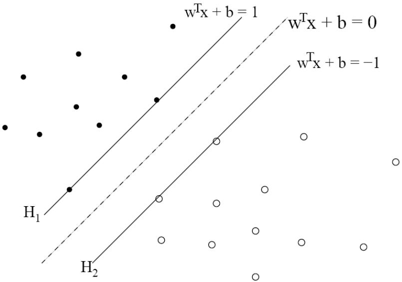
Support Vector Machines, Constrained Optimization

CSci 5525: Machine Learning

Instructor: Paul Schrater

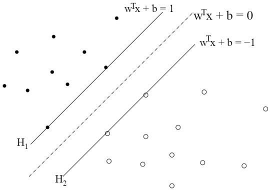
Linear SVM



- Distance of central hyperplane from origin $= \frac{|b|}{\|\mathbf{w}\|}$
- Distance of parallel hyperplanes are $\frac{|b-1|}{\|\mathbf{w}\|}$ and $\frac{|b+1|}{\|\mathbf{w}\|}$
- Distance between hyperplanes $= \frac{2}{\|\mathbf{w}\|}$
- Main Idea:

Choose **w** *to maximize class separation*

Linear SVM: Separable Case



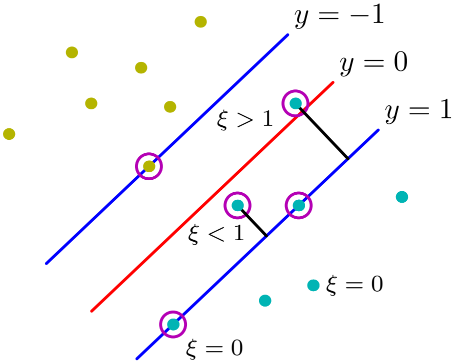
- The Main Idea can be formulated as

$$\min \; \frac{1}{2}\|\mathbf{w}\|^2 \quad \text{such that} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \forall i$$

- The choice of "1" as a constant is *wlog*
  - Any other choice can be reduced to the above form
- The main problem is a "quadratic program"

($\forall$: for all, $\exists$: there exists, wlog: without loss of generality)

Linear SVM: Non-Separable Case



Linear SVM: Non-Separable Case

- Separability assumption: $\exists \mathbf{w}, \forall i \; y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$
- If not true, the problem formulation is infeasible
- For the general case, we will introduce *slack variables*

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \forall i$$

- Note that $\sum_i \xi_i$ is an upper bound on the training error
- In general, the problem can be formulated as

$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_i \xi_i \quad \text{such that} \quad y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \xi_i \geq 0$$

- Perspective: constrained optimization

SVM loss, Revisited

- The prediction $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$
- The (primal) non-separable case

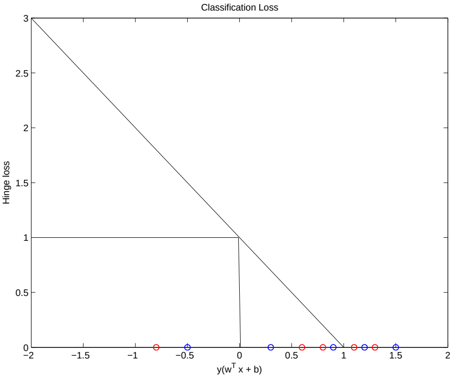
$$\min_{\mathbf{w},\{\xi_i\}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \right\} \text{ such that } y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \xi_i \geq 0$$

- Alternative viewpoint as a regularized hinge loss

$$\min_{\mathbf{w}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i f(\mathbf{x}_i)\} + \lambda\|\mathbf{w}\|^2 \right\}$$

- Regularized loss minimization with two terms
  - First term: Margin loss on the training set
  - Second term: Regularization
- Perspective: unconstrained "non-smooth" optimization

The Hinge Loss

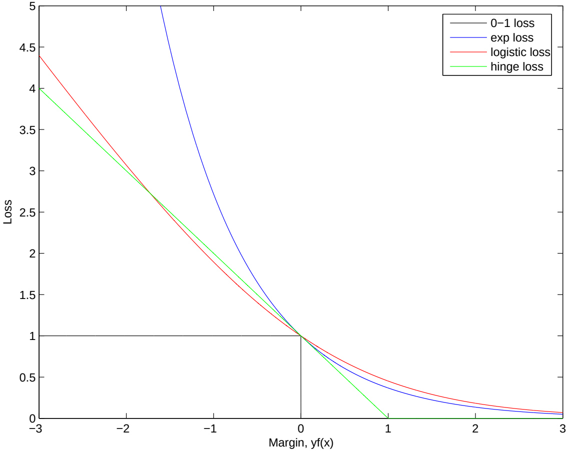


- The goal is to minimize

$$\min_{\mathbf{w}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y_i f(\mathbf{x}_i)\} + \lambda \|\mathbf{w}\|^2 \right\}$$

- The hinge-loss: $h(\mathbf{x}_i, y_i, f) = \max\{0, 1 - y_i f(\mathbf{x}_i)\}$

Upper Bounds on Training Error



- SVM maximizes minimum margin
- SVM is a L_2 regularized fit using hinge loss
- Logistic and Hinge losses are very similar

Constrained Optimization

- The equality & inequality constrained optimization problem

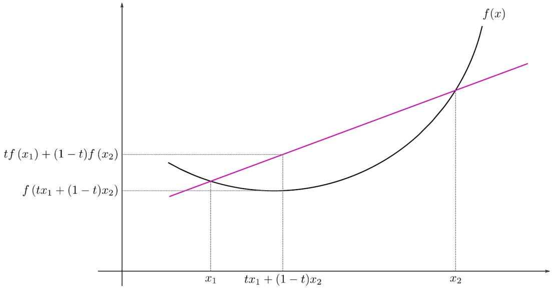
$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) = 0 \qquad i = 1, \ldots, m \\ & g_j(\mathbf{x}) \leq 0 \qquad j = 1, \ldots, n \end{aligned}$$

  - Domain $\mathcal{D} = \mathrm{dom}(f) \cap \bigcap_{i=1}^{m} \mathrm{dom}(h_i) \cap \bigcap_{j=1}^{n} \mathrm{dom}(g_j)$
  - Called the "primal" or primal problem
- Feasible set $\mathcal{F} \subseteq \mathcal{D}$: $\mathbf{x} \in \mathcal{F}$ satisfies $h_i(\mathbf{x}) = 0, g_j(\mathbf{x}) \leq 0$
- The Lagrangian

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= f(\mathbf{x}) + \boldsymbol{\lambda}^T h(\mathbf{x}) + \boldsymbol{\nu}^T g(\mathbf{x}) \\ &= f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j g_j(\mathbf{x}) \end{aligned}$$

  - Domain $\mathrm{dom}(L) = \mathcal{D} \times \mathbb{R}^m \times \mathbb{R}^n$
- $\{\lambda_i\}_{i=1}^{m}, \{\nu_j\}_{j=1}^{n}$ are the Lagrange multipliers

Background: Convex Functions



- f is convex if $\forall x_1, x_2 \in \mathrm{dom}(f), \forall t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

- If f is differentiable, then

$$f(x_1) \geq f(x_2) + (x_1 - x_2)^T \nabla f(x_2)$$

- Examples: $f(x) = \frac{1}{2}||x||_2^2, f(x) = -\log x, f(x) = \|x\|_1$
- f is concave if $-f$ is convex

Lagrange Dual

- The Lagrange dual function

$$\begin{aligned} L^*(\boldsymbol{\lambda}, \boldsymbol{\nu}) &= \inf_{\mathbf{x} \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &= \inf_{\mathbf{x} \in \mathcal{D}} \left(f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j g_j(\mathbf{x}) \right) \end{aligned}$$

- Let p^* be the constrained optimum of $f(\mathbf{x})$
- The Lagrange dual L^* is
  - A concave function, even when original problem is not convex
  - A lower bound to the optimum p^*:

$$L^*(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^* , \qquad \forall \boldsymbol{\nu} \geq 0$$

- How close is the maximum of $L^*(\boldsymbol{\lambda}, \boldsymbol{\nu})$ to p^*?

Lagrange Dual: Concave, Lower Bound to Primal

- $L^*(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is a concave function
  - Consider a function
$$\eta^*(\mathbf{v}) = \sup_{\mathbf{x} \in \mathcal{D}} \left(\langle \mathbf{v}, \phi(\mathbf{x}) \rangle - f(\mathbf{x}) \right)$$
  - Vectors $\mathbf{v} = [\boldsymbol{\lambda} \quad \boldsymbol{\nu}]$ and $\phi(\mathbf{x}) = [-\mathbf{h}(\mathbf{x}) \quad -\mathbf{g}(\mathbf{x})]$
  - Then, $\eta^*(\mathbf{v})$ is always convex: sup of affine functions
  - Hence, $L^*(\boldsymbol{\lambda}, \boldsymbol{\nu}) = -\eta^*(\mathbf{v})$ is concave
- Lower bound: for $\boldsymbol{\nu} \geq 0$, $L^*(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$
  - When $\mathbf{x}$ is feasible, i.e., $\mathbf{x} \in \mathcal{F}$, $h_i(\mathbf{x}) = 0, g_j(\mathbf{x}) \leq 0$
  - When $\mathbf{x} \in \mathcal{F}$, since $\nu_j \geq 0$
$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j g_j(\mathbf{x}) \quad \leq \quad f(\mathbf{x})$$
  - When $\mathbf{x} \in \mathcal{D}$, since $\mathcal{F} \subseteq \mathcal{D}$
$$\inf_{x \in \mathcal{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \;\leq\; \inf_{x \in \mathcal{F}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \;\leq\; \inf_{x \in \mathcal{F}} f(x) \;=\; p^*$$
  - As a result: $L^*(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$

Example: Quadratic Problems with equality constraints

$$\text{minimize} \quad \mathbf{x}^T\mathbf{x}$$
$$\text{subject to} \quad A\mathbf{x} = b$$

- Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x}^T\mathbf{x} + \boldsymbol{\lambda}^T(A\mathbf{x} - b)$
- Recall that $L^*(\boldsymbol{\lambda}) = \inf_{\mathbf{x}} \ L(\mathbf{x}, \boldsymbol{\lambda})$
- Setting gradient to 0, $\mathbf{x} = -\frac{1}{2}A^T\boldsymbol{\lambda}$
- Hence, the dual

$$L^*(\boldsymbol{\lambda}) = L\left(-\frac{1}{2}A^T\boldsymbol{\lambda}, \boldsymbol{\lambda}\right) = -\frac{1}{4}\boldsymbol{\lambda}^T AA^T\boldsymbol{\lambda} - \boldsymbol{\lambda}^T b$$

- $L^*(\boldsymbol{\lambda})$ is a lower bounding concave function

Example: General Quadratic Programs

$$\begin{aligned} \text{minimize} \quad & \mathbf{x}^T\mathbf{x} \\ \text{subject to} \quad & A\mathbf{x} \leq \mathbf{b} \end{aligned}$$

- Lagrange dual

$$L^*(\boldsymbol{\nu}) = \inf_{\mathbf{x}} \left(\mathbf{x}^T\mathbf{x} + \boldsymbol{\nu}^T(A\mathbf{x} - \mathbf{b}) \right) = -\frac{1}{4}\boldsymbol{\nu}^T A A^T \boldsymbol{\nu} - b^T\boldsymbol{\nu}$$

- Dual problem

$$\begin{aligned} \text{maximize} \quad & -\frac{1}{4}\boldsymbol{\nu}^T A A^T \boldsymbol{\nu} - b^T\boldsymbol{\nu} \\ \text{subject to} \quad & \boldsymbol{\nu} \geq 0 \end{aligned}$$

The Lagrange Dual Problem

$$\text{maximize } L^*(\boldsymbol{\lambda}, \boldsymbol{\nu})$$
$$\text{subject to } \boldsymbol{\nu} \geq 0$$

- Best lower bound to p^*, the optimal of the primal
- Concave optimization problem with maximum d^*
- Constraints are $\boldsymbol{\nu} \geq 0$ and $(\boldsymbol{\lambda}, \boldsymbol{\nu}) \in \text{dom}(L^*)$
- For example, in quadratic programming

$$\text{minimize } \mathbf{x}^T\mathbf{x} \qquad \text{maximize } -\frac{1}{4}\boldsymbol{\lambda}^T A A^T \boldsymbol{\lambda} - b^T \boldsymbol{\lambda}$$
$$\text{subject to } A\mathbf{x} \leq \mathbf{b} \qquad \text{subject to } \nu \geq 0$$

Weak and Strong Duality

- Weak Duality: $d^* \leq p^*$
  - Always holds
  - Non-trivial lower bounds for hard problems
  - Used in approximation algorithms
- Strong Duality: $d^* = p^*$
  - Does not hold in general
  - If it holds, it is sufficient to solve the dual
  - How to check it if holds?
- Constraint Qualification
  - Normally true on convex problems
  - True if the convex problem is strictly feasible, e.g.,

$$\exists x \in \mathrm{relint}(\mathcal{D}) \quad \text{s.t.} \quad Ax = b, \quad g_j(x) < 0, \quad \text{for (non-affine) } g_j$$

  - Slater's Condition for strong duality
  - Example: Quadratic programs

Complementary Slackness

- If strong duality holds, $\mathbf{x}^*$ for primal, $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ for dual

$$
\begin{aligned}
f(\mathbf{x}^*) = L^*(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = & \inf_{\mathbf{x}} \left(f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* h_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j^* g_j(\mathbf{x}) \right) \\
\leq & \; f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* h_i(\mathbf{x}^*) + \sum_{j=1}^{n} \nu_j^* g_j(\mathbf{x}^*) \\
\leq & \; f(\mathbf{x}^*)
\end{aligned}
$$

- The two inequalities *must* hold with equality
  - $\mathbf{x}^*$ minimizes the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$
  - $\nu_j^* g_j(\mathbf{x}^*) = 0$ for all $j = 1, \ldots, n$ so that

$$
\nu_j^* > 0 \Rightarrow g_j(\mathbf{x}^*) = 0, \qquad \text{and} \qquad g_j(\mathbf{x}^*) < 0 \Rightarrow \nu_j^* = 0
$$

Karush-Kuhn-Tucker (KKT) Conditions

Necessary conditions satisfied by any primal and dual optimal pairs $\tilde{\mathbf{x}}$ and $(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$

- Primal Feasibility:

$$h_i(\tilde{\mathbf{x}}) = 0, i = 1, \ldots, n, \quad g_j(\tilde{\mathbf{x}}) \leq 0, j = 1, \ldots, m$$

- Dual Feasibility:

$$\tilde{\nu}_j \geq 0, j = 1, \ldots, m$$

- Complementary Slackness:

$$\tilde{\nu}_j g_j(\tilde{\mathbf{x}}) = 0, j = 1, \ldots, m$$

- Gradient condition:

$$\nabla f(\tilde{\mathbf{x}}) + \sum_{i=1}^{n} \tilde{\lambda}_i \nabla h_i(\tilde{\mathbf{x}}) + \sum_{j=1}^{m} \tilde{\nu}_j \nabla g_j(\tilde{\mathbf{x}}) = 0$$

- The conditions are <u>sufficient</u> for a convex problem

SVM Lagrange Dual: Separable Case

- The Lagrangian

$$L([\mathbf{w} \quad b], \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i y_i(\mathbf{w}^T \mathbf{x}_i + b) + \sum_i \alpha_i$$

- Setting gradient w.r.t. $[\mathbf{w} \quad b]$ to 0, we get

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \qquad \sum_i \alpha_i y_i = 0$$

- Substituting these back, we get the Lagrange dual ($\boldsymbol{\alpha} \geq 0$)

$$L^*(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

- Constraints for the dual optimization

$$\begin{aligned} \alpha_i &\geq 0, \forall i \\ \sum_i \alpha_i y_i &= 0 \,. \end{aligned}$$

Learning and Prediction: Separable Case

- The Lagrange dual ($\boldsymbol{\alpha} \geq 0$)

$$L^*(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

- Recall complementary slackness $\alpha_i g_i(\mathbf{x}) = 0$ for $g_i(\mathbf{x}) \leq 0$

$\alpha_i > 0 \Rightarrow y_i(\mathbf{w}^T \mathbf{x}_i + b) = 1$ $\quad$ $\mathbf{x}_i$ is a support vector

Otherwise $y_i(\mathbf{w}^T \mathbf{x}_i + b) > 1$ $\quad$ $\mathbf{x}_i$ is not a support vector

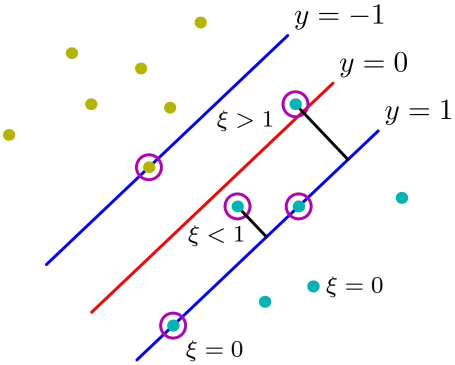
- If $\alpha_i > 0$, the constraint holds with equality
- The resulting $\mathbf{w}$ is given by

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i = \sum_{i:\alpha_i > 0} \alpha_i y_i \mathbf{x}_i$$

- b can be obtained using complimentary slackness
- For any future point $\mathbf{x}$, prediction is

$$\text{sign}(\mathbf{w}^T \mathbf{x} + b)$$

Linear SVM: Non-Separable Case



$$\min \frac{1}{2}\|\mathbf{w}\|^2 + C\sum \xi_i \quad \text{such that} \quad y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \xi_i \geq 0$$

Lagrange Dual: Non-Separable Case

- The Lagrangian

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i - \sum_i \alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i) - \sum_i \mu_i \xi_i$$

- Setting gradient w.r.t. $[\mathbf{w} \quad b \quad \xi]$ to 0, we get

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i \qquad \sum_i \alpha_i y_i = 0 \qquad 0 \leq \alpha_i \leq C$$

- Substituting, we get the Lagrange dual $(0 \leq \boldsymbol{\alpha} \leq C)$

$$L^*(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

- One additional set of box constraints on α_i

The KKT Conditions

- Primal feasibility:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i \geq 0$$
$$\xi_i \geq 0$$

- Dual feasibility:

$$\alpha_i \geq 0$$
$$\mu_i \geq 0$$

- Complementary slackness:

$$\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i) = 0$$
$$\mu_i\xi_i = 0$$

- Gradient condition:

$$\mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_i = 0$$
$$\sum_i \alpha_i y_i = 0$$
$$\alpha_i + \mu_i - C = 0$$

Prediction

- The set of support vectors have $\alpha_i > 0$
- The trained classifier has weight

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i = \sum_{i:\alpha_i > 0} \alpha_i y_i \mathbf{x}_i$$

- KKT conditions can be used to compute b
- The prediction on a new point $\mathbf{x}$

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = \sum_{i:\alpha_i > 0} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$$

- The prediction is in terms of dot-products $\mathbf{x}_i^T \mathbf{x}$
- The dual was also in terms of dot-products $\mathbf{x}_i^T \mathbf{x}_j$