Slides adapted from Prof Carpuat and Duraiswami

CMSC 422 Introduction to Machine Learning

Lecture 23 Support Vector Machines I

Furong Huang / furongh@umd.edu

UNIVERSITY OF MARYLAND

Back to linear classification

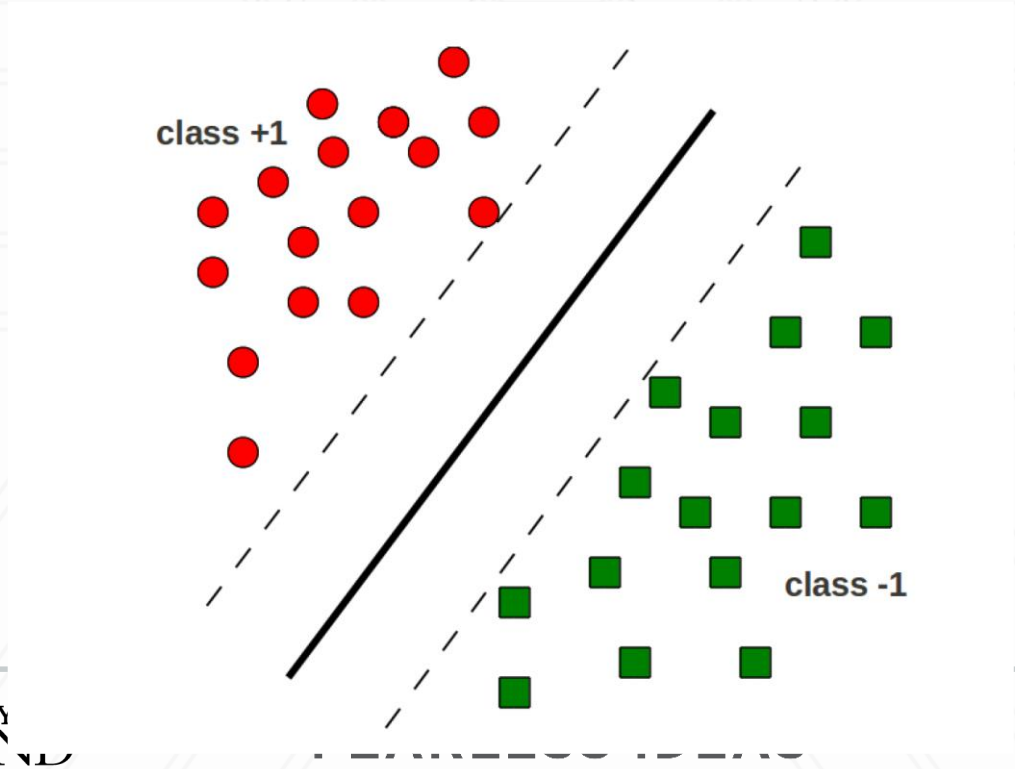
Last time: we've seen that kernels can help capture non-linear patterns in data while keeping the advantages of a linear classifier

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 $w^\top \mathbf{x} + \mathbf{b}$**

$$margin(\mathbf{D}, w, b) = \begin{cases} \min_{(x,y)\in\mathbf{D}} \dfrac{y(w^\top x + b)}{||w||} & \text{if } w \text{ separates } \mathbf{D} \\ -\infty & \text{otherwise} \end{cases} \quad (3.8)$$

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

$$\min\left(y(w^\top x + b)\right) \geq 0$$

- **Margin of a dataset D**

$$margin(\mathbf{D}) = \sup_{w,b} margin(\mathbf{D}, w, b) \quad (3.9)$$

Largest attainable margin on D

Support Vector Machine (SVM)

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

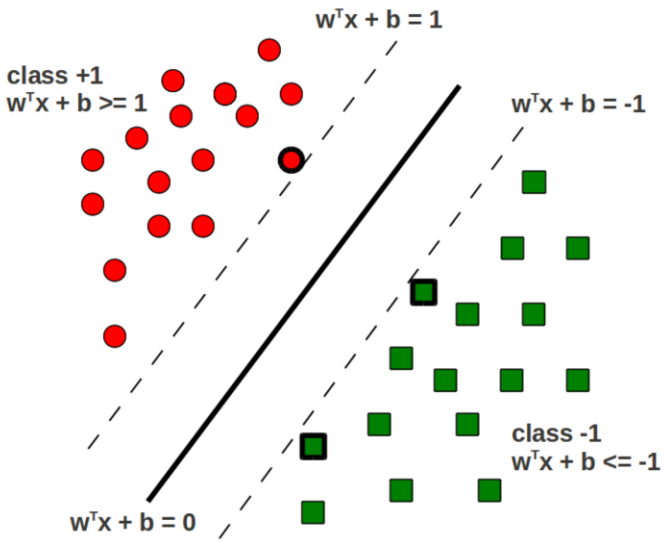
Prediction rule: $y = sign(\mathbf{w}^T\mathbf{x} + b)$

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)

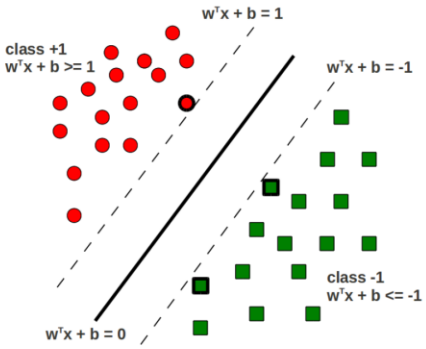


- Assume the hyperplane is such that
  - $\mathbf{w}^T\mathbf{x}_n + b \geq 1$ for $y_n = +1$
  - $\mathbf{w}^T\mathbf{x}_n + b \leq -1$ for $y_n = -1$
  - Equivalently, $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$
    $\Rightarrow \min_{1\leq n\leq N} |\mathbf{w}^T\mathbf{x}_n + b| = 1$
  - The hyperplane's margin:

$$\gamma = \min_{1\leq n\leq N} \frac{|\mathbf{w}^T\mathbf{x}_n + b|}{||\mathbf{w}||} = \frac{1}{||\mathbf{w}||}$$

The Optimization Problem

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



Maximizing the margin γ = minimizing $||\mathbf{w}||$ (the norm)

Our optimization problem would be:

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

$$\text{subject to} \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1, \qquad \forall\, n = 1, \ldots, N$$

Large Margin = Good Generalization

Intuitively, large margins mean good generalization

- Large margin => small $||\mathbf{w}||$
- small $||\mathbf{w}||$ => regularized/simple solutions

(Learning theory gives a more formal justification)

Solving the SVM Optimization Problem

Our optimization problem is:

$$\begin{aligned} &\text{Minimize} \quad f(\mathbf{w}, b) = \frac{||\mathbf{w}||^2}{2} \\ &\text{subject to} \quad 1 \leq y_n(\mathbf{w}^T \mathbf{x}_n + b), \qquad \forall\, n = 1, \ldots, N \end{aligned}$$

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

$$\begin{aligned} &\text{Minimize} \quad L(\mathbf{w}, b, \alpha) = \frac{||\mathbf{w}||^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(\mathbf{w}^T \mathbf{x}_n + b)\} \\ &\text{subject to} \quad \alpha_n \geq 0; \; \forall\; n = 1, \ldots, N \end{aligned}$$

Lagrange Dual Function

An optimization problem in standard form:

$$\begin{aligned} \text{minimize} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leq 0, \quad i = 1,2,\dots,m \\ & h_i(x) = 0, \quad i = 1,2,\dots,p \end{aligned}$$

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 (λ, μ). What is the best lower bound that can be obtained from Lagrange dual function?

$$\text{maxmize } g(\lambda, \mu)$$
$$\text{subject to } \lambda \geq 0$$

  This is the Lagrange dual problem with dual variables (λ, μ).

- Dual objective function is always a concave function since it's the infimum of a family of affine functions in (λ, μ). Therefore: convex optimization

Solving the SVM Optimization Problem

Our optimization problem is:

$$\begin{aligned} &\text{Minimize} \quad f(\mathbf{w}, b) = \frac{||\mathbf{w}||^2}{2} \\ &\text{subject to} \quad 1 \leq y_n(\mathbf{w}^T \mathbf{x}_n + b), \qquad \forall\, n = 1, \ldots, N \end{aligned}$$

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

$$\begin{aligned} &\text{Minimize} \quad L(\mathbf{w}, b, \alpha) = \frac{||\mathbf{w}||^2}{2} + \sum_{n=1}^{N} \alpha_n \{1 - y_n(\mathbf{w}^T \mathbf{x}_n + b)\} \\ &\text{subject to} \quad \alpha_n \geq 0; \; \forall\; n = 1, \ldots, N \end{aligned}$$

Solving the SVM Optimization Problem

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

$$\frac{\partial L_P}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n, \qquad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

Substituting these in the Primal Lagrangian L_P gives the Dual Lagrangian

$$\text{Maximize} \quad L_D(\mathbf{w}, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (\mathbf{x}_m^T \mathbf{x}_n)$$

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

http://cs229.stanford.edu/notes/cs229-notes3.pdf

Solving the SVM Optimization Problem

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

$$= \sum_{n=1}^{N} \alpha_n y_n \mathbf{x}_n, \qquad \frac{\partial L_P}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y_n = 0$$

A Quadratic Program for which many off-the-shelf solvers exist

Substituting thes the Primal Lagrangian L_P gives the Dual Lagrangian

$$\text{Maximize} \quad L_D(\mathbf{w}, b, \alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{m,n=1}^{N} \alpha_m \alpha_n y_m y_n (\mathbf{x}_m^T \mathbf{x}_n)$$

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

SVM: the solution!

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

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

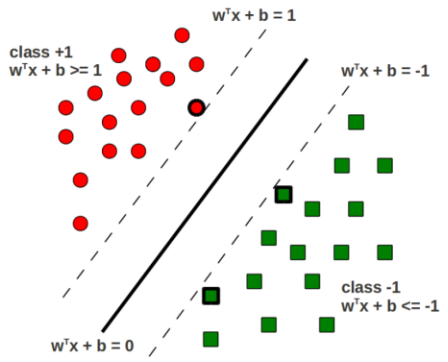
$$b = -\frac{1}{2}\left(\min_{n:y_n=+1} \mathbf{w}^T\mathbf{x}_n + \max_{n:y_n=-1} \mathbf{w}^T\mathbf{x}_n\right)$$

Note: Most α_n's in the solution are zero (sparse solution)

- Reason: Karush-Kuhn-Tucker (KKT) conditions
- For the optimal α_n's

$$\alpha_n\{1 - y_n(\mathbf{w}^T\mathbf{x}_n + b)\} = 0$$

- α_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
  $\Rightarrow$ by *minimizing* the sum of slack variables $(\sum_{n=1}^{N} \xi_n)$

The optimization problem for the non-separable case

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

$$\text{subject to} \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \quad \xi_n \geq 0 \qquad n = 1, \ldots, N$$

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: assme the data is linearly separable (today’s lecture)
- Soft margin: more general case (next time!)

Recall KKT conditions

Remember duality

Given a minimization problem

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \dots m \\ & \ell_j(x) = 0, \quad j = 1, \dots r \end{aligned}$$

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:

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

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, $f^\star$ and $g^\star$, always satisfy weak duality: $f^\star \geq g^\star$
- 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: $f^\star = g^\star$. (Can be further refined to strict inequalities over nonaffine h_i, $i = 1, \ldots m$)

Slides credit to Geoff Gordon & Ryan Tibshirani

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

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

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 $f(x) - f^\star \leq \epsilon$

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

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & h_i(x) \leq 0, \quad i = 1, \ldots m \\ & \ell_j(x) = 0, \quad j = 1, \ldots r \end{aligned}$$

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

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_{x \in \mathbb{R}^n} \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\leq f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &\leq f(x^\star) \end{aligned}$$

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 $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the **stationarity** condition
- We must have $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly **complementary slackness**

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

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

(Note that this statement assumes nothing a priori about convexity of our problem, i.e. of f, h_i, ℓ_j)

Slides credit to Geoff Gordon & Ryan Tibshirani

Recall KKT conditions

Sufficiency

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

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

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

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

> If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Slides credit to Geoff Gordon & Ryan Tibshirani

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 contraints),

$$x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions}$$
$$\Leftrightarrow \quad x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions}$$

(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

UNIVERSITY OF
MARYLAND

Furong Huang
4124 IRB, College Park, MD 20740
301.405.8010 / furongh@umd.edu