Extra Topic (Reading Material)
Lagrange Multipliers
& Duality of SVMs

Rishabh Iyer

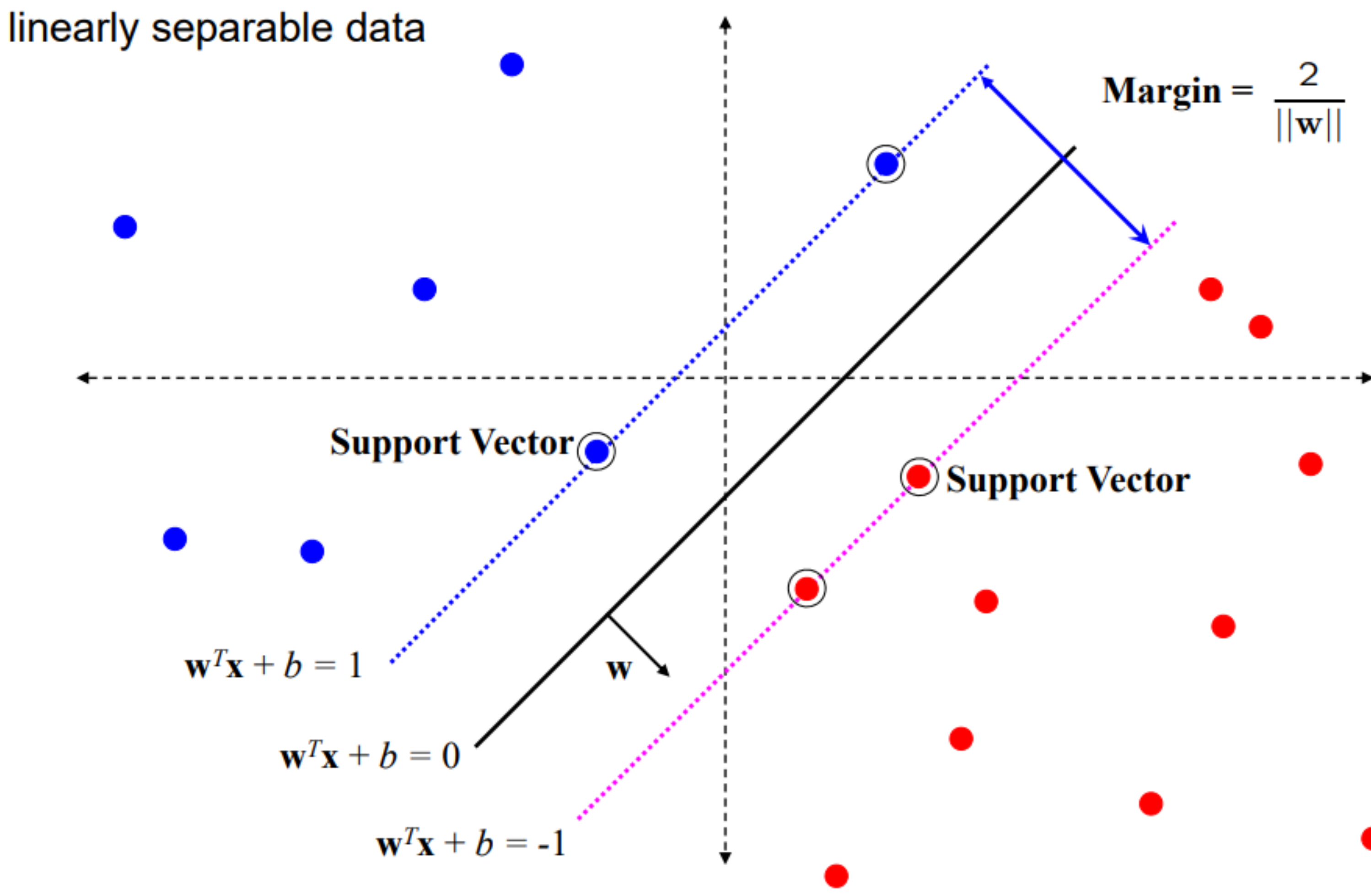
University of Texas at Dallas

Most Slides are from Nick Rouzzi! Thanks Nick!!

The Strategy So Far...

- Choose hypothesis space
- Construct loss function (ideally convex)
- Minimize loss to “learn” correct parameters

Recap: SVM



linearly separable data
Margin = $\frac{2}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}^T\mathbf{x} + b = 1$
$\mathbf{w}$
$\mathbf{w}^T\mathbf{x} + b = 0$
$\mathbf{w}^T\mathbf{x} + b = -1$

SVM Optimization Problem

- Recall: The SVM optimization problem:

$$\min_{w,b} ||w||^2$$

such that

$$y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, \text{for all } i$$

- This is a standard quadratic programming problem
  - Falls into the class of convex optimization problems
  - Can be solved with many specialized optimization tools (e.g., quadprog() in MATLAB)

Constrained Optimization

A mathematical detour, we'll come back to SVMs soon!

$$\min_{x \in \mathbb{R}^n} f_0(x) \leftarrow \frac{1}{2}\|w\|^2$$

(w, b)

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \ldots, m \leftarrow 1 - y^{(i)}(w^T x^{(i)} + b) \leq 0$$

$$h_i(x) = 0, \qquad i = 1, \ldots, p$$

Constrained Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

f_0 is not necessarily convex

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \ldots, m$$
$$h_i(x) = 0, \qquad i = 1, \ldots, p$$

General Optimization

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$\begin{aligned} f_i(x) &\le 0, \quad & i &= 1, \ldots, m \\ h_i(x) &= 0, \quad & i &= 1, \ldots, p \end{aligned}$$

Constraints do not need to be linear

Example

$$\min_{x\in\mathbb{R}^{\cancel{3}2}} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$\mathbb{R}^2$

$$x_1 + x_2 = 1$$
$$x_1 \geq 0$$
$$x_2 \geq 0$$

Example

$$\min_{x\in\mathbb{R}^3} \overbrace{x_1 \log x_1 + x_2 \log x_2}^{f_0(x)}$$

subject to:

$$h_1(x) = 0 \longrightarrow 1 - x_1 - x_2 = 0 \quad h_i(x)$$

$$\left.\begin{array}{l} f_1(x) \le 0 \longrightarrow -x_1 \le 0 \\ f_2(x) \le 0 \longrightarrow -x_2 \le 0 \end{array}\right\} f_i(x)$$

Lagrangian

Obj　　Ineq Const　　Eq Const

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

$h_i(x) = 0$

$f_i(x) \le 0$

- Incorporate constraints into a new objective function
- $\lambda \geq 0$ and ν are vectors of ***Lagrange multipliers***
- The Lagrange multipliers can be thought of as enforcing soft constraints

Example

$$\min_{x\in\mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$

$\nu_1 \to$ eq.

$$-x_1 \leq 0$$

$$-x_2 \leq 0$$

$\left.\begin{array}{l}\lambda_1\\ \lambda_2\end{array}\right\}$ ineq.

$$\begin{aligned} L(x_1, x_2, \nu_1, \lambda_1, \lambda_2) \\ &= x_1 \log x_1 + x_2 \log x_2 + \nu_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2 \end{aligned}$$

Duality

- Construct a dual function by minimizing the Lagrangian over the primal variables

min

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

- $g(\lambda, \nu) = -\infty$ whenever the Lagrangian is not bounded from below for a fixed λ and ν

Example

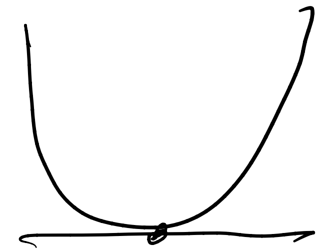
$$\min_{x \in \mathbb{R}^2} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$
$$-x_1 \leq 0$$
$$-x_2 \leq 0$$

Example

$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$



subject to:

$$1 - x_1 - x_2 = 0$$
$$-x_1 \leq 0$$
$$-x_2 \leq 0$$

$$L(x_1, x_2, v_1, \lambda_1, \lambda_2)$$
$$= x_1 \log x_1 + x_2 \log x_2 + v_1 \cdot (1 - x_1 - x_2) - \lambda_1 x_1 - \lambda_2 x_2$$

$$\frac{\partial L}{\partial x_1} = 1 + \log x_1 - v_1 - \lambda_1 = 0 \Rightarrow x_1 = \exp(v_1 + \lambda_1 - 1)$$

$$\frac{\partial L}{\partial x_2} = 1 + \log x_2 - v_1 - \lambda_2 = 0 \Rightarrow x_2 = \exp(v_1 + \lambda_2 - 1)$$

$g(\lambda, v)$ = Subst. x_1 & x_2 into $L(x_1, x_2, \lambda_1, \lambda_2, v_1)$

Example

$$\min_{x \in \mathbb{R}^3} x_1 \log x_1 + x_2 \log x_2$$

subject to:

$$1 - x_1 - x_2 = 0$$
$$-x_1 \leq 0$$
$$-x_2 \leq 0$$

The Primal Problem

$$\min_{x \in \mathbb{R}^n} f_0(x) = \max_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

subject to:

$$f_i(x) \leq 0, \qquad i = 1, \dots, m$$
$$h_i(x) = 0, \qquad i = 1, \dots, p$$

Equivalently,

$$\overset{\min}{\inf_x} \; \overset{\max}{\sup_{\lambda \geq 0, \nu}} L(x, \lambda, \nu)$$

Why are these equivalent?

The Primal Problem

$$\min_{x \in \mathbb{R}^n} f_0(x)$$

subject to:

$$f_i(x) \le 0, \qquad i = 1, \dots, m$$
$$h_i(x) = 0, \qquad i = 1, \dots, p$$

Original Problem

Equivalently,

min max

$$\inf_x \sup_{\lambda \ge 0, \nu} L(x, \lambda, \nu)$$

$$\sup_{\lambda \ge 0, \nu} \left[f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right] = \infty$$

whenever x violates the constraints

The Dual Problem

$$\overset{\text{max}}{\sup_{\lambda \geq 0, \nu}} \; g(\lambda, \nu)$$

Equivalently,
$$\overset{\text{max}}{\sup_{\lambda \geq 0, \nu}} \; \overbrace{\overset{\text{min}}{\inf_{x}} L(x, \lambda, \nu)}$$

- The dual problem is always concave, even if the primal problem is not convex
  - For each x, $L(x, \lambda, \nu)$ is a linear function in λ and ν
  - Minimum (or infimum) of linear functions is concave!

Primal vs. Dual

$$\underbrace{\sup_{\lambda\geq 0,\nu} \overbrace{\inf_x L(x,\lambda,\nu)}^{g(\lambda,\nu)}}_{\text{Dual.}} \leq \underbrace{\inf_x \overbrace{\sup_{\lambda\geq 0,\nu} L(x,\lambda,\nu)}^{f_0(x)}}_{\text{Primal.}}$$

- Why?
  - $g(\lambda,\nu) \leq L(x,\lambda,\nu)$ for all x $\qquad g(\lambda,\nu) = \min_x L(x,\lambda,\nu)$
  - $L(x',\lambda,\nu) \leq f_0(x')$ for any feasible x', $\lambda \geq 0$
    - x is feasible if it satisfies all of the constraints
  - Let x^* be the optimal solution to the primal problem and $\lambda \geq 0$

$$g(\lambda,\nu) \leq L(x^*,\lambda,\nu) \leq f_0(x^*)$$

$$L(x,\lambda,\nu) = f_0(x) + \underbrace{\sum_{i=1}^{m} \lambda_i f_i(x)}_{\leq 0} + \cancel{\sum_{i=1}^{p} \nu_i h_i(x)} = 0$$

Duality

- Under certain conditions, the two optimization problems are equivalent

$$\sup_{\lambda \geq 0, \nu} \inf_x L(x, \lambda, \nu) = \inf_x \sup_{\lambda \geq 0, \nu} L(x, \lambda, \nu)$$

  - This is called strong duality

- If the inequality is strict, then we say that there is a duality gap

  - Size of gap measured by the difference between the two sides of the inequality

Slater's Condition

For any optimization problem of the form

$$\min_{x \in \mathbb{R}^n} \underline{f_0(x)} \leftarrow \text{Convex}$$

subject to:

Convex
$\downarrow$

$$\underline{f_i(x)} \le 0, \qquad i = 1, \dots, m$$
$$Ax = b \leftarrow \text{Linear Equality.}$$

where $f_0, \dots, f_m$ are convex functions, strong duality holds if there exists an x such that

$$f_i(x) < 0, \qquad i = 1, \dots, m$$
$$Ax = b$$

Dual SVM

$$\min_{w} \frac{1}{2} ||w||^2 \leftarrow \text{Objective}$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1, \text{for all } i \leftarrow \text{InEq Constraints}$$

- Note that Slater's condition holds as long as the data is linearly separable

Dual SVM

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_i \lambda_i (1 - y_i (w^T x^{(i)} + b))$$

Convex in w, so take derivatives to form the dual

$$\frac{\partial L}{\partial w_k} = w_k + \sum_i -\lambda_i y_i x_k^{(i)} = 0$$

$$\frac{\partial L}{\partial b} = \sum_i -\lambda_i y_i = 0$$

Dual SVM

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_i \lambda_i (1 - y_i (w^T x^{(i)} + b))$$

Convex in w, so take derivatives to form the dual

$$w = \sum_i \lambda_i y_i x^{(i)}$$

$$\sum_i \lambda_i y_i = 0$$

Dual SVM

$$\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j \underbrace{x^{(i)^T} x^{(j)}}_{K(x^{(i)},\, x^{(j)})} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- By strong duality, solving this problem is equivalent to solving the primal problem
  - Given the optimal λ, we can easily construct w (b can be found by complementary slackness...)

x_{test}, $\quad y_{pred} = \text{Sign}(w^T x_{test} + b)$

Complementary Slackness

- Suppose that there is zero duality gap
- Let x^* be an optimum of the primal and (λ^*, ν^*) be an optimum of the dual

Primal = Dual.

$$
\begin{aligned}
f_0(x^*) &= g(\lambda^*, \nu^*) \\
&= \inf_x \left[f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \right] \\
&\le f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \underbrace{\sum_{i=1}^{p} \nu_i^* h_i(x^*)}_{=0} \\
&= f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) \\
&\le f_0(x^*)
\end{aligned}
$$

$\lambda_i^* \geq 0, \; f_i(x^*) \leq 0$

Complementary Slackness

- This means that

$$\sum_{i=1}^{m} \overbrace{\lambda_i^* f_i(x^*)}^{\leq 0} = 0$$

- As $\lambda \geq 0$ and $f_i(x_i^*) \leq 0$, this can only happen if $\lambda_i^* f_i(x^*) = 0$ for all i
- Put another way,
  - If $f_i(x^*) < 0$ (i.e., the constraint is not tight), then $\lambda_i^* = 0$
  - If $\lambda_i^* > 0$, then $f_i(x^*) = 0$ ← i is a support vector
  - ONLY applies when there is no duality gap

Dual SVM (Obtaining b)

$$\max_{\lambda \geq 0} -\frac{1}{2}\sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)^T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

- By complementary slackness, $\lambda_i^* > 0$ means that $x^{(i)}$ is a support vector (can then solve for b using w)
- In particular,

$$b = y_i - w.x_i$$

for any i where $\lambda_i > 0$ [support vectors]

Dual SVM

$$\max_{\lambda \geq 0} -\frac{1}{2}\sum_{i=1}^{M}\sum_{j=1}^{M}\lambda_i\lambda_j y_i y_j x^{(i)^T} x^{(j)} + \sum_i \lambda_i$$

such that

$$\sum_i \lambda_i y_i = 0$$

O(m²)

- Takes $\cancel{O(n^2)}$ time just to evaluate the objective function
  - Active area of research to try to speed this up