Machine Learning and Computational Statistics
Support Vector Machine - SVM

Nhung Le

Intro: This document consists of concepts and exercises related to SVM.

1 Key Concepts

1. SVM set-up
   - Hypothesis space $\mathcal{F} = \langle f(x) = w^T x + b | w \in \mathbf{R}^d, b \in \mathbf{R} \rangle$
   - l_2 regularization
   - The margin: $m = yf(x)$, which is a measure of how correct we are. We want to maximize the margin.
   - Hinge Loss: $l_{Hinge} = max\{1-m, 0\} = \{1-m\}_+$. Hinge loss is convex, upper bound on 0-1 loss, but not differentiable at $m = 1$. We have margin error when m < 1
   - Objective function

     $$J(w) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n} max(0, 1 - y_i[w^T x_i + b])$$

   - SVM Key Take-aways
2. SVM Visualization
   - Sides of the hyperplane $w^T v = a$, where a defines the value of the hyperplane $w^T v$.
   - Signed distance from x to Hyperplane $w^T x$: If we have a vector $x \in \mathbf{R}^d$ and a hyperplane $cH = \{v | w^T v = b\}$, we can measure the distance from x to H by:

     $$d(x, cH) = |\frac{w^T x - b}{\| w \|}|$$

     Without the absolute value we get the signed distance.
   - Hard Margin SVM: require linearly separable data.
     - Linearly separable: we say (x_i, y_i) for $i = 1, \cdots, n$ are linearly separable if there is a $w \in \mathbf{R}^d$ and $b \in \mathbf{R}$ such that $y_i(w^T x_i - b) > 0$ for all i. The set $\langle v \in \mathbf{R}^d | w^T v - b = 0 \rangle$ is called a separating hyperplane.

- Geometric margin: Let cH be a hyperplane that separates the data (x_i, y_i) for $i = 1, \cdots, n$. The geometric margin of this hyperplane is $\min_i d(x_i, cH)$, or the distance from the hyperplane to the closest data point.
- Maximizing margin

Maximizing margin

We want to:

$$\text{maximize} \min_i d(x_i, H)$$

Remember:

$$d(x_i, H) = \left| \frac{w^T x_i - b}{\|w\|_2} \right| = \frac{y_i(w^T x_i - b)}{\|w\|_2}.$$

So:

$$\text{maximize}_{w,b} \min_i \frac{y_i(w^T x_i - b)}{\|w\|_2}.$$

Note, if $M = \min_i \frac{y_i(w^T x_i - b)}{\|w\|_2}$, then $\frac{y_i(w^T x_i - b)}{\|w\|_2} \geq M$ for all i

Maximizing margin

We can rewrite this in a more standard form:

$$\begin{array}{ll} \text{maximize}_{w,b,M} & M \\ \text{subject to} & \frac{y_i(w^T x_i - b)}{\|w\|_2} \geq M \quad \text{for all } i. \end{array}$$

fix $\|w\|_2 = 1/M$ to obtain

$$\begin{array}{ll} \text{maximize}_{w,b} & 1/\|w\|_2 \\ \text{subject to} & y_i(w^T x_i - b) \geq 1 \quad \text{for all } i. \end{array}$$

To find the optimal w, a we can instead solve the minimization problem

$$\begin{array}{ll} \text{minimize}_{w,b} & \|w\|_2^2 \\ \text{subject to} & y_i(w^T x_i - b) \geq 1 \quad \text{for all } i. \end{array}$$

- Soft Margin SVM: remove restriction on linearly separable data, and allow vectors to violate the geometric margin requirements, but at a penalty.
  - Objective SVM function:
$$\min_{w,a} \parallel w \parallel + \frac{C}{n} \sum_{i=1}^{n} \epsilon_i$$
$$\text{subject to } y_i(w^T x_i + a) \geq 1 - \epsilon_i \text{ for all } i$$
$$\epsilon_i \geq 0 \texttt{for all i}$$
- Slack variable $\epsilon_i \geq 0$ corresponding x_i violates the geometric margin condition. ϵ_i measures the size of the violation in multiples of the geometric margin. For example, $\epsilon_i = 1$ means x_i lies on the decision hyperplane $w^T v + a = 0$, and $\epsilon_i = 3$ means x_i lies 2 margin widths past the decision hyperplane $w^T v + a = 0$.
- Violation penalty C
- Support vectors: some subset of the x_i that either lie on the margin boundary $y_i(w^T x_i + a) = 1$, or violate the margin boundary $y_i(w^T x_i + a) < 1, \epsilon_i > 0$

3. SVM Notes
4. Uniqueness of SVM Solution
5. The SVM Dual Problem

SVM as a Quadratic Program

- The SVM optimization problem is equivalent to

$$
\begin{aligned}
\text{minimize} \quad & \frac{1}{2}\|w\|^2+\frac{c}{n}\sum_{i=1}^{n}\xi_i \\
\text{subject to} \quad & -\xi_i \leqslant 0 \text{ for } i=1,\ldots,n \\
& \left(1-y_i\left[w^Tx_i+b\right]\right)-\xi_i \leqslant 0 \text{ for } i=1,\ldots,n
\end{aligned}
$$

- Differentiable objective function
- $n+d+1$ unknowns and $2n$ affine constraints.
- A quadratic program that can be solved by any off-the-shelf QP solver.
- Let's learn more by examining the dual.

SVM Lagrange Multipliers

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i \\
\text{subject to} & -\xi_i \leqslant 0 \text{ for } i = 1, \ldots, n \\
& \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1, \ldots, n
\end{array}
$$

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0$

$$
L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}||w||^2 + \frac{c}{n}\sum_{i=1}^{n} \xi_i + \sum_{i=1}^{n} \alpha_i\left(1 - y_i\left[w^T x_i + b\right] - \xi_i\right) + \sum_{i=1}^{n} \lambda_i\left(-\xi_i\right)
$$

SVM Lagrangian

- The Lagrangian for this formulation is

$$
\begin{aligned}
& L(w,b,\xi,\alpha,\lambda) \\
= \; & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1-y_i\left[w^Tx_i+b\right]-\xi_i\right) - \sum_{i}\lambda_i\xi_i \\
= \; & \frac{1}{2}w^Tw + \sum_{i=1}^{n}\xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right) + \sum_{i=1}^{n}\alpha_i\left(1-y_i\left[w^Tx_i+b\right]\right).
\end{aligned}
$$

- Primal and dual:

$$
\begin{aligned}
p^* &= \inf_{w,\xi,b}\sup_{\alpha,\lambda\succeq 0} L(w,b,\xi,\alpha,\lambda) \\
&\geqslant \sup_{\alpha,\lambda\succeq 0}\inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) = d^*
\end{aligned}
$$

- Do we have $p^* = d^*$?

Strong Duality by Slater's constraint qualification

- The SVM optimization problem:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & -\xi_i \leqslant 0 \text{ for } i = 1,\ldots,n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leqslant 0 \text{ for } i = 1,\ldots,n \end{array}$$

- Convex problem + affine constraints $\Longrightarrow$ strong duality iff problem is feasible
- Constraints are satisfied by $w = b = 0$ and $\xi_i = 1$ for $i = 1,\ldots,n$,
  - so **we have strong duality** $\Longrightarrow$

$$\begin{aligned} p^* &= \inf_{w,\xi,b} \sup_{\alpha,\lambda\succeq 0} L(w,b,\xi,\alpha,\lambda) \\ &= \sup_{\alpha,\lambda\succeq 0} \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) = d^* \end{aligned}$$

SVM Dual Function

- Lagrange dual is the inf over primal variables of the Lagrangian:

$$\begin{aligned} g(\alpha,\lambda) &= \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) \\ &= \inf_{w,b,\xi}\left[\frac{1}{2}w^Tw+\sum_{i=1}^n \xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right)+\sum_{i=1}^n \alpha_i\left(1-y_i\left[w^Tx_i+b\right]\right)\right] \end{aligned}$$

- Taking inf of convex and differentiable function of w, b, ξ.
    - Quadratic in w and linear in ξ and b.
- Thus optimal point iff $\partial_w L = 0$ $\partial_b L = 0$ $\partial_\xi L = 0$
- Note: $g(\alpha,\lambda) = -\infty$ when $\frac{c}{n}-\alpha_i-\lambda_i \neq 0$. (send $\xi_i \to \pm\infty$). This inf is NOT an optimum because it is never attained.

SVM Dual Function: First Order Conditions

Lagrange dual function is the inf over primal variables of L:

$$
\begin{aligned}
g(\alpha,\lambda) &= \inf_{w,b,\xi} L(w,b,\xi,\alpha,\lambda) \\
&= \inf_{w,b,\xi}\left[\frac{1}{2}w^Tw+\sum_{i=1}^n \xi_i\left(\frac{c}{n}-\alpha_i-\lambda_i\right)+\sum_{i=1}^n \alpha_i\left(1-y_i\left[w^Tx_i+b\right]\right)\right]
\end{aligned}
$$

$$\partial_w L = 0 \iff w-\sum_{i=1}^n \alpha_i y_i x_i = 0 \iff \boxed{w=\sum_{i=1}^n \alpha_i y_i x_i}$$

$$\partial_b L = 0 \iff -\sum_{i=1}^n \alpha_i y_i = 0 \iff \boxed{\sum_{i=1}^n \alpha_i y_i = 0}$$

$$\partial_{\xi_i} L = 0 \iff \frac{c}{n}-\alpha_i-\lambda_i = 0 \iff \boxed{\alpha_i+\lambda_i=\frac{c}{n}}$$

SVM Dual Function

- Substituting these conditions back into L, the second term disappears.
- First and third terms become

$$
\begin{aligned}
\frac{1}{2}w^Tw &= \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
\sum_{i=1}^{n}\alpha_i(1-y_i\left[w^Tx_i+b\right]) &= \sum_{i=1}^{n}\alpha_i - \sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i - b\underbrace{\sum_{i=1}^{n}\alpha_i y_i}_{=0}.
\end{aligned}
$$

- Putting it together, the dual function is

$$
g(\alpha,\lambda) = \begin{cases} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_j^T x_i & \substack{\sum_{i=1}^{n}\alpha_i y_i = 0 \\ \alpha_i+\lambda_i=\frac{c}{n},\ \text{all } i} \\ -\infty & \text{otherwise.} \end{cases}
$$

SVM Dual Problem

- The **dual function** is

$$g(\alpha,\lambda) = \begin{cases} \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_j^T x_i & \substack{\sum_{i=1}^n \alpha_i y_i = 0 \\ \alpha_i + \lambda_i = \frac{c}{n},\ \text{all } i} \\ -\infty & \text{otherwise.} \end{cases}$$

- The **dual problem** is $\sup_{\alpha,\lambda \succeq 0} g(\alpha,\lambda)$:

$$\begin{aligned} \sup_{\alpha,\lambda} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i\alpha_j y_i y_j x_j^T x_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & \alpha_i + \lambda_i = \frac{c}{n} \quad \alpha_i, \lambda_i \geqslant 0,\ i = 1,\dots,n \end{aligned}$$

6. SVM and Complementary Slackness

Support Vectors and The Margin

- Recall "**slack variable**" $\xi_i^* = \max(0, 1 - y_i f^*(x_i))$ is the hinge loss on (x_i, y_i).
- Suppose $\xi_i^* = 0$.
- Then $y_i f^*(x_i) \geqslant 1$
  - "on the margin" $(=1)$, or
  - "on the good side" (>1)

Complementary Slackness Conditions

- Recall our primal constraints and Lagrange multipliers:

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leqslant 0$
α_i	$(1 - y_i f(x_i)) - \xi_i \leqslant 0$

- Recall first order condition $\nabla_{\xi_i} L = 0$ gave us $\lambda_i^* = \frac{c}{n} - \alpha_i^*$.
- By strong duality, we must have **complementary slackness**:

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

Consequences of Complementary Slackness

- By strong duality, we must have **complementary slackness**:

$$\alpha_i^* \left(1 - y_i f^*(x_i) - \xi_i^*\right) = 0$$
$$\left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

- If $y_i f^*(x_i) > 1$ then the margin loss is $\xi_i^* = 0$, and we get $\alpha_i^* = 0$.
- If $y_i f^*(x_i) < 1$ then the margin loss is $\xi_i^* > 0$, so $\alpha_i^* = \frac{c}{n}$.
- If $\alpha_i^* = 0$, then $\xi_i^* = 0$, which implies no loss, so $y_i f^*(x) \geqslant 1$.
- If $\alpha_i^* \in \left(0, \frac{c}{n}\right)$, then $\xi_i^* = 0$, which implies $1 - y_i f^*(x_i) = 0$.

Complementary Slackness Results: Summary

$$
\begin{aligned}
\alpha_i^* = 0 \quad &\Longrightarrow \quad y_i f^*(x_i) \geqslant 1 \\
\alpha_i^* \in \left(0, \frac{c}{n}\right) \quad &\Longrightarrow \quad y_i f^*(x_i) = 1 \\
\alpha_i^* = \frac{c}{n} \quad &\Longrightarrow \quad y_i f^*(x_i) \leqslant 1
\end{aligned}
$$

$$
\begin{aligned}
y_i f^*(x_i) < 1 \quad &\Longrightarrow \quad \alpha_i^* = \frac{c}{n} \\
y_i f^*(x_i) = 1 \quad &\Longrightarrow \quad \alpha_i^* \in \left[0, \frac{c}{n}\right] \\
y_i f^*(x_i) > 1 \quad &\Longrightarrow \quad \alpha_i^* = 0
\end{aligned}
$$

Support Vectors

- If α^* is a solution to the dual problem, then primal solution is

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

  with $\alpha_i^* \in [0, \frac{c}{n}]$.
- The x_i's corresponding to $\alpha_i^* > 0$ are called **support vectors.**
- Few margin errors or "on the margin" examples $\implies$ **sparsity in input examples**.

Complementary Slackness To Get b^*

The Bias Term: b

- For our SVM primal, the complementary slackness conditions are:

$$\alpha_i^* \left(1 - y_i \left[x_i^T w^* + b\right] - \xi_i^*\right) = 0 \tag{1}$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0 \tag{2}$$

- Suppose there's an i such that $\alpha_i^* \in \left(0, \frac{c}{n}\right)$.
- (2) implies $\xi_i^* = 0$.
- (1) implies

$$\begin{aligned} & y_i \left[x_i^T w^* + b^*\right] = 1 \\ \iff \quad & x_i^T w^* + b^* = y_i \ (\text{use } y_i \in \{-1, 1\}) \\ \iff \quad & \boxed{b^* = y_i - x_i^T w^*} \end{aligned}$$

The Bias Term: b

- The optimal b is
$$b^* = y_i - x_i^T w^*$$
- We get the same b^* for any choice of i with $\alpha_i^* \in \left(0, \frac{c}{n}\right)$
  - **With exact calculations!**
- With numerical error, more robust to average over all eligible i's:
$$b^* = \text{mean}\left\{ y_i - x_i^T w^* \mid \alpha_i^* \in \left(0, \frac{c}{n}\right) \right\}.$$
- If there are no $\alpha_i^* \in \left(0, \frac{c}{n}\right)$?
  - Then we have a **degenerate SVM training problem**[1] ($w^* = 0$).

[1] See Rifkin et al.'s "A Note on Support Vector Machine Degeneracy", an MIT AI Lab Technical Report.