CMSC 422 Introduction to Machine Learning
Lecture 23 Support Vector Machines I

Furong Huang / furongh@cs.umd.edu

Slides adapted from Prof Carpuat and Duraiswami
Back to linear classification

Last time: we’ve seen that kernels can help capture non-linear patterns in data while keeping the advantages of a linear classifier.

Today: Support Vector Machines

A hyperplane-based classification algorithm
Highly influential
Backed by solid theoretical grounding (Vapnik & Cortes, 1995)
Easy to kernelize
The Maximum Margin Principle

Find the hyperplane with maximum separation margin on the training data
Margin of a data set $D$

- Margin of a dataset $D$ with respect to a hyperplane $w^T x + b$

\[
\text{margin}(D, w, b) = \begin{cases} 
\min_{(x,y) \in D} \frac{y(w^T x + b)}{||w||} & \text{if } w \text{ separates } D \\
-\infty & \text{otherwise}
\end{cases}
\]

Distance between the hyperplane $(w,b)$ and the nearest point in $D$

- Margin of a dataset $D$

\[
\text{margin}(D) = \sup_{w,b} \text{margin}(D, w, b)
\]

Largest attainable margin on $D$

\[
\text{margin}(D) \geq 0
\]
Support Vector Machine (SVM)

A hyperplane based linear classifier defined by $\mathbf{w}$ and $b$

Prediction rule: $y = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$

**Given:** Training data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$

**Goal:** Learn $\mathbf{w}$ and $b$ that achieve the maximum margin
Characterizing the margin

Let’s assume the entire training data is correctly classified by \((w, b)\) that achieve the maximum margin.

- Assume the hyperplane is such that:
  - \(w^T x_n + b \geq 1\) for \(y_n = +1\)
  - \(w^T x_n + b \leq -1\) for \(y_n = -1\)
  - Equivalently, \(y_n (w^T x_n + b) \geq 1\)
    \[ \Rightarrow \min_{1 \leq n \leq N} |w^T x_n + b| = 1 \]
- The hyperplane’s margin:
  \[ \gamma = \min_{1 \leq n \leq N} \frac{|w^T x_n + b|}{|w|} = \frac{1}{|w|} \]
The Optimization Problem

We want to maximize the margin $\gamma = \frac{1}{||w||}$

Maximizing the margin $\gamma = \text{minimizing} \quad ||w||$ (the norm)

Our optimization problem would be:

Minimize $f(w, b) = \frac{||w||^2}{2}$

subject to $y_n(w^T x_n + b) \geq 1$, $\forall \ n = 1, \ldots, N$
Large Margin = Good Generalization

Intuitively, large margins mean good generalization

Large margin $\Rightarrow$ small $||w||$
small $||w|| \Rightarrow$ regularized/simple solutions

(Learning theory gives a more formal justification)
Solving the SVM Optimization Problem

Our optimization problem is:

\[
\text{Minimize } f(w) = \frac{\|w\|^2}{2} \\
\text{subject to } 1 \leq y_n(w^T x_n + b), \quad \forall \ n = 1, \ldots, N
\]

Introducing Lagrange Multipliers \(\alpha_n\) \((n = \{1, \ldots, N\})\), one for each constraint, leads to the Lagrangian:

\[
\text{Minimize } L(w, b, \alpha) = \frac{\|w\|^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(w^T x_n + b)\} \\
\text{subject to } \alpha_n \geq 0; \quad \forall \ n = 1, \ldots, N
\]
Lagrange Dual Function

An optimization problem in standard form:

\[
\begin{align*}
& \text{minimize} & & f_0(x) \\
& \text{subject to} & & f_i(x) \leq 0, \quad i = 1, 2, \ldots, m \\
& & & h_i(x) = 0, \quad i = 1, 2, \ldots, p \\
\end{align*}
\]

Variables: \( x \in \mathbb{R}^n \). Assume nonempty feasible set

Optimal value: \( p^* \). Optimizer: \( x^* \).

Idea: augment objective with a weighted sum of constraints

- **Lagrangian**: \( L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \mu_i h_i(x) \)
- **Lagrange multipliers (dual variables)**: \( \lambda \geq 0, \mu \)
- **Lagrange dual function**: \( g(\lambda, \mu) = \inf_x L(x, \lambda, \mu) \)
- **Lower bound on Optimal Value**: \( g(\lambda, \mu) \leq p^*, \forall \lambda \geq 0, \mu \)
Lagrange Dual Problem

- Lower bound from Lagrange dual function depends on $(\lambda, \mu)$. What is the best lower bound that can be obtained from Lagrange dual function?

\[
\begin{align*}
\text{maximize} & \quad g(\lambda, \mu) \\
\text{subject to} & \quad \lambda \geq 0
\end{align*}
\]

This is the Lagrange dual problem with dual variables $(\lambda, \mu)$.

- Dual objective function is always a concave function since it’s the infimum of a family of affine functions in $(\lambda, \mu)$. Therefore: convex optimization
Solving the SVM Optimization Problem

Our optimization problem is:

\[
\text{Minimize } f(w, b) = \frac{\|w\|^2}{2}
\]
subject to \(1 \leq y_n(w^T x_n + b), \quad \forall \ n = 1, \ldots, N\)

Introducing Lagrange Multipliers \(\alpha_n\) \((n = \{1, \ldots, N\})\), one for each constraint, leads to the Lagrangian:

\[
\text{Minimize } L(w, b, \alpha) = \frac{\|w\|^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(w^T x_n + b)\}
\]
subject to \(\alpha_n \geq 0; \quad \forall \ n = 1, \ldots, N\)
Solving the SVM Optimization Problem

Take (partial) derivatives of $L_P$ w.r.t. $w$, $b$ and set them to zero

$$\frac{\partial L_P}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^{N} \alpha_n y_n x_n, \quad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting these in the Primal Lagrangian $L_P$ gives the Dual Lagrangian

Maximize $L_D(w, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m, n=1}^{N} \alpha_m \alpha_n y_m y_n (x_m^T x_n)$

subject to $\sum_{n=1}^{N} \alpha_n y_n = 0, \quad \alpha_n \geq 0; \quad n = 1, \ldots, N$
Solving the SVM Optimization Problem

Take (partial) derivatives of $L_P$ w.r.t. $w$, $b$ and set them to zero

$$L_P = \sum_{n=1}^{N} \alpha_n y_n x_n, \quad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting these into the Primal Lagrangian $L_P$ gives the Dual Lagrangian

Maximize $L_D(w, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (x_m^T x_n)$

subject to $\sum_{n=1}^{N} \alpha_n y_n = 0, \quad \alpha_n \geq 0; \quad n = 1, \ldots, N$

A Quadratic Program for which many off-the-shelf solvers exist
SVM: the solution!

Once we have the $\alpha_n$'s, $\mathbf{w}$ and $b$ can be computed as:

$$
\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n
$$

$$
b = -\frac{1}{2} \left( \min_{n:y_n=+1} \mathbf{w}^T \mathbf{x}_n + \max_{n:y_n=-1} \mathbf{w}^T \mathbf{x}_n \right)
$$

**Note:** Most $\alpha_n$'s in the solution are zero (sparse solution)

- **Reason:** Karush-Kuhn-Tucker (KKT) conditions
- For the optimal $\alpha_n$'s
  $$
  \alpha_n \{1 - y_n (\mathbf{w}^T \mathbf{x}_n + b)\} = 0
  $$

  $\alpha_n$ is non-zero only if $\mathbf{x}_n$ lies on one of the two margin boundaries, i.e., for which $y_n (\mathbf{w}^T \mathbf{x}_n + b) = 1$

  These examples are called support vectors
- Support vectors “support” the margin boundaries
What if the data is not separable?

Non-separable case: We will allow misclassified training examples but we want their number to be minimized by minimizing the sum of slack variables \( \sum_{n=1}^{N} \xi_n \).

The optimization problem for the non-separable case is:

Minimize \( f(w, b) = \frac{||w||^2}{2} + C \sum_{n=1}^{N} \xi_n \)

subject to \( y_n(w^T x_n + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \quad n = 1, \ldots, N \)
Support Vector Machines

Find the max margin linear classifier for a dataset

Discovers “support vectors”, the training examples that “support” the margin boundaries

Hard margin vs soft margin SVM

- Hard margin: assume the data is linearly separable (today’s lecture)
- Soft margin: more general case (next time!)
Recall KKT conditions

Remember duality

Given a minimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

subject to $h_i(x) \leq 0, \ i = 1, \ldots m$

$$\ell_j(x) = 0, \ j = 1, \ldots r$$

we defined the **Lagrangian:**

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

and **Lagrange dual function:**

$$g(u, v) = \min_{x \in \mathbb{R}^n} L(x, u, v)$$

Slides credit to Geoff Gordon & Ryan Tibshirani
Recall KKT conditions

The subsequent dual problem is:

$$\max_{u \in \mathbb{R}^m, v \in \mathbb{R}} g(u, v)$$
subject to $u \geq 0$

Important properties:

- Dual problem is always convex, i.e., $g$ is always concave (even if primal problem is not convex)
- The primal and dual optimal values, $f^*$ and $g^*$, always satisfy weak duality: $f^* \geq g^*$
- Slater’s condition: for convex primal, if there is an $x$ such that

$$h_1(x) < 0, \ldots, h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots, \ell_r(x) = 0$$

then strong duality holds: $f^* = g^*$. (Can be further refined to strict inequalities over nonaffine $h_i$, $i = 1, \ldots, m$)
Recall KKT conditions

**Duality gap**

Given primal feasible $x$ and dual feasible $u, v$, the quantity

$$f(x) - g(u, v)$$

is called the **duality gap** between $x$ and $u, v$. Note that

$$f(x) - f^* \leq f(x) - g(u, v)$$

so if the duality gap is zero, then $x$ is primal optimal (and similarly, $u, v$ are dual optimal).

From an algorithmic viewpoint, provides a stopping criterion: if $f(x) - g(u, v) \leq \epsilon$, then we are guaranteed that $f(x) - f^* \leq \epsilon$.

Very useful, especially in conjunction with iterative methods ... more dual uses in coming lectures.

Slides credit to Geoff Gordon & Ryan Tibshirani
Recall KKT conditions

Karush-Kuhn-Tucker conditions

Given general problem

\[
\min_{x \in \mathbb{R}^n} f(x)
\]

subject to \( h_i(x) \leq 0, \ i = 1, \ldots m \)
\( \ell_j(x) = 0, \ j = 1, \ldots r \)

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- \( 0 \in \partial f(x) + \sum_{i=1}^{m} u_i \partial h_i(x) + \sum_{j=1}^{r} v_j \partial \ell_j(x) \) (stationarity)
- \( u_i \cdot h_i(x) = 0 \) for all \( i \) (complementary slackness)
- \( h_i(x) \leq 0, \ \ell_j(x) = 0 \) for all \( i, j \) (primal feasibility)
- \( u_i \geq 0 \) for all \( i \) (dual feasibility)

Slides credit to Geoff Gordon & Ryan Tibshirani
Recall KKT conditions

Necessity

Let $x^*$ and $u^*, v^*$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater’s condition). Then

$$f(x^*) = g(u^*, v^*)$$

$$= \min_{x \in \mathbb{R}^n} f(x) + \sum_{i=1}^{m} u^*_i h_i(x) + \sum_{j=1}^{r} v^*_j \ell_j(x)$$

$$\leq f(x^*) + \sum_{i=1}^{m} u^*_i h_i(x^*) + \sum_{j=1}^{r} v^*_j \ell_j(x^*)$$

$$\leq f(x^*)$$

In other words, all these inequalities are actually equalities.

Slides credit to Geoff Gordon & Ryan Tibshirani
Recall KKT conditions

Two things to learn from this:

- The point $x^*$ minimizes $L(x, u^*, v^*)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^*, v^*)$ must contain 0 at $x = x^*$—this is exactly the **stationarity** condition.

- We must have $\sum_{i=1}^m u_i^* h_i(x^*) = 0$, and since each term here is $\leq 0$, this implies $u_i^* h_i(x^*) = 0$ for every $i$—this is exactly **complementary slackness**.

Primal and dual feasibility obviously hold. Hence, we’ve shown:

> If $x^*$ and $u^*, v^*$ are primal and dual solutions, with zero duality gap, then $x^*, u^*, v^*$ satisfy the KKT conditions.

(Note that this statement assumes nothing a priori about convexity of our problem, i.e. of $f, h_i, \ell_j$)
Recall KKT conditions

**Sufficiency**

If there exists $x^*, u^*, v^*$ that satisfy the KKT conditions, then

$$g(u^*, v^*) = f(x^*) + \sum_{i=1}^{m} u_i^* h_i(x^*) + \sum_{j=1}^{r} v_j^* \ell_j(x^*)$$

$$= f(x^*)$$

where the first equality holds from stationarity, and the second holds from complementary slackness.

Therefore duality gap is zero (and $x^*$ and $u^*, v^*$ are primal and dual feasible) so $x^*$ and $u^*, v^*$ are primal and dual optimal. I.e., we've shown:

If $x^*$ and $u^*, v^*$ satisfy the KKT conditions, then $x^*$ and $u^*, v^*$ are primal and dual solutions.
Recall KKT conditions

Putting it together

In summary, KKT conditions:
- always sufficient
- necessary under strong duality

Putting it together:

For a problem with strong duality (e.g., assume Slater’s condition: convex problem and there exists $x$ strictly satisfying non-affine inequality constraints),

$x^*$ and $u^*, v^*$ are primal and dual solutions

$\iff x^*$ and $u^*, v^*$ satisfy the KKT conditions

(Warning, concerning the stationarity condition: for a differentiable function $f$, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless $f$ is convex)

Slides credit to Geoff Gordon & Ryan Tibshirani