langle w^*, x_i\rangle + b = y_i$) for all support vectors.

$w^* = \sum_{i \in SV} \lambda_i^* y_i x_i$, where $SV \subseteq [n]$ denotes the index set of all the support vectors. ρ

For any $i \in SV$, $\langle w^*, x_i\rangle + b = y_i$. Thus, $\underline{b = y_i - \langle w^*, x_i\rangle} = y_i - \sum_{j=1}^n \lambda_i^* y_j \langle x_j, x_i\rangle$.

Exercise. Show that the largest margin of separation is equal to $\left[\frac{1}{\sum_{i \in SV} \lambda_i^*}\right]$. $= \rho$

Nonlinearly separable points

Suppose that the dataset $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is NOT linearly separable. That is, the convex hulls of $S^+ = (x \mid (x, y) \in S\ ,\ \ y = +1)$ and $S^- = (x \mid (x, y) \in S\ ,\ \ y = -1)$ are NOT disjoint.

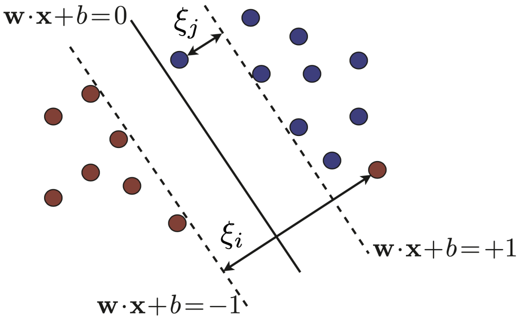


Figure: Nonlinearly separable points.

Linear SVM with soft margin

Slack variables. Introduce nonnegative variables ξ_i, $i = 1, \dots, n$ such that for all i,

$$y_i(w^\intercal x_i + b) + \xi_i \geq 1$$

We call ξ_i's as *slack variables*.

Linear SVM with soft margin

Slack variables. Introduce nonnegative variables $\xi_i, i = 1, \dots, n$ such that for all i,

$$y_i(w^\intercal x_i + b) + \xi_i \geq 1$$

$\|(\xi_1, \dots, \xi_n)\|_0$

We call ξ_i's as *slack variables*.

In this case, we train a linear SVM with **soft margin** using the following QP:

$$\begin{aligned} \underset{\xi\in\mathbb{R}^n,\ w\in\mathbb{R}^m,\ b\in\mathbb{R}}{\text{Minimize}} \quad & \frac{1}{2}\|w\|^2 + \boxed{\Psi(\xi_1, \dots, \xi_n)} \\ \text{Subject to} \quad & 1 - y_i(w^\intercal x_i + b) - \xi_i \leq 0 \quad, \ i = 1, \dots, n \\ & -\xi \preceq \mathbf{0} \end{aligned}$$

→ minimum misclassified pts.

(Soft-margin SVM)

where, $\Psi(\xi_1, \dots, \xi_n)$ is convex and typically taken to be $c \sum_{i=1}^{n} \xi_i^r$ for $r \geq 1$.

Linear SVM with soft margin

Slack variables. Introduce nonnegative variables $\xi_i, i = 1, \ldots, n$ such that for all i,

$$y_i(w^\intercal x_i + b) + \xi_i \geq 1$$

We call ξ_i's as *slack variables*.

In this case, we train a linear SVM with **soft margin** using the following QP:

$$\begin{array}{cll} \underset{\xi \in \mathbb{R}^n,\ w \in \mathbb{R}^m,\ b \in \mathbb{R}}{\text{Minimize}} & \frac{1}{2}\|w\|^2 + \Psi(\xi_1, \ldots, \xi_n) & \\ \text{Subject to} & 1 - y_i(w^\intercal x_i + b) - \xi_i \leq 0 & ,\ i = 1, \ldots, n \\ & -\xi \preceq \mathbf{0} & \end{array} \qquad \text{(Soft-margin SVM)}$$

where, $\Psi(\xi_1, \ldots, \xi_n)$ is convex and typically taken to be $c \sum_{i=1}^{n} \xi_i^r$ for $r \geq 1$.

We want ξ to be sparse. This can be approximately obtained by using $r = 1$.

Further readings

The lecture notes are based on Chapter 5 of "Foundations of Machine Learning" by M. Mohri, A. Rostamizadeh, and A. Talwalkar.

You may want to check out the following:

- **Learning guarantee of SVM:** Section 5.2.4 on *leave-one-out* (or *uniform stability*) analysis of SVM.
- **Generalizability of SVM:** Section 5.4 on *margin theory*.
- **Soft margin linear SVM:** More details can be found in Section 5.3.