SUPPORT VECTOR MACHINES (SVM)


Intuition: Don't solve a harder problem as an intermediate step towards your goal.

If you care about low-loss (misclassification rate), then optimize for that.

Goal: Classification

1. Generative approach: estimate \( \Pr(Y|X) \Rightarrow \text{Bayes} \Rightarrow P(Y|X) \Rightarrow \hat{y}_{\text{MAP}} \)

2. Discriminative approach: estimate \( \text{arg max} \ P(Y|X) \)

3. Discriminative #2: estimate \( \text{score}(x) = w^T x \) \( \text{sign}(w^T x + b) \)

Notation change: \( y \in \{+1, -1\} \) (rather than \( 0, 1 \))

\( w \) no longer has bias weight included

So \( w^T x + b > 0 \) rather than \( b \geq 0 \)

\( w \in \mathbb{R}^d \)
SVM [Linear SVM for now]

Many different \((\mathbf{w}, b)\) work well (actually perfectly) on training data. Which one should we choose?

Ans: The one with the largest margin!

Why?

- Trust me, I'm Vapnik
- Math will be easy & elegant
- There will be bounds on generalization
- Many different interpretations lead to this; has to be correct!

Mathematically (or pseudo-mathematically for now)

\[
\text{max } \frac{\mathbf{w}, b}{\text{Margin}} \\
\text{subject to } \mathbf{t} \text{ (Correct Classification on training data)}
\]
What is a correct classification?

$\Rightarrow \begin{cases} \langle w, x_i \rangle + b \geq 0 & \forall \ y_i = +1 \\ \langle w, x_i \rangle + b < 0 & \forall \ y_i = -1 \end{cases}$

Together:

$y_i (\langle w, x_i \rangle + b) \geq 0 \ \forall \ i$

What is Margin?

$\langle w, x \rangle + b = +r$

$\langle w, x \rangle + b = 0$

$\langle w, x \rangle + b = -r$

Parallel lines

$\Rightarrow$ Notice that we can always scale both $w$ & $b$ to get arbitrary $\pm r$. So let fix $r = +1$

Now:

$\langle w, x \rangle + b = +1$

$\langle w, x \rangle + b = 0$

$\langle w, x \rangle + b = -1$
Let $x^+$ be a pt on $w^T x + b = +1$

Let $x^-$ be a pt on $w^T x + b = -1$

Let line from $x^+$ to $x^-$ be orthogonal to our hyperplane $w^T x + b = 0$

$\Rightarrow ||x^+ - x^-|| = \text{Margin}$

We know $(x^+ - x^-) \perp$ hyperplane

$\Rightarrow (x^+ - x^-) \parallel w$

$\Rightarrow (x^+ - x^-) = \lambda w$

Let's left & right multiply by $w^T$

$w^T (x^+ - x^-) = \lambda w^T w$

$w^T x^+ - w^T x^- = \lambda w^T w$

$\Rightarrow (1-b) - (-1-b) = \lambda w^T w$

$\Rightarrow \lambda = \frac{2}{w^T w}$

$\Rightarrow \text{Margin} = ||x^+ - x^-||$

$= \frac{2}{w^T w} \cdot ||w|| = \frac{2}{||w||}$
Finally, Linear SVM with a "hard margin"

\[
\max_{\mathbf{\omega}, b} \frac{2}{\|\mathbf{\omega}\|} \quad \text{not very nice to optimize}
\]

\[
s.t. \quad y_i (\mathbf{\omega}^T \mathbf{x}_i + b) \geq 0
\]

2 changes

\[
\begin{cases}
\min_{\mathbf{\omega}, b} \frac{1}{2}\|\mathbf{\omega}\|^2 \\
\text{s.t.} \quad y_i (\mathbf{\omega}^T \mathbf{x}_i + b) \geq (1) \quad \forall i
\end{cases}
\]

why?

Very well studied optimization problem

→ Quadratic Program (QP)

→ \( \mathbf{\omega}^T \mathbf{\omega} \) = quadratic in \( \mathbf{\omega} \) to \( b \)

→ Convex

→ Hessian \( = \frac{1}{2} \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix} \geq 0 \) (PSD)

→ \( d+1 \) variables, \( N \) constraints

→ Standard QP solvers will solve this optimally for you.
③ Soft-Margin Linear SVM

→ What if data is not linearly separable?
→ OP will be infeasible ⇒ no solution

→ How can we allow for some mistakes?
→ Idea: allow some “slack”

allow every data-point to violate constraint by some “slack” \( \epsilon_i \).

Picture:

\[
\begin{align*}
\text{min} & \quad \frac{1}{2} w^T w + C \sum_{i=1}^{N} \epsilon_i \\
\text{subject to} & \quad w^T x_i + b \geq 1 - \epsilon_i, \\
& \quad \epsilon_i \geq 0
\end{align*}
\]

\( w, b, \epsilon_i \) 

\( w^T x_i + b \geq 1 - \epsilon_i \) 

\( \epsilon_i \geq 0 \) 

→ violation in margin 

→ can’t set slack to negative to give the objective
Still a convex QP

\[ \begin{align*}
\min & \quad d + 1 + N \\
\text{subject to} & \quad \langle w, e_i \rangle \\
& \quad b \\
\end{align*} \]

2N constraints

C = trade-off parameter

C = 0 \implies w^* = 0, b^* = 0 (opt soln)

C = \infty \implies Hard-Margin SVM (might be impossible for linearly non-separable data)

Hinge-Loss

Hinge Function is defined as:

\[ h(z) = \max \{ 0, 1 - z \} \]

In SVM QP

\[ \begin{align*}
& \quad \max \left\{ e_i \geq 1 - y_i (w^T x_i + b) \right\} \\
& \quad e_i \geq 0 \\
\end{align*} \]

\[ \implies e_i \geq \max \left\{ 0, 1 - y_i (w^T x_i + b) \right\} \]

Almost like \( h(y_i (w^T x_i + b)) \) but inequality lets fix that
Let $S_i = \text{score by SVM of } i\text{th data-point} = w^T x_i + b$

$$\epsilon_i^* = \max\{0, 1 - y_i S_i\}$$

Claim: Let $(w^*, b^*, \epsilon_i^*)$ be QP optimum.

$$\epsilon_i^* = \max\{0, 1 - y_i S_i^*\}$$

Proof: By contradiction. Assume equality doesn't hold

$$\Rightarrow \epsilon_i^* > \max\{0, 1 - y_i (w_i^* x_i + b^*)\}$$

Define $\epsilon_i^{\text{new}} = \max\{0, 1 - y_i (w_i^* x_i + b)\}$

By hinge loss construction

$$\left\{ \begin{array}{l}
\epsilon_i^{\text{new}} > 0 \\
\epsilon_i^{\text{new}} > 1 - y_i (w_i^* x_i + b)
\end{array} \right\}$$

$$\Rightarrow (w^*, b^*, \epsilon_i^{\text{new}}) \text{ is a FEASIBLE soln to SVM QP}$$

Also $(w^*, b^*, \epsilon_i^{\text{new}})$ reduces objective by

$$C \sum_{i=1}^n (\epsilon_i^* - \epsilon_i^{\text{new}}) > 0$$

$$\Rightarrow (w^*, b^*, \epsilon^*) \text{ is NOT OPT.}$$

$$\Rightarrow \text{contradiction}$$

$$\Rightarrow \text{QED}$$
So \( a^* = \max \{ 0, 1 - y_i (w^T x_i + b^*) \} \)

\[
\Rightarrow \text{SVM QP can be re-written as}
\]

\[
\min_{w, b} \frac{1}{2} w^T w + C \sum_{i=1}^{N} h(y_i (w^T x_i + b))
\]

Unconstrained QP where

\[ h(z) = \max \{ 0, 1 - z \} \quad \text{Hinge-Loss} \]

\[ h(t) = \begin{cases} 1 & \text{if } t < 1 \\ 0 & \text{otherwise} \end{cases} \]

\[ h(s) = 0 \]

Comparison between LR and SVM

\[ \Rightarrow \text{SVM: } \min_{w} \text{hinge-loss} + \frac{1}{2} w^T w \]

\[ \Rightarrow \text{LR: } \max_{w} \log P(y | x, w) - \lambda w^T w \]

Can we compare the fuel teams?

\[ \Rightarrow \text{Consider just 1 training pt } (x_1, y_1 = +1) \]

[Datasets are never this small, but just for illustration]
\[ P(Y = 1 | x, w) = \frac{1}{1 + e^{-wx - b}} \]

\[ w, b \sim N(0, \sigma I) \]

\[ b_{\text{MAP}} = \arg \max \log P(Y = 1 | x, w, b) - \frac{1}{2} \sigma^2 \]

\[ = \arg \max \log \left[ \frac{1}{1 + e^{-wx - b}} \right] - \frac{1}{2} \sigma^2 \]

\[ = \arg \min \frac{1}{2} \| w \|^2 + \text{hinge}(wx + b) \]

\[ = \arg \min \frac{1}{2} \| w \|^2 + \log (1 + e^{-wx - b}) \]

\[ h(s_i) \]

\[ \text{log-softmax} \]

\[ S_i = w^T x_i + b \]

\[ \text{max Margin} \]

\[ \text{Correct Class} \]