Machine Learning: Chenhao Tan

University of Colorado Boulder

LECTURE 10

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Roadmap

- Last time: linear SVM formulation when data is linearly separable
- This time:
  - Introduce duality
  - Make linear SVM work when data is not linearly separable
  - Introduce an efficient algorithm for finding weights
- Next time: Kernel trick

Overview

Duality

Slack variables

Sequential Mimimal Optimization

Recap

Outline

Duality

Slack variables

Sequential Mimimal Optimization

Recap

Binary classification

Given: $S_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m}$ training examples, $\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \{-1, 1\}$
Goal: Find hypothesis function $h : X \to Y$
Linear SVM: learn a linear decision rule of the form $\boldsymbol{w} \cdot \boldsymbol{x} + b$

Optimizing the objective function

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \tag{1}$$

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$

Optimizing Constrained Functions

The Method of Lagrange Multipliers

Constrained problem (Primal problem)

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$
$$\text{s.t. } g_i(\boldsymbol{x}) \geq 0, i \in [1, n]$$

Lagrange Multiplier

$$\mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = f(\boldsymbol{x}) - \sum_{i=1}^{n} \alpha_i g_i(\boldsymbol{x}),$$
$$\alpha_i \geq 0, i \in [1, n]$$

Lagrange Multiplier

p^*: the optimal value in the primal problem
We claim that

$$p^* = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \sum_{i=1}^{n} \alpha_i g_i(\boldsymbol{x})$$

Lagrange Multiplier

p^*: the optimal value in the primal problem
We claim that

$$p^* = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha}} f(\boldsymbol{x}) - \sum_{i=1}^{n} \alpha_i g_i(\boldsymbol{x})$$

This is because

$$\max -\boldsymbol{\alpha}\boldsymbol{y} = \begin{cases} 0 & \boldsymbol{y} \geq 0 \\ +\infty & \text{otherwise} \end{cases}$$

Lagrange Multiplier

What happens if we reverse min and max:

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha}) \geq or \leq \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha})$$

Lagrange Multiplier

What happens if we reverse min and max:

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{x}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha}) \leq \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha}} \mathscr{L}(\boldsymbol{x}, \boldsymbol{\alpha})$$

The left leads to the dual problem.

Primal vs. Dual

Derive the function for dual problem.

Primal problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2$$

$$\text{s.t. } y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$$

Primal vs. Dual

Replace $\boldsymbol{w}, b$ with stationarity conditions.

Primal problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2$$

$$\text{s.t. } y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$$

Primal vs. Dual

Primal problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2$$

$$\text{s.t. } y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$$

Dual problem

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i)$$

$$\text{s.t. } \alpha_i \geq 0, i \in [1, m]$$

$$\sum_i \alpha_i y_i = 0$$

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \alpha_i \geq 0 \tag{2}$$

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \alpha_i \geq 0 \tag{2}$$

Stationarity

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3}$$

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, \alpha_i \geq 0 \tag{2}$$

Stationarity

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{3}$$

Complementary slackness

$$\alpha_i = 0 \vee y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) = 1 \tag{4}$$

Outline

Duality

Slack variables

Sequential Mimimal Optimization

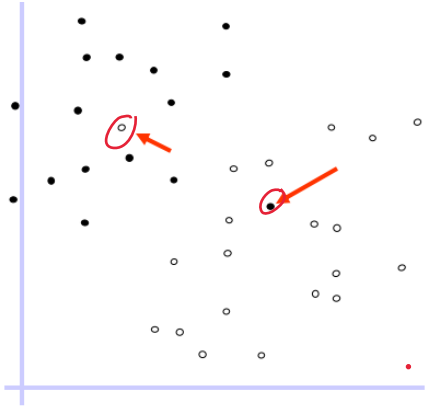
Recap

Old objective function

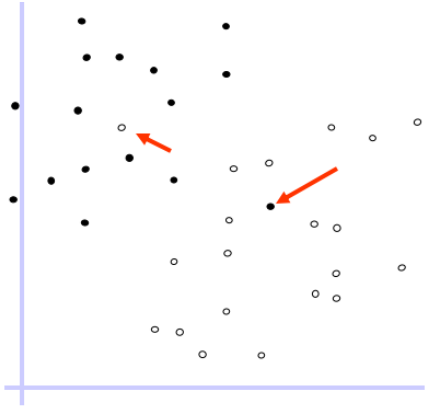
$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \tag{5}$$

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$

Can SVMs Work Here?

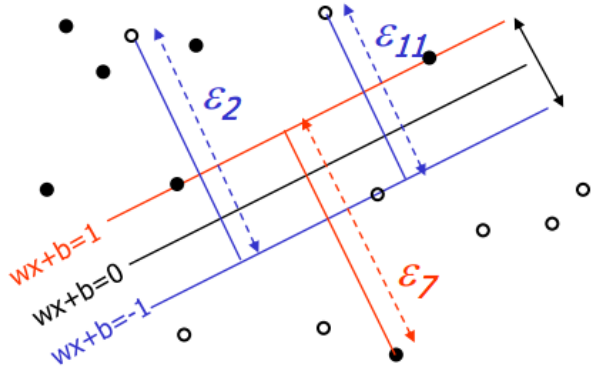


Can SVMs Work Here?



$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \qquad \xi_i \geq 0 \tag{6}$$

Trick: Allow for a few bad apples



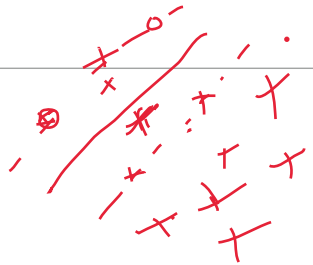
Old objective function

$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2 \tag{7}$$

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$

$\xi_i \geq 0$

Relaxing the constraint



$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i$$

- $\xi_i = 0$ means at least one margin on correct side of decision boundary
- $\xi_i = 1/2$ means at least one-half margin on correct side of decision boundary
- $\xi_i = 2$ means at least one margin on wrong side of decision boundary

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1} \xi_i^p \tag{8}$$

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1} \xi_i^p \tag{8}$$

subject to $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

- Standard margin

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1} \xi_i^p \tag{8}$$

subject to $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

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

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1} \xi_i^p \tag{8}$$

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

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

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1} \xi_i^p \quad (8)$$

subject to $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

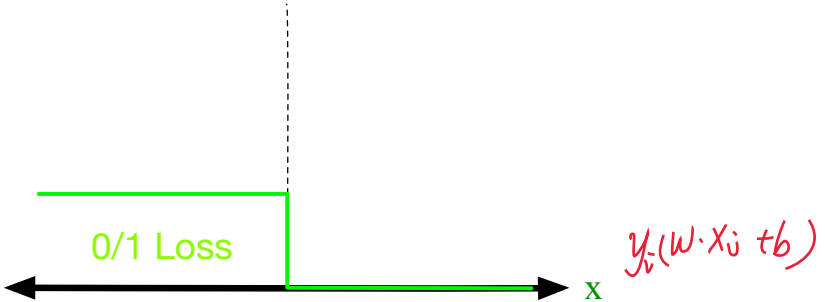
- Standard margin
- How wrong a point is (slack variables)
- Tradeoff between margin and slack variables
- How bad wrongness scales

Aside: Loss Functions

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

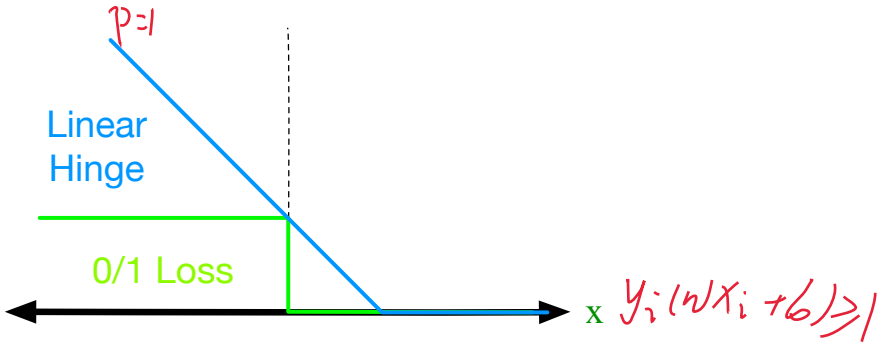
Aside: Loss Functions

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



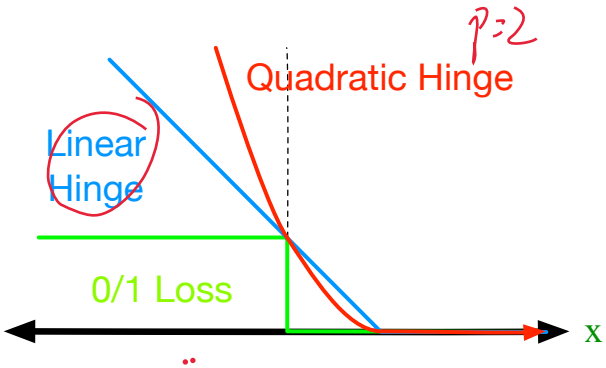
Aside: Loss Functions

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



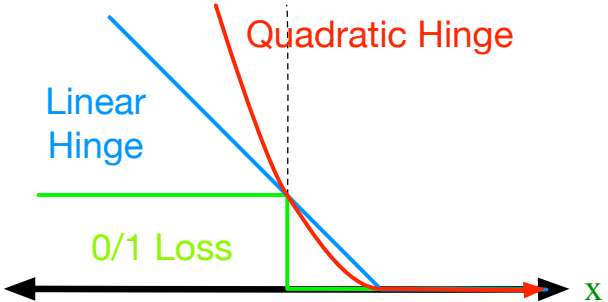
Aside: Loss Functions

- 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, set $p = 1$

What is the role of C?

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1} \xi_i \tag{9}$$

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i \wedge \xi_i \geq 0, i \in [1, m]$

A. $C \uparrow \Rightarrow$ low bias, low variance

B. $C \uparrow \Rightarrow$ low bias, high variance

C. $C \uparrow \Rightarrow$ high bias, low variance

D. $C \uparrow \Rightarrow$ high bias, high variance

New Lagrangian

$$\mathscr{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{10}$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(\boldsymbol{w}\cdot x_i + b) - 1 + \xi_i\right] \tag{11}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{12}$$

New Lagrangian

$$\mathscr{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{10}$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(\boldsymbol{w}\cdot x_i + b) - 1 + \xi_i\right] \tag{11}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{12}$$

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

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \tag{13}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{14}$$

$$\alpha_i + \beta_i = C \tag{15}$$

New Lagrangian

$$\mathscr{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{10}$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(\boldsymbol{w}\cdot x_i + b) - 1 + \xi_i\right] \tag{11}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{12}$$

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

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \tag{13}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{14}$$

$$\alpha_i + \beta_i = C \tag{15}$$

New Lagrangian

$$\mathscr{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{10}$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(\boldsymbol{w}\cdot x_i + b) - 1 + \xi_i\right] \tag{11}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{12}$$

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

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \tag{13}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{14}$$

$$\alpha_i + \beta_i = C \tag{15}$$

New Lagrangian

$$\mathscr{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \tag{10}$$

$$- \sum_{i=1}^{m}\alpha_i\left[y_i(\boldsymbol{w}\cdot x_i + b) - 1 + \xi_i\right] \tag{11}$$

$$- \sum_{i=1}^{m}\beta_i\xi_i \tag{12}$$

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

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \tag{13}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{14}$$

$$\alpha_i + \beta_i = C \tag{15}$$

Simplifying dual objective

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i \qquad \sum_{i=1}^{m} \alpha_i y_i = 0 \qquad \alpha_i + \beta_i = C$$

Simplifying dual objective

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$

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

$$\alpha_i + \beta_i = C$$

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

Dual Problem

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i)$$

$$\text{s.t. } C \geq \alpha_i \geq 0, i \in [1, m]$$

$$\sum_i \alpha_i y_i = 0$$

Dual Problem

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i)$$

$$\text{s.t. } C \geq \alpha_i \geq 0, i \in [1, m]$$

$$\sum_i \alpha_i y_i = 0$$

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, C \geq \alpha_i \geq 0, \beta_i \geq 0 \tag{16}$$

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, C \geq \alpha_i \geq 0, \beta_i \geq 0 \tag{16}$$

Stationarity

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i + \beta_i = C \tag{17}$$

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, C \geq \alpha_i \geq 0, \beta_i \geq 0 \tag{16}$$

Stationarity

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i + \beta_i = C \tag{17}$$

Complementary slackness

$$\alpha_i[y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i] = 0, \beta_i \xi_i = 0 \tag{18}$$

More on Complementary Slackness

$$\alpha_i[y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i] = 0, \beta_i \xi_i = 0 \tag{19}$$

- $\boldsymbol{x}_i$ satisfies the margin $\Rightarrow y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$

More on Complementary Slackness

$$\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1+\xi_i]=0, \beta_i\xi_i=0 \tag{19}$$

- $\boldsymbol{x}_i$ satisfies the margin, $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)>1 \Rightarrow \alpha_i=0$
- $\boldsymbol{x}_i$ does not satisfy the margin, $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)<1 \Rightarrow \alpha_i=C$

$\xi_i>0 \quad \beta_i=0 \quad \alpha_i+\beta_i=C$

More on Complementary Slackness

$$\alpha_i[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)-1+\xi_i]=0, \beta_i\xi_i=0 \tag{19}$$

- $\boldsymbol{x}_i$ satisfies the margin, $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)>1 \Rightarrow \alpha_i=0$
- $\boldsymbol{x}_i$ does not satisfy the margin, $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)<1 \Rightarrow \alpha_i=C$
- $\boldsymbol{x}_i$ is on the margin, $y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i+b)=1 \Rightarrow 0\leq\alpha_i\leq C$

Outline

Duality

Slack variables

Sequential Mimimal Optimization

Recap

Sequential Mimimal Optimization

Trivia

- Invented by John Platt in 1998 at Microsoft Research
- Called Minimal due to solving small sub-problems

Dual problem

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i)$$

$$\text{s.t. } C \geq \alpha_i \geq 0, i \in [1, m]$$

$$\sum_i \alpha_i y_i = 0$$

Brief Interlude: Coordinate Ascent

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i)$$

$$\text{s.t. } C \geq \alpha_i \geq 0, i \in [1, m]$$

$$\sum_i \alpha_i y_i = 0$$

Loop over each training example, change α_i to maximize the above function

Brief Interlude: Coordinate Ascent

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_j \cdot \boldsymbol{x}_i)$$

$$\text{s.t. } C \geq \alpha_i \geq 0, i \in [1, m]$$

$$\sum_i \alpha_i y_i = 0$$

Loop over each training example, change α_i to maximize the above function
Although coordinate ascent works OK for lots of problems, we have the constraint $\sum_i \alpha_i y_i = 0$

Outline for SVM Optimization (SMO)

1. Select two examples i, j
2. Update α_j, α_i to maximize the above function

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

$$y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, C \geq \alpha_i \geq 0, \beta_i \geq 0 \tag{20}$$

Stationarity

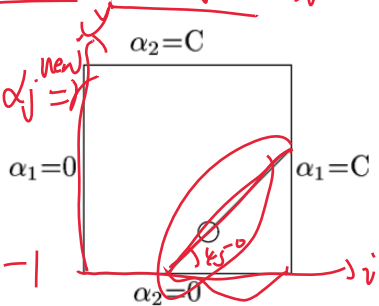
$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i, \sum_{i=1}^{m} \alpha_i y_i = 0, \alpha_i + \beta_i = C \tag{21}$$

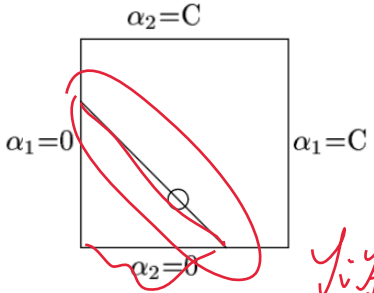
Complementary slackness

$$\alpha_i [y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i] = 0, \beta_i \xi_i = 0 \tag{22}$$

Outline for SVM Optimization (SMO)

$$y_i \alpha_i^{\text{new}} + y_j \alpha_j^{\text{new}} = y_i \alpha_i^{\text{old}} + y_j \alpha_j^{\text{old}} = \gamma$$





Step 2: Optimize α_j

1. Compute upper (H) and lower (L) bounds that ensure $0 \leq \alpha_j \leq C$.

$y_i \neq y_j$

$$L = \max(0, \alpha_j - \alpha_i) \tag{23}$$
$$H = \min(C, C + \alpha_j - \alpha_i) \tag{24}$$

$y_i = y_j$

$$L = \max(0, \alpha_i + \alpha_j - C) \tag{25}$$
$$H = \min(C, \alpha_j + \alpha_i) \tag{26}$$

Step 2: Optimize α_j

1. Compute upper (H) and lower (L) bounds that ensure $0 \leq \alpha_j \leq C$.

$y_i \neq y_j$

$$L = \max(0, \alpha_j - \alpha_i) \quad (23)$$
$$H = \min(C, C + \alpha_j - \alpha_i) \quad (24)$$

$y_i = y_j$

$$L = \max(0, \alpha_i + \alpha_j - C) \quad (25)$$
$$H = \min(C, \alpha_j + \alpha_i) \quad (26)$$

This is because the update for α_i is based on $y_i y_j$ (sign matters)

Step 2: Optimize α_j

Compute errors for i and j

$$E_k \equiv f(x_k) - y_k \tag{27}$$

$$\eta = 2x_i \cdot x_j - x_i \cdot x_i - x_j \cdot x_j \tag{28}$$

for new value for α_j

$$\alpha_j^{\text{new}} = \alpha_j^{\text{old}} - \frac{y_j(E_i - E_j)}{\eta} \tag{29}$$

Step 3: Optimize α_i

Set α_i:

$$\alpha_i^{\text{new}} = \alpha_i^{\text{old}} + y_i y_j \left(\alpha_j^{\text{old}} - \alpha_j^{\text{new}} \right) \tag{30}$$

$$\alpha_i^{new} + y_i y_j \alpha_j^{new} = \alpha_i^{old} + y_i y_j \alpha_i^{old}$$

Step 3: Optimize α_i

Set α_i:

$$\alpha_i^{\text{new}} = \alpha_i^{\text{old}} + y_i y_j \left(\alpha_j^{\text{old}} - \alpha_j^{\text{new}} \right) \tag{30}$$

This balances out the move that we made for α_j.

Overall algorithm

Iterate over $i = \{1, \ldots m\}$

Repeat until KKT conditions are met

Choose j randomly from $m - 1$ other options

Update α_i, α_j

Find $\boldsymbol{w}, b$ based on stationarity conditions

$$w = \sum_i \alpha_i y_i x_i$$

$$b = y_i - w x_i + b \geq 1$$

Iterations / Details

- What if i doesn't violate the KKT conditions?
- What if $\eta \geq 0$? $\eta \leq 0$
- When do we stop?

Iterations / Details

- What if i doesn't violate the KKT conditions? **Skip it!**
- What if $\eta \geq 0$?
- When do we stop?

Iterations / Details

- What if i doesn't violate the KKT conditions? **Skip it!**
- What if $\eta \geq 0$? **Skip it! (should not happen except for numerical instability)**
- When do we stop?

Iterations / Details

- What if i doesn't violate the KKT conditions? **Skip it!**
- What if $\eta \geq 0$? **Skip it! (should not happen except for numerical instability)**
- When do we stop? **Until we go through α's without changing anything**

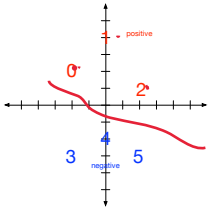
SMO Algorithm

Positive

(-2, 2)
(0, 4)
(2, 1)

Negative

(-2, -3)
(0, -1)
(2, -3)



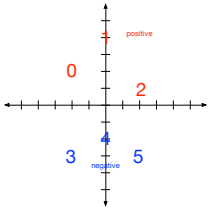
SMO Algorithm

Positive

(-2, 2)
(0, 4)
(2, 1)

Negative

(-2, -3)
(0, -1)
(2, -3)



- Initially, all alphas are zero

$$\boldsymbol{\alpha} = < 0, 0, 0, 0, 0, 0 >$$

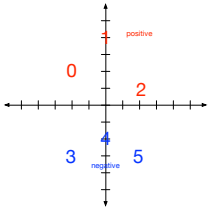
SMO Algorithm

Positive

(-2, 2)
(0, 4)
(2, 1)

Negative

(-2, -3)
(0, -1)
(2, -3)

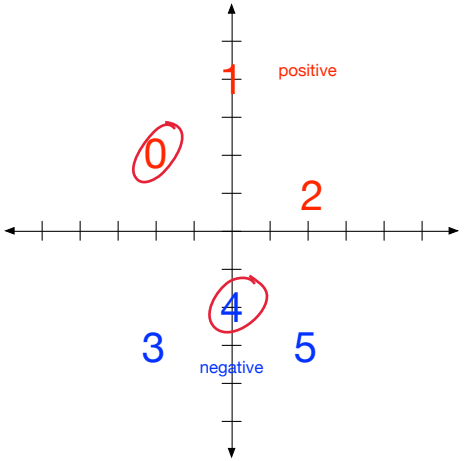


- Initially, all alphas are zero

$$\boldsymbol{\alpha} =< 0, 0, 0, 0, 0, 0 >$$

- Intercept b is also zero
- Capacity $C = \pi$

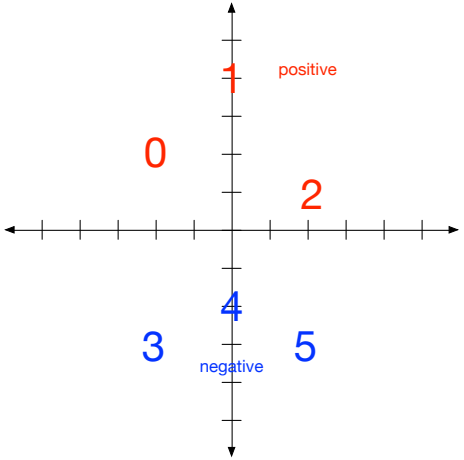
SMO Optimization for $i = 0, j = 4$: Predictions and Step



α_0 α_4

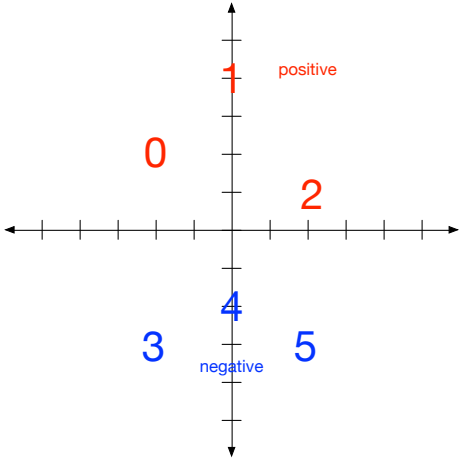
- Prediction: $f(x_0)$
- Prediction: $f(x_4)$
- Error: E_0
- Error: E_4

SMO Optimization for $i = 0, j = 4$: Predictions and Step



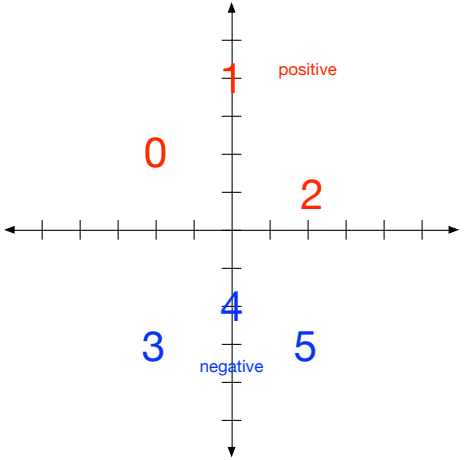
- Prediction: $f(x_0) = 0$
- Prediction: $f(x_4)$
- Error: E_0
- Error: E_4

SMO Optimization for $i = 0, j = 4$: Predictions and Step



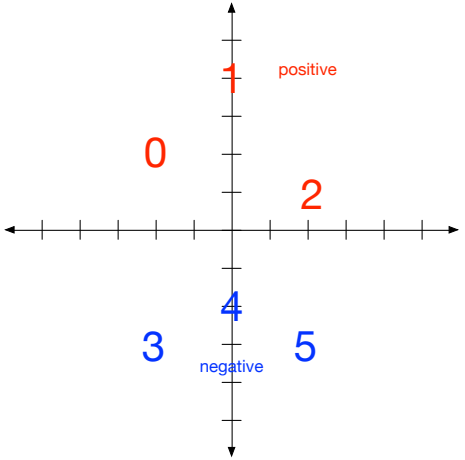
- Prediction: $f(x_0) = 0$
- Prediction: $f(x_4) = 0$
- Error: E_0
- Error: E_4

SMO Optimization for $i = 0, j = 4$: Predictions and Step



- Prediction: $f(x_0) = 0$
- Prediction: $f(x_4) = 0$
- Error: $E_0 = -1$
- Error: $E_4 = +1$

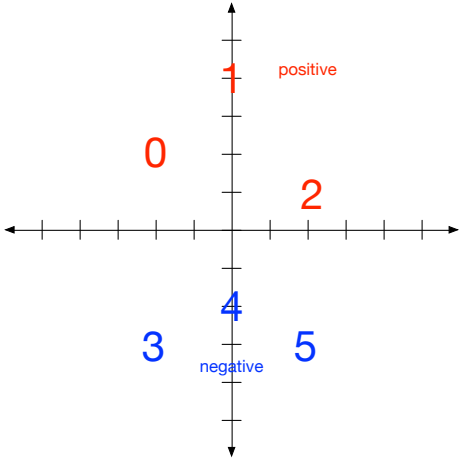
SMO Optimization for $i = 0, j = 4$: Predictions and Step



- Prediction: $f(x_0) = 0$
- Prediction: $f(x_4) = 0$
- Error: $E_0 = -1$
- Error: $E_4 = +1$

$$\eta = 2\langle x_0, x_4 \rangle - \langle x_0, x_0 \rangle - \langle x_4, x_4 \rangle$$

SMO Optimization for $i = 0, j = 4$: Predictions and Step



- Prediction: $f(x_0) = 0$
- Prediction: $f(x_4) = 0$
- Error: $E_0 = -1$
- Error: $E_4 = +1$

$$\eta = 2\langle x_0, x_4 \rangle - \langle x_0, x_0 \rangle - \langle x_4, x_4 \rangle$$
$$= 2 \cdot -2 - 8 - 1 = -13$$

SMO Optimization for $i = 0, j = 4$: Bounds

- Lower and upper bounds for α_j

$$L = \max(0, \alpha_j - \alpha_i) \tag{31}$$

$$H = \min(C, C + \alpha_j - \alpha_i) \tag{32}$$

SMO Optimization for $i = 0, j = 4$: Bounds

- Lower and upper bounds for α_j

$$L = \max(0, \alpha_j - \alpha_i) = 0 \tag{31}$$

$$H = \min(C, C + \alpha_j - \alpha_i) \tag{32}$$

SMO Optimization for $i = 0, j = 4$: Bounds

- Lower and upper bounds for α_j

$$L = \max(0, \alpha_j - \alpha_i) = 0 \tag{31}$$

$$H = \min(C, C + \alpha_j - \alpha_i) = \pi \tag{32}$$

SMO Optimization for $i = 0, j = 4$: α update

New value for α_j

$$\alpha_j^* = \alpha_j - \frac{y_j(E_i - E_j)}{\eta} \tag{33}$$

(34)

SMO Optimization for $i = 0, j = 4$: α update

New value for α_j

$$\alpha_j^* = \alpha_j - \frac{y_j(E_i - E_j)}{\eta} = \frac{-2}{\eta} = \frac{2}{13} \tag{33}$$

(34)

SMO Optimization for $i = 0, j = 4$: α update

New value for α_j

$$\alpha_j^* = \alpha_j - \frac{y_j(E_i - E_j)}{\eta} = \frac{-2}{\eta} = \frac{2}{13} \tag{33}$$

New value for α_i

(34)

SMO Optimization for $i = 0, j = 4$: α update

New value for α_j

$$\alpha_j^* = \alpha_j - \frac{y_j(E_i - E_j)}{\eta} = \frac{-2}{\eta} = \frac{2}{13} \tag{33}$$

New value for α_i

$$\alpha_i^* = \alpha_i + y_i y_j \left(\alpha_j^{(old)} - \alpha_j\right) \tag{34}$$

SMO Optimization for $i = 0, j = 4$: α update

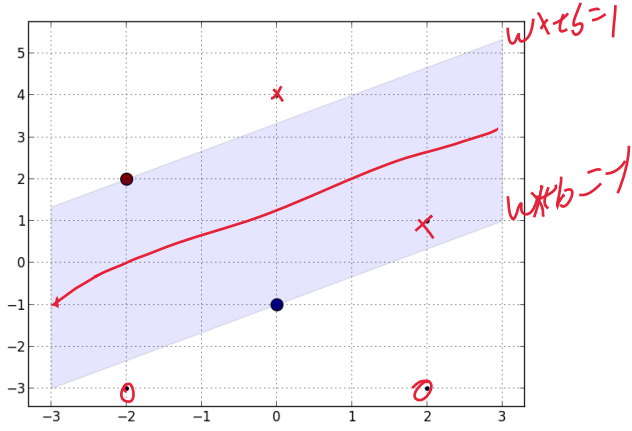
New value for α_j

$$\alpha_j^* = \alpha_j - \frac{y_j(E_i - E_j)}{\eta} = \frac{-2}{\eta} = \frac{2}{13} \quad (33)$$

New value for α_i

$$\alpha_i^* = \alpha_i + y_i y_j \left(\alpha_j^{(old)} - \alpha_j\right) = \alpha_j = \frac{2}{13} \quad (34)$$

Margin



Find weight vector and bias

- Weight vector

$$\vec{w} = \sum_{i}^{m} \alpha_i y_i \vec{x}_i \tag{35}$$

- Bias

$$b = b^{(old)} - E_i - y_i(\alpha_i^* - \alpha_i^{(old)})x_i \cdot x_i - y_j(\alpha_j^* - \alpha_j^{(old)})x_i \cdot x_j \tag{36}$$

$$\tag{37}$$

Find weight vector and bias

- Weight vector

$$\vec{w} = \sum_{i}^{m} \alpha_i y_i \vec{x_i} = \frac{2}{13}\begin{bmatrix} -2 \\ 2 \end{bmatrix} - \frac{2}{13}\begin{bmatrix} 0 \\ -1 \end{bmatrix} \tag{35}$$

- Bias

$$b = b^{(old)} - E_i - y_i(\alpha_i^* - \alpha_i^{(old)})x_i \cdot x_i - y_j(\alpha_j^* - \alpha_j^{(old)})x_i \cdot x_j \tag{36}$$

$$\tag{37}$$

Find weight vector and bias

- Weight vector

$$\vec{w} = \sum_{i}^{m} \alpha_i y_i \vec{x}_i = \frac{2}{13}\begin{bmatrix}-2\\2\end{bmatrix} - \frac{2}{13}\begin{bmatrix}0\\-1\end{bmatrix} = \begin{bmatrix}\frac{-4}{13}\\\frac{6}{13}\end{bmatrix} \tag{35}$$

- Bias

$$b = b^{(old)} - E_i - y_i(\alpha_i^* - \alpha_i^{(old)})x_i \cdot x_i - y_j(\alpha_j^* - \alpha_j^{(old)})x_i \cdot x_j \tag{36}$$

$$\tag{37}$$

Find weight vector and bias

- Weight vector

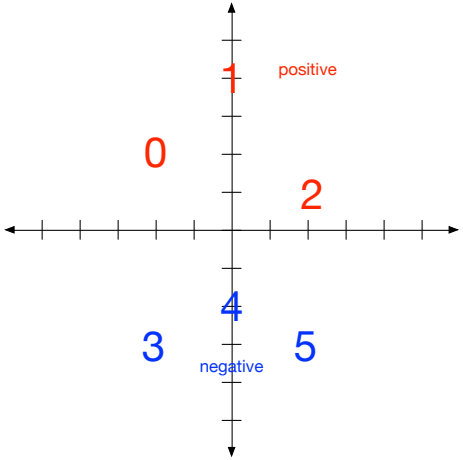
$$\vec{w} = \sum_{i}^{m} \alpha_i y_i \vec{x}_i = \frac{2}{13}\begin{bmatrix}-2\\2\end{bmatrix} - \frac{2}{13}\begin{bmatrix}0\\-1\end{bmatrix} = \begin{bmatrix}\frac{-4}{13}\\\frac{6}{13}\end{bmatrix} \tag{35}$$

- Bias

$$b = b^{(old)} - E_i - y_i(\alpha_i^* - \alpha_i^{(old)})x_i \cdot x_i - y_j(\alpha_j^* - \alpha_j^{(old)})x_i \cdot x_j \tag{36}$$

$$= 1 - \frac{2}{13} \cdot 8 + \frac{2}{13} \cdot -2 = -0.54 \tag{37}$$

SMO Optimization for $i = 2, j = 4$

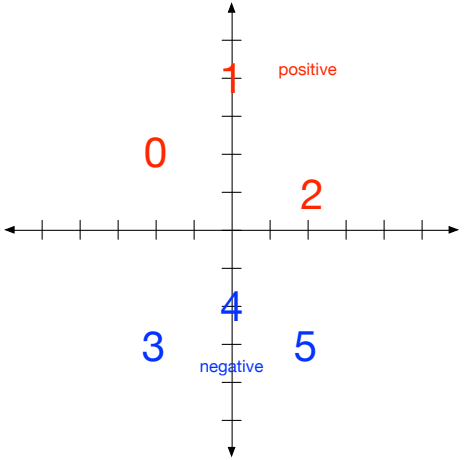


Let's skip the boring stuff

- $E_2 = -1.69$
- $E_4 = 0.00$
- $\eta = -8$

- $\alpha_4 = \alpha_j^{(old)} - \frac{y_j(E_i - E_j)}{\eta}$
- $\alpha_2 = \alpha_i^{(old)} + y_i y_j \left(\alpha_j^{(old)} - \alpha_j \right)$

SMO Optimization for $i = 2, j = 4$

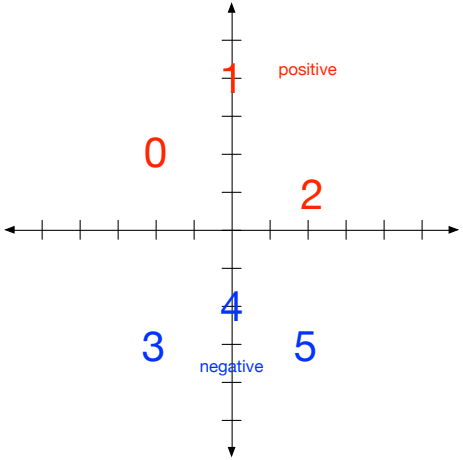


Let's skip the boring stuff

- $E_2 = -1.69$
- $E_4 = 0.00$
- $\eta = -8$

- $\alpha_4 = \alpha_j^{(old)} - \frac{y_j(E_i - E_j)}{\eta}$
- $\alpha_2 = \alpha_i^{(old)} + y_i y_j \left(\alpha_j^{(old)} - \alpha_j \right)$

SMO Optimization for $i = 2, j = 4$

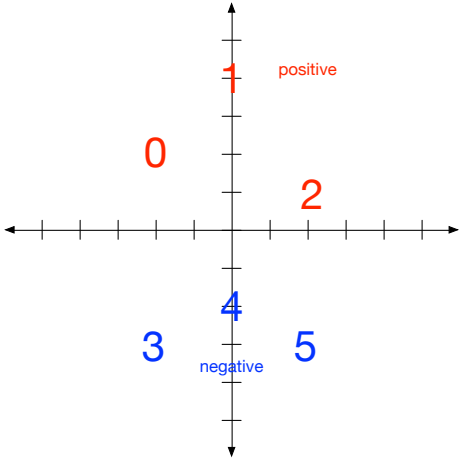


Let's skip the boring stuff

- $E_2 = -1.69$
- $E_4 = 0.00$
- $\eta = -8$

- $\alpha_4 = \alpha_j^{(old)} - \frac{y_j(E_i - E_j)}{\eta} = 0.15 + \frac{-1.69}{-8} = 0.37$
- $\alpha_2 = \alpha_i^{(old)} + y_i y_j \left(\alpha_j^{(old)} - \alpha_j\right)$

SMO Optimization for $i = 2, j = 4$

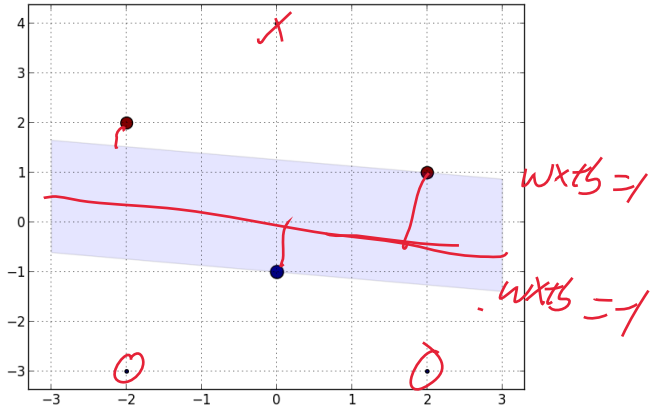


Let's skip the boring stuff

- $E_2 = -1.69$
- $E_4 = 0.00$
- $\eta = -8$

- $\alpha_4 = \alpha_j^{(old)} - \frac{y_j(E_i - E_j)}{\eta} = 0.15 + \frac{-1.69}{-8} = 0.37$
- $\alpha_2 = \alpha_i^{(old)} + y_i y_j \left(\alpha_j^{(old)} - \alpha_j\right) = 0 - (0.15 - 0.37) = 0.21$

Margin



Weight vector and bias

- Bias $b = -0.12$
- Weight vector

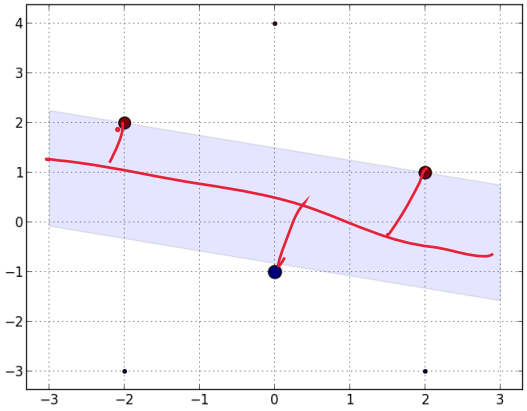
$$\vec{w} = \sum_{i}^{m} \alpha_i y_i \vec{x}_i \tag{38}$$

Weight vector and bias

- Bias $b = -0.12$
- Weight vector

$$\vec{w} = \sum_{i}^{m} \alpha_i y_i \vec{x}_i = \begin{bmatrix} 0.12 \\ 0.88 \end{bmatrix} \tag{38}$$

Another Iteration $(i = 0, j = 2)$



SMO Algorithm

- Convenient approach for solving: vanilla, slack, kernel approaches
- Convex problem
- Scalable to large datasets (implemented in scikit learn)
- What we didn't do:
  - Check KKT conditions
  - Randomly choose indices

Outline

Duality

Slack variables

Sequential Mimimal Optimization

Recap

Recap

- Duality
- Slack variables

Simpler form

Soft margin

Recap

- SMO: Optimize objective function for two data points
- Convex problem: Will converge
- Relatively fast
- Gives good performance

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 up: ~~simple algorithm for finding SVMs~~

kernels a