Machine Learning: Yoshinari Fujinuma

University of Colorado Boulder

LECTURE 19

Slides adapted from Chenhao Tan, Jordan Boyd-Graber, Chris Ketelsen

Logistics

- Homework 3 is due next Monday
- Final project proposal is due next Friday

Outline

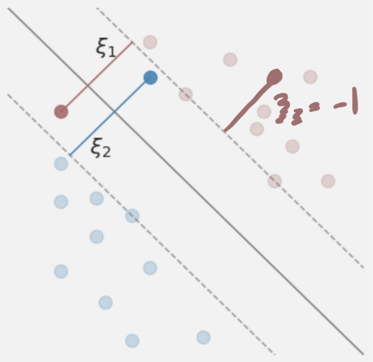
Recap: Soft-margin SVM and intuition of ξ

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

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

$$\xi_i \geq 0, i \in [1, m]$$

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



Primal problem

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

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

$$\xi_i \geq 0, i \in [1, m]$$

Dual problem

Our goal today: derive the following dual problem from the primal SVM problem

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

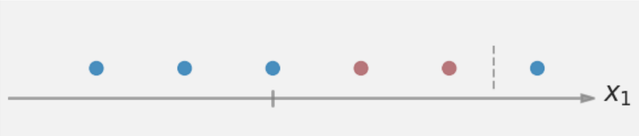
subject to $0 \leq \alpha_i \leq C, i \in [1, m]$

$$\sum_i \alpha_i y_i = 0$$

Pros: $\boldsymbol{x}_j^T \boldsymbol{x}_i$ is a dot product, using a projection ϕ i.e., $\phi(\boldsymbol{x}_j)^T \phi(\boldsymbol{x}_i)$ would make SVM non-linear classifier.

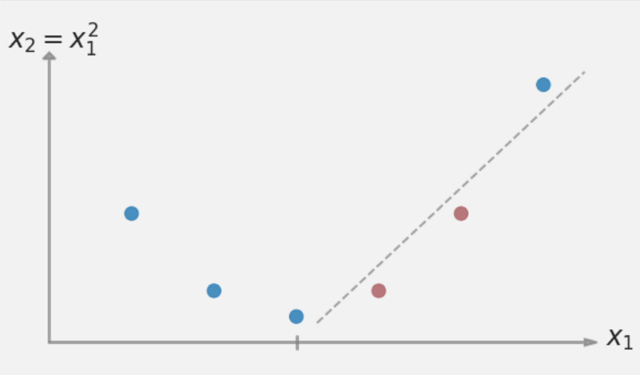
Kernels: Motivating Why do We want Dual Problem

What can we do if the data is clearly not linearly separable?



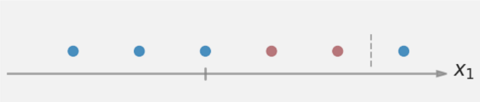
Kernels: Motivating Why do We want Dual Problem

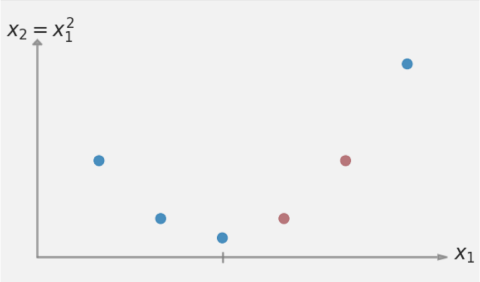
Add a dimension.



Derived features

We started with the original feature vector, $\boldsymbol{x} = (x_1)$,
and we created a new derived feature vector, $\phi(\boldsymbol{x}) = (x_1, x_1^2)$.





Dual problem

Our goal today: derive the following dual problem

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

subject to $0 \leq \alpha_i \leq C, i \in [1, m]$

$$\sum_i \alpha_i y_i = 0$$

Pros: $\boldsymbol{x}_j^T \boldsymbol{x}_i$ is a dot product, using a projection ϕ i.e., $\phi(\boldsymbol{x}_j)^T \phi(\boldsymbol{x}_i)$ would make SVM non-linear classifier.
But let's start from figuring out what are α_i...

Background: Convex Optimization Problem with Inequality Constraints

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f(w) \\ \text{subject to} \quad & g_i(w) \leq 0, i \in [1, m] \end{aligned}$$

Define a modified objective function called **Lagrangian**:

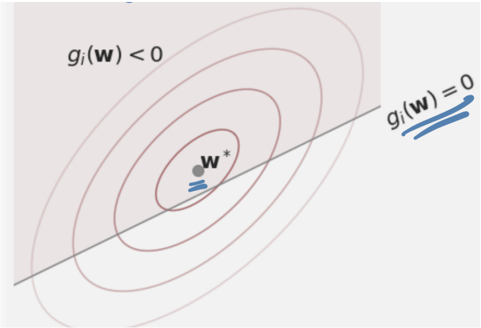
$$L(\boldsymbol{w}, \alpha) = f(\boldsymbol{w}) + \sum_i \alpha_i g_i(\boldsymbol{w})$$

where $\alpha_i \geq 0$ are called **Lagrange Multipliers**

Background: Intuition of α_i

$$L(\boldsymbol{w}, \alpha) = f(\boldsymbol{w}) + \sum_i \alpha_i g_i(\boldsymbol{w})$$

Suppose $\boldsymbol{w}^*$ minimizes $f(\boldsymbol{w})$
The constraint $g_i(\boldsymbol{w}) < 0$ is inactive
The associated Lagrange multiplier
$\alpha_i = 0$



Background: Intuition of α_i

$$L(\boldsymbol{w}, \alpha) = f(\boldsymbol{w}) + \sum_i \alpha_i g_i(\boldsymbol{w})$$

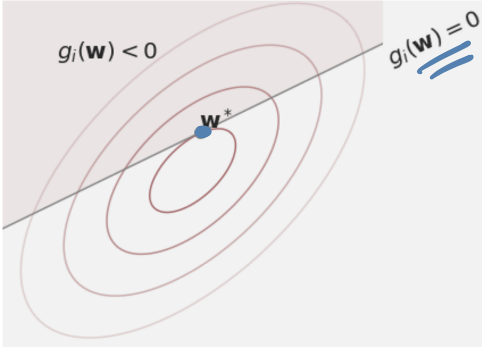
Suppose $\boldsymbol{w}^*$ minimizes $f(\boldsymbol{w})$

The constraint $g_i(\boldsymbol{w}) < 0$ is active

The associated Lagrange multiplier

$\alpha_i > 0$

In SVM: Only support vectors affect the weights ($\alpha_i > 0$).



Background: Lagrangian Formulation of the Primal Problem

$$\min_{\boldsymbol{w}} \quad \max_{\boldsymbol{\alpha}} f(\boldsymbol{w}) + \sum_i \alpha_i g_i(\boldsymbol{w})$$

$$\text{subject to} \quad \alpha_i(\boldsymbol{w}) \geq 0, i \in [1, m]$$

Let's switch $\min$ and $\max$ to help leading to the dual problem

$$\max_{\boldsymbol{\alpha}} \quad \min_{\boldsymbol{w}} f(\boldsymbol{w}) + \sum_i \alpha_i g_i(\boldsymbol{w})$$

$$\text{subject to} \quad \alpha_i(\boldsymbol{w}) \geq 0, i \in [1, m]$$

Background: Primal and Dual Problem

By defining $h(\alpha) = \min_{\boldsymbol{w}} f(\boldsymbol{w}) + \sum_i \alpha_i g_i(\boldsymbol{w})$,

Primal Problem

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f(\boldsymbol{w}) \\ \text{subject to} \quad & g_i(\boldsymbol{w}) \leq 0, i \in [1, m] \end{aligned}$$

Dual Problem

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & h(\alpha) \\ \text{subject to} \quad & \alpha_i \geq 0, i \in [1, m] \end{aligned}$$

The optimal values of primal and dual problems matches under **KKT condition**
But let's first derive a dual problem of SVM.

Lagrangian of SVM Optimization Problem

Recap of the SVM Optimization Problem:

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

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

$$\xi_i \geq 0, i \in [1, m]$$

Lagrangian of SVM

Using Lagrange multipliers α and β for the two constraints,

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

Lagrangian of SVM

Using Lagrange multipliers α and β for the two constraints,

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

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

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

Lagrangian of SVM

Using Lagrange multipliers α and β for the two constraints,

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

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

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

Lagrangian of SVM

Using Lagrange multipliers α and β for the two constraints,

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

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

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

Lagrangian of SVM

Using Lagrange multipliers α and β for the two constraints,

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

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

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

Lagrangian of SVM

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

Using $\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i$, $\sum_{i=1}^{m}\alpha_i y_i = 0$, and $\alpha_i + \beta_i = C$, the dual objective function is

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

Karush-Kuhn-Tucker (KKT) conditions for Soft-Margin SVM

Primal and dual feasibility

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

Karush-Kuhn-Tucker (KKT) conditions for Soft-Margin SVM

Primal and dual feasibility

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

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 for Soft-Margin SVM

Primal and dual feasibility

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

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

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

More on Complementary Slackness

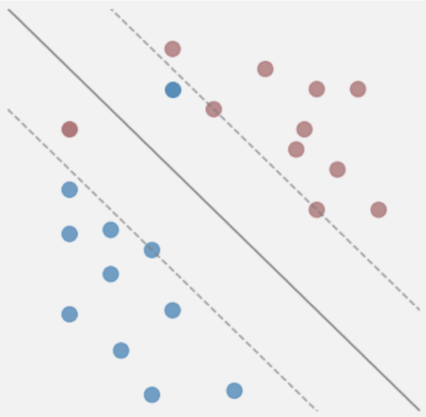
$$\alpha_i[y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 + \xi_i] = 0, \beta_i\xi_i = 0$$

Also, $\alpha_i + \beta_i = C$

Also looking at the feasibility conditions:

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

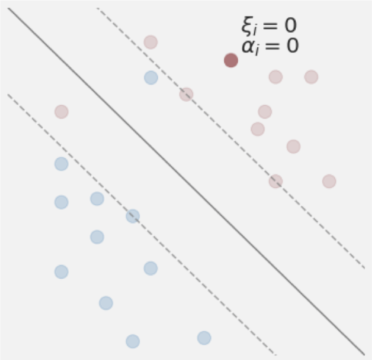
E.g., When $\beta_i > 0$ then $\xi_i = 0$.
Therefore, it's called "complementary"



More on Complementary Slackness

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

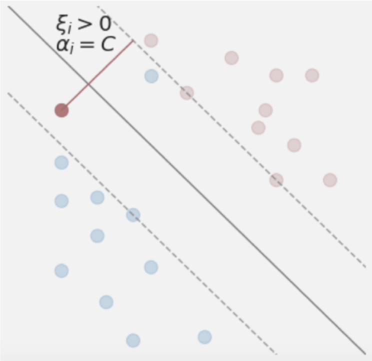
- $\boldsymbol{x}_i$ satisfies the margin ($\xi = 0$), $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$



More on Complementary Slackness

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

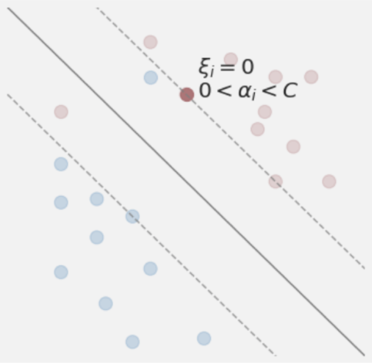
- $\boldsymbol{x}_i$ satisfies the margin ($\xi = 0$), $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$
- $\boldsymbol{x}_i$ does not satisfy the margin ($\xi > 0$), then $\beta = 0$, $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < 1 \Rightarrow \alpha_i = C$



More on Complementary Slackness

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

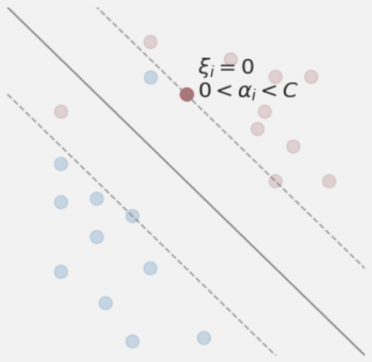
- $\boldsymbol{x}_i$ satisfies the margin ($\xi = 0$), $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$
- $\boldsymbol{x}_i$ does not satisfy the margin ($\xi > 0$), then $\beta = 0$, $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < 1 \Rightarrow \alpha_i = C$
- $\boldsymbol{x}_i$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



More on Complementary Slackness

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

- $\boldsymbol{x}_i$ satisfies the margin ($\xi = 0$), $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 1 \Rightarrow \alpha_i = 0$
- $\boldsymbol{x}_i$ does not satisfy the margin ($\xi > 0$), then $\beta = 0$, $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) < 1 \Rightarrow \alpha_i = C$
- $\boldsymbol{x}_i$ is on the margin ($\xi = 0$), $\beta > 0$, $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1 \Rightarrow 0 \leq \alpha_i \leq C$
- Gathering everything, we end up with the constraint $0 \leq \alpha_i \leq C$



Dual problem

Our goal today: derive the following dual problem

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

subject to $0 \leq \alpha_i \leq C, i \in [1, m]$

$$\sum_i \alpha_i y_i = 0$$