Machine Learning: Chenhao Tan

University of Colorado Boulder

LECTURE 17

Slides adapted from Jordan Boyd-Graber, Chris Ketelsen

Logistics

- Homework 3 is due on Sunday!
- Project team matches

Roadmap

- Last time: linear SVM formulation when data is linearly separable
- This time:
  - Make linear SVM work when data is not linearly separable
  - Introduce duality
- Next week: KKT conditions & Kernel tricks

Overview

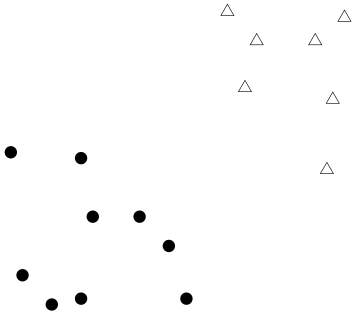
Soft-margin SVM

Duality

Outline

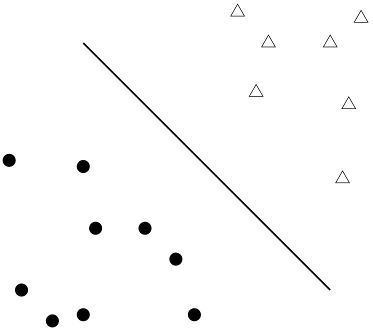
Support Vector Machines

- 2-class training data



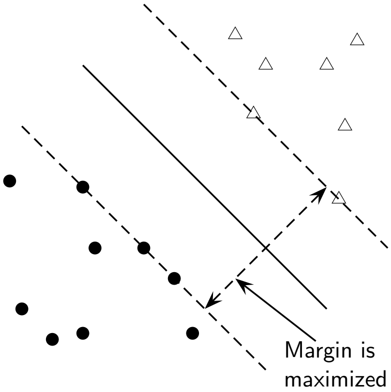
Support Vector Machines

- 2-class training data
- decision boundary $\rightarrow$ **linear separator**



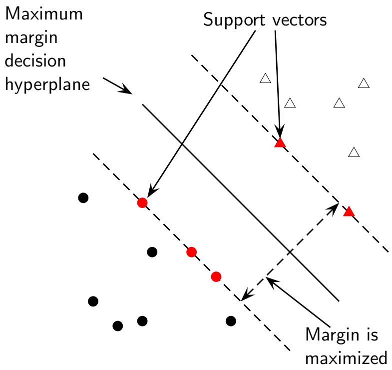
Support Vector Machines

- 2-class training data
- decision boundary $\rightarrow$ **linear separator**
- criterion: being maximally far away from any data point $\rightarrow$ determines classifier **margin**



Support Vector Machines

- 2-class training data
- decision boundary $\rightarrow$ **linear separator**
- criterion: being maximally far away from any data point $\rightarrow$ determines classifier **margin**
- linear separator position defined by **support vectors**
- other points have no impact on the decision boundary



Objective function for hard-margin SVM

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

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

Theoretical evidence that suggests SVMs will Work

- Leave-one-out error
- Margin analysis (omitted)
- VC Dimension (omitted for now)

Leave-one-out error (sketch)

Leave one out error is the error by using one point as your test set (averaged over all such points).

$$\hat{R}_{LOO} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\left[h_{s-\{x_i\}} \neq y_i\right]$$

Leave-one-out error (sketch)

Leave one out error is the error by using one point as your test set (averaged over all such points).

$$\hat{R}_{LOO} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\left[h_{s-\{x_i\}} \neq y_i\right]$$

This serves as an unbiased estimate of generalization error for samples of size $m-1$.

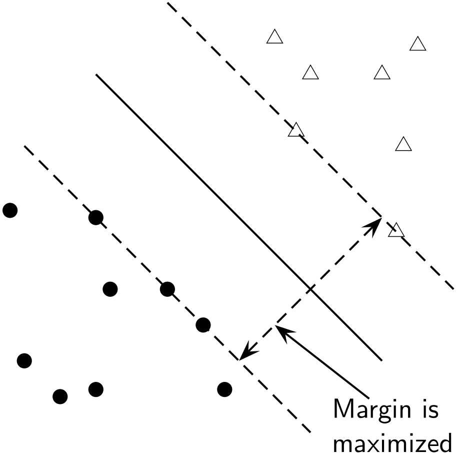
Leave-one-out error (sketch)

Leave-one-out error is bounded by the number of support vectors.

$$\mathbb{E}_{S \sim D^{m-1}}\left[R(h_s)\right] \leq \mathbb{E}_{S \sim D^m}\left[\frac{N_{SV}(S)}{m}\right]$$

Consider the held out error for x_i.

Pictorial proof



Leave-one-out error (sketch)

Leave-one-out error is bounded by the number of support vectors.

$$\mathbb{E}_{S\sim D^{m-1}}\left[R(h_s)\right] \leq \mathbb{E}_{S\sim D^m}\left[\frac{N_{SV}(S)}{m}\right]$$

Consider the held out error for x_i.

- If x_i was not a support vector, the answer doesn't change.
- If x_i was a support vector, it could change the answer; this is when we can have an error.

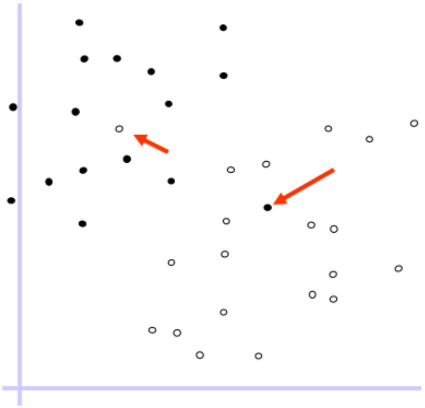
There are N_{SV} support vectors and thus N_{SV} possible errors.

Objective function for hard-margin SVM

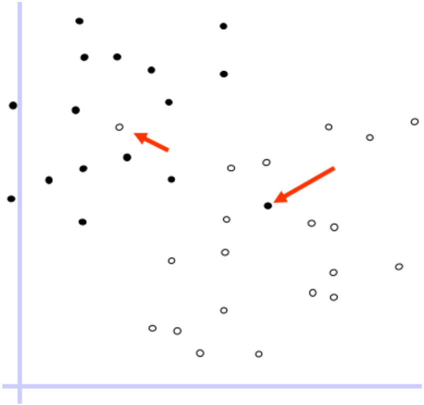
$$\min_{\boldsymbol{w},b} \frac{1}{2}||\boldsymbol{w}||^2$$

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

Can SVMs Work Here?

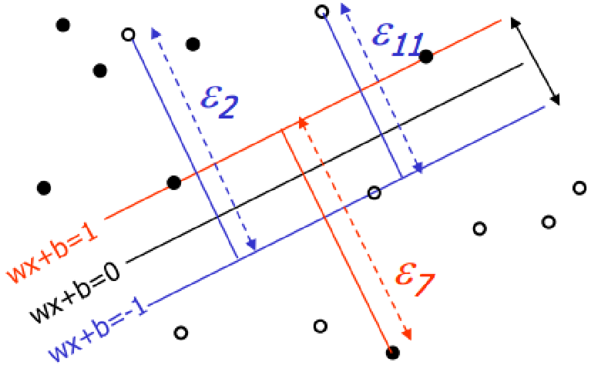


Can SVMs Work Here?



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

Trick: Allow for a few bad apples



Hard-margin objective function

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

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

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

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

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

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

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

- Standard margin

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

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

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

New objective function

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

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

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

What is the role of C?

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

subject to

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

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

B. $C \uparrow \Rightarrow$ decrease bias, increase variance

C. $C \uparrow \Rightarrow$ increase bias, decrease variance

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

Outline

Soft-margin SVM

Duality

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

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

subject to $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) \geq 1, i \in [1, m]$
This is a quadratic objective function with linear inequality constraints. Many off-the-shelf optimization methods are available.

Optimizing Constrained Functions

The Method of Lagrange Multipliers

Constrained problem (Primal problem)

$$\min_{\boldsymbol{x}} f(\boldsymbol{x})$$

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

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

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

Primal problem

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

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

Derive the function for dual problem. Replace $\boldsymbol{w}, b$ with stationarity conditions. (There will be detailed derivations for the soft-margin case later.)

Primal vs. Dual

Primal problem

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

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

Dual problem

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

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

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

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

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

Stationarity

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

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

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

Stationarity

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

Remember that two properties about support vector machine directly follows from this:

- Only support vectors affect the weights ($\alpha_i > 0$).
- There must be both positive and negative support vectors.

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

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

Stationarity

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

Complementary slackness

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

What is the dual problem of soft-margin SVM?

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

subject to

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

New Lagrangian

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

New Lagrangian

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

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

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

New Lagrangian

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

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

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

New Lagrangian

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \\
& - \sum_{i=1}^{m}\alpha_i\left[y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) - 1 + \xi_i\right] \\
& - \sum_{i=1}^{m}\beta_i\xi_i
\end{aligned}
$$

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

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

New Lagrangian

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m} \xi_i \\
& - \sum_{i=1}^{m} \alpha_i \left[y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i \right] \\
& - \sum_{i=1}^{m} \beta_i \xi_i
\end{aligned}
$$

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

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

Simplifying dual objective

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

Simplifying dual objective

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

$$\begin{aligned}
\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m} \xi_i \\
& - \sum_{i=1}^{m} \alpha_i \left[y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) - 1 + \xi_i \right] \\
& - \sum_{i=1}^{m} \beta_i \xi_i
\end{aligned}$$

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

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

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

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

Stationarity

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

Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

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

Stationarity

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

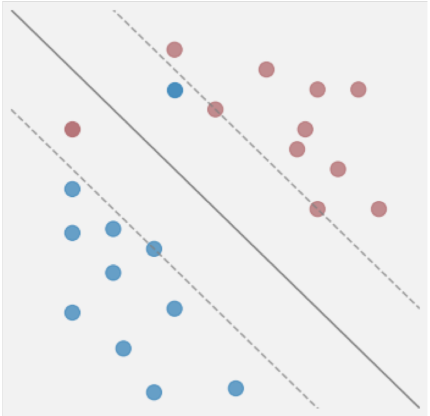
Complementary slackness

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

More on Complementary Slackness

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

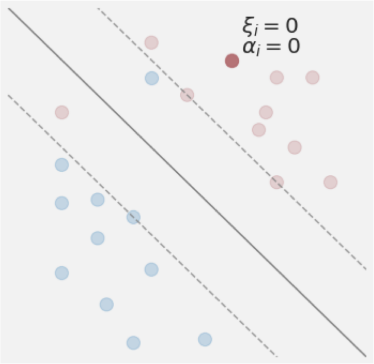
Also, $\alpha_i + \beta_i = C$



More on Complementary Slackness

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

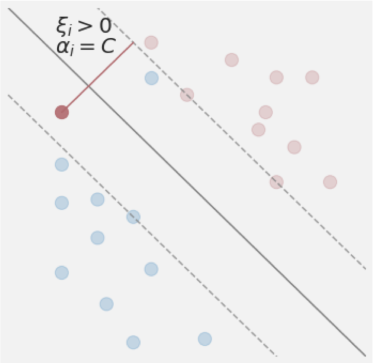
- $\boldsymbol{x}_i$ satisfies the margin, $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$



More on Complementary Slackness

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

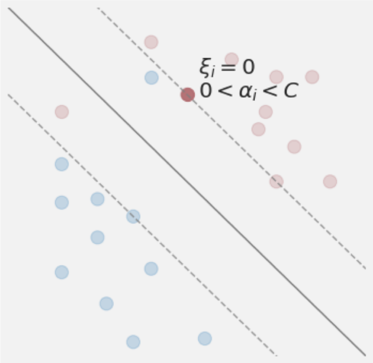
- $\boldsymbol{x}_i$ satisfies the margin,
  $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$
- $\boldsymbol{x}_i$ does not satisfy the margin,
  $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) < 1 \Rightarrow \alpha_i = C$



More on Complementary Slackness

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

- $\boldsymbol{x}_i$ 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, $y_i(\boldsymbol{w} \cdot \boldsymbol{x}_i + b) = 1 \Rightarrow 0 \leq \alpha_i \leq C$



Karush-Kuhn-Tucker (KKT) conditions

Primal and dual feasibility

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

Stationarity

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

Complementary slackness

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