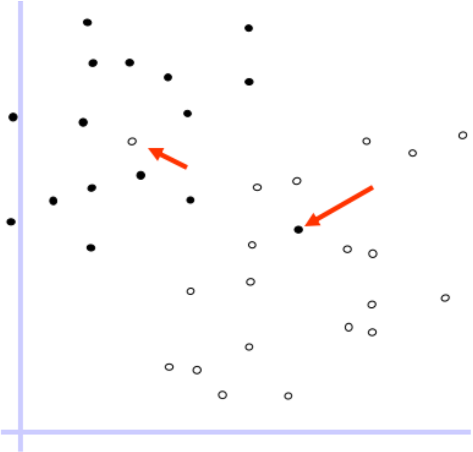
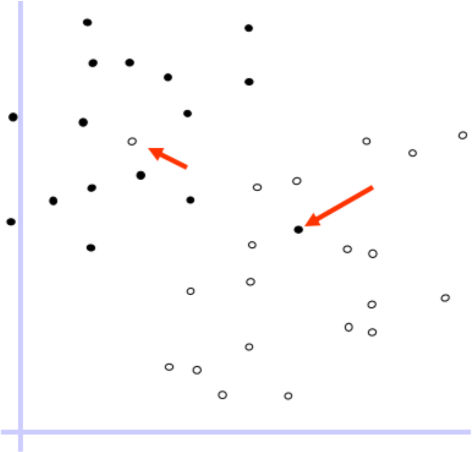
Slack SVMs

Jordan Boyd-Graber
University of Colorado Boulder
LECTURE 9

Can SVMs Work Here?

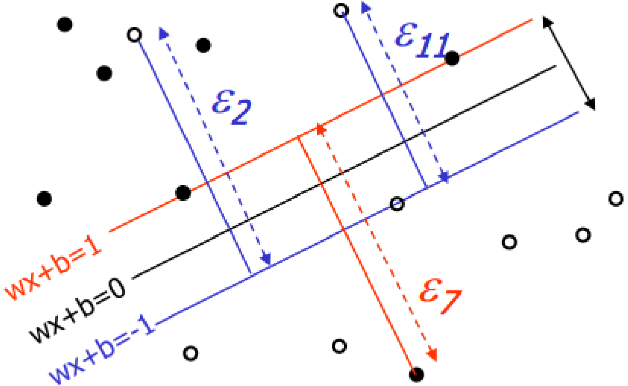


Can SVMs Work Here?



$$y_i(w \cdot x_i + b) \geq 1 \quad (1)$$

Trick: Allow for a few bad apples



New objective function

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1} {\xi_i}^p \tag{2}$$

subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

New objective function

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1} {\xi_i}^p \quad (2)$$

subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

- Standard margin

New objective function

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1} {\xi_i}^p \tag{2}$$

subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

- Standard margin
- How wrong a point is (slack variables)

New objective function

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C \sum_{i=1} {\xi_i}^p \tag{2}$$

subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i \land \xi_i \geq 0, i \in [1, m]$

- Standard margin
- How wrong a point is (slack variables)
- Tradeoff between margin and slack variables

New objective function

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sum_{i=1} {\xi_i}^{p} \tag{2}$$

subject to $y_i(w \cdot x_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

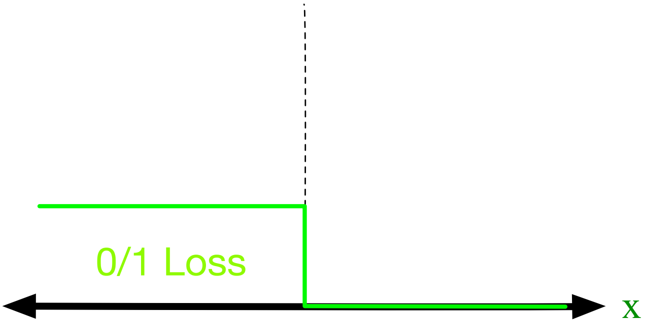
- Standard margin
- How wrong a point is (slack variables)
- Tradeoff between margin and slack variables
- How bad wrongness scales

Aside: Loss Functions

- Losses measure how bad a mistake is
- Important for slack as well

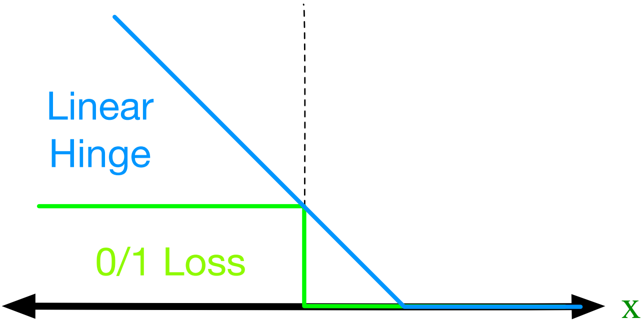
Aside: Loss Functions

- Losses measure how bad a mistake is
- Important for slack as well



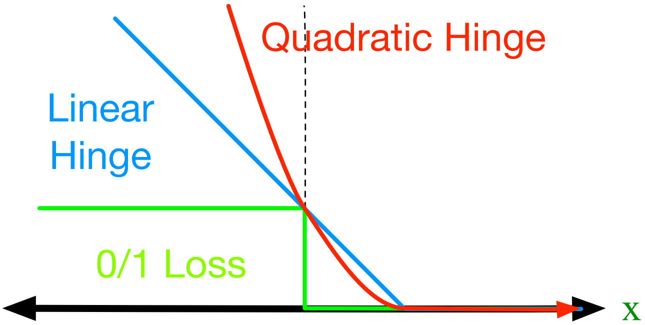
Aside: Loss Functions

- Losses measure how bad a mistake is
- Important for slack as well



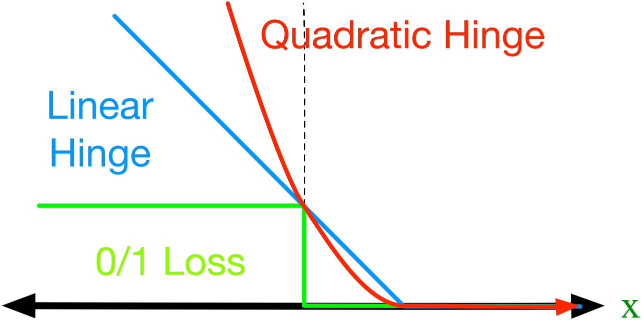
Aside: Loss Functions

- Losses measure how bad a mistake is
- Important for slack as well



Aside: Loss Functions

- Losses measure how bad a mistake is
- Important for slack as well



We'll focus on linear hinge loss

Optimizing Constrained Functions

Theorem: Lagrange Multiplier Method

Given functions $f(x_1, \dots x_n)$ and $g(x_1, \dots x_n)$, the critical points of f restricted to the set $g = 0$ are solutions to equations:

$$\begin{aligned} \frac{\partial f}{\partial x_i}(x_1, \dots x_n) =& \lambda \frac{\partial g}{\partial x_i}(x_1, \dots x_n) \quad \forall i \\ g(x_1, \dots x_n) =& 0 \end{aligned}$$

This is $n+1$ equations in the $n+1$ variables $x_1, \dots x_n, \lambda$.

Lagrange Example

Maximize $f(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

Lagrange Example

Maximize $f(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial f}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial f}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

Lagrange Example

Maximize $f(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial f}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \qquad \frac{\partial g}{\partial x} = 20$$

$$\frac{\partial f}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \qquad \frac{\partial g}{\partial y} = 10$$

- Create new systems of equations

Lagrange Example

Maximize $f(x, y) = \sqrt{xy}$ subject to the constraint $20x + 10y = 200$.

- Compute derivatives

$$\frac{\partial f}{\partial x} = \frac{1}{2}\sqrt{\frac{y}{x}} \quad \frac{\partial g}{\partial x} = 20$$
$$\frac{\partial f}{\partial y} = \frac{1}{2}\sqrt{\frac{x}{y}} \quad \frac{\partial g}{\partial y} = 10$$

- Create new systems of equations

$$\frac{1}{2}\sqrt{\frac{y}{x}} = 20\lambda$$
$$\frac{1}{2}\sqrt{\frac{x}{y}} = 10\lambda$$
$$20x + 10y = 200$$

Lagrange Example

- Dividing the first equation by the second gives us

$$\frac{y}{x} = 2 \tag{3}$$

- which means $y = 2x$, plugging this into the constraint equation gives:

$$\begin{aligned} 20x + 20(2x) &= 200 \\ x = 5 &\Rightarrow y = 10 \end{aligned}$$

New Lagrangian

$$\mathscr{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \tag{4}$$

$$- \sum_{i=1}^{m}\alpha_i \left[y_i(w \cdot x_i + b) - 1 + \xi_i\right] \tag{5}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{6}$$

New Lagrangian

$$\mathscr{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \tag{4}$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(w \cdot x_i + b) - 1 + \xi_i\right] \tag{5}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{6}$$

Taking the gradients $(\nabla_w\mathscr{L}, \nabla_b\mathscr{L}, \nabla_{\xi_i})$ and solving for zero gives us

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \quad (7) \qquad \vec{w} = \sum_{i=1}^{m}\alpha_i y_i x_i \quad (8) \qquad \alpha_i + \beta_i = C \quad (9)$$

New Lagrangian

$$\mathscr{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \tag{4}$$

$$-\sum_{i=1}^{m}\alpha_i\left[y_i(w \cdot x_i + b) - 1 + \xi_i\right] \tag{5}$$

$$-\sum_{i=1}^{m}\beta_i\xi_i \tag{6}$$

Taking the gradients ($\nabla_w\mathscr{L}, \nabla_b\mathscr{L}, \nabla_{\xi_i}$) and solving for zero gives us

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{7}$$

$$\vec{w} = \sum_{i=1}^{m}\alpha_i y_i x_i \tag{8}$$

$$\alpha_i + \beta_i = C \tag{9}$$

New Lagrangian

$$
\begin{aligned}
\mathscr{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i && (4)\\
& - \sum_{i=1}^{m}\alpha_i\left[y_i(w \cdot x_i + b) - 1 + \xi_i\right] && (5)\\
& - \sum_{i=1}^{m}\beta_i\xi_i && (6)
\end{aligned}
$$

Taking the gradients $(\nabla_w \mathscr{L}, \nabla_b \mathscr{L}, \nabla_{\xi_i})$ and solving for zero gives us

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \quad (7) \qquad \vec{w} = \sum_{i=1}^{m}\alpha_i y_i x_i \quad (8) \qquad \alpha_i + \beta_i = C \quad (9)$$

New Lagrangian

$$\mathscr{L}(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \quad (4)$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(w \cdot x_i + b) - 1 + \xi_i\right] \quad (5)$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \quad (6)$$

Taking the gradients ($\nabla_w \mathscr{L}, \nabla_b \mathscr{L}, \nabla_{\xi_i}$) and solving for zero gives us

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \quad (7) \qquad \vec{w} = \sum_{i=1}^{m}\alpha_i y_i x_i \quad (8) \qquad \alpha_i + \beta_i = C \quad (9)$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = \frac{1}{2} \left\| \vec{w}_i \right\| - \sum_{i}^{m} \alpha_i y_i \vec{w} \cdot \vec{x}_i - \sum_{i}^{m} \alpha_i y_i b - \sum_{i=1}^{m} \beta_i \xi_i \tag{10}$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \right\| - \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) - \sum_{i}^{m} \alpha_i y_i b - \sum_{i=1}^{m} \beta_i \xi_i \tag{10}$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \right\| - \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) - \sum_{i}^{m} \alpha_i y_i b - \sum_{i=1}^{m} \beta_i \xi_i \tag{10}$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \right\| - \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) - \sum_{i}^{m} \alpha_i y_i b + \sum_{i}^{m} \alpha_i \tag{10}$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \right\| - \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) - \sum_{i}^{m} \alpha_i y_i b + \sum_{i}^{m} \alpha_i \tag{10}$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \right\| - \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) - \sum_{i}^{m} \alpha_i y_i b + \sum_{i}^{m} \alpha_i \tag{10}$$

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathcal{L} = \frac{1}{2} \left\| \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \right\| - \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) + \sum_{i}^{m} \alpha_i \tag{10}$$

First two terms are the same!

Simplifying dual objective

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \vec{w} = \sum_{i=1}^{m} \alpha_i y_i x_i \qquad \alpha_i + \beta_i = C$$

$$\mathscr{L} = -\frac{1}{2} \sum_{i}^{m} \sum_{j}^{m} \alpha_i \alpha_j y_i y_j (\vec{x}_j \cdot \vec{x}_i) + \sum_{i}^{m} \alpha_i \tag{10}$$

Just like separable case, except that we add the constraint that $\alpha_i \leq C$!

Wrapup

- Adding slack variables don't break the SVM problem
- Very popular algorithm
  - SVMLight (many options)
  - Libsvm / Liblinear (very fast)
  - Weka (friendly)
  - pyml (Python focused, from Colorado)
- Next time: simple algorithm for finding SVMs