

## Lecture 12: Classification with Support Vector Machines

Yi, Yung (이웅)

Mathematics for Machine Learning  
<https://yung-web.github.io/home/courses/mathml.html>  
KAIST EE

April 3, 2021

Please watch this tutorial video by Luis Serrano on Support Vector Machine.

[https://youtu.be/Lpr\\_\\_X8zuE8](https://youtu.be/Lpr__X8zuE8)

April 3, 2021 1 / 28

April 3, 2021 2 / 28

- (1) Story and Separating Hyperplanes
- (2) Primal SVM: Hard SVM
- (3) Primal SVM: Soft SVM
- (4) Dual SVM
- (5) Kernels
- (6) Numerical Solution

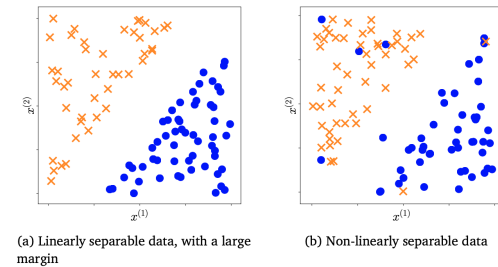
- (1) Story and Separating Hyperplanes
- (2) Primal SVM: Hard SVM
- (3) Primal SVM: Soft SVM
- (4) Dual SVM
- (5) Kernels
- (6) Numerical Solution

April 3, 2021 3 / 28

L12(1)

April 3, 2021 4 / 28

- (Binary) classification vs. regression
- A Classification predictor  $f : \mathbb{R}^D \mapsto \{+1, -1\}$ , where  $D$  is the dimension of features.
- Supervised learning as in the regression with a given dataset  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ , where our task is to learn the model parameters which produces the smallest classification errors.
- SVM
  - Geometric way of thinking about supervised learning
  - Relying on empirical risk minimization
  - Binary classification = Drawing a separating hyperplane
  - Various interpretation from various perspectives: geometric view, loss function view, the view from convex hulls of data points



- Hard SVM: Linearly separable, and thus, allow no classification error
- Soft SVM: Non-linearly separable, thus, allow some classification error

L12(1)

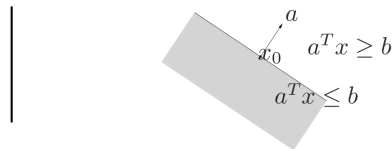
April 3, 2021 5 / 28

L12(1)

April 3, 2021 6 / 28

## Separating Hyperplane

- **Hyperplane** in  $\mathbb{R}^D$  is a set:  $\{x \mid a^T x = b\}$  where  $a \in \mathbb{R}^n, a \neq 0, b \in \mathbb{R}$  L7(3)  
In other words,  $\{x \mid a^T(x - x_0) = 0\}$ , where  $x_0$  is any point in the hyperplane, i.e.,  $a^T x_0 = b$ .



- Divides  $\mathbb{R}^D$  into two **halfspaces**:  $\{x \mid a^T x \leq b\}$  and  $\{x \mid a^T x > b\}$

- In our problem, we consider the hyperplane  $\mathbf{w}^T \mathbf{x} + b = 0$ , where  $\mathbf{w}$  and  $b$  are the parameters of the model.

- Classification logic

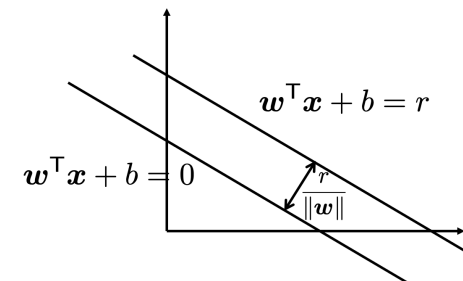
$$\begin{cases} \mathbf{w}^T \mathbf{x}_n + b \geq 0 & \text{when } y_n = +1 \\ \mathbf{w}^T \mathbf{x}_n + b < 0 & \text{when } y_n = -1 \end{cases} \implies y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 0$$

L12(1)

April 3, 2021 7 / 28

## Distance between Two Hyperplanes

- Consider two hyperplanes  $\mathbf{w}^T \mathbf{x} - b = 0$  and  $\mathbf{w}^T \mathbf{x} - b = r$ , where assume  $r > 0$ .
- **Question**. What is the distance<sup>1</sup> between two hyperplanes? Answer:  $\frac{r}{\|\mathbf{w}\|}$



<sup>1</sup>Shortest distance between two hyperplanes.

L12(1)

April 3, 2021 8 / 28

- (1) Story and Separating Hyperplanes
- (2) **Primal SVM: Hard SVM**
- (3) Primal SVM: Soft SVM
- (4) Dual SVM
- (5) Kernels
- (6) Numerical Solution

- Assume that the data points are linearly separable.
- Goal: Find the hyperplane that maximizes the margin between the positive and the negative samples
- Given the training dataset  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$  and a hyperplane  $\mathbf{w}^T \mathbf{x} + b = 0$ , what is the constraint that all data points are  $\frac{r}{\|\mathbf{w}\|}$ -away from the hyperplane?

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq \frac{r}{\|\mathbf{w}\|}$$

- Note that  $r$  and  $\|\mathbf{w}\|$  are scaled together, so if we fix  $\|\mathbf{w}\| = 1$ , then

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq r$$

## Hard SVM: Formulation 1

- Maximize the margin, such that all the training data points are well-classified into their classes (+ or -)

$$\begin{aligned} & \max_{\mathbf{w}, b, r} \quad r \\ & \text{subject to} \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq r, \text{ for all } n = 1, \dots, N, \quad \|\mathbf{w}\| = 1, \quad r > 0 \end{aligned}$$

## Formulation 2 (1)

$$\begin{aligned} & \max_{\mathbf{w}, b, r} \quad r \\ & \text{subject to} \quad y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq r, \text{ for all } n = 1, \dots, N, \quad \|\mathbf{w}\| = 1, \quad r > 0 \end{aligned}$$

- Since  $\|\mathbf{w}\| = 1$ , reformulate  $\mathbf{w}$  by  $\mathbf{w}'$  as:  $y_n\left(\frac{\mathbf{w}'^T}{\|\mathbf{w}'\|} \mathbf{x}_n + b\right) \geq r$
- Change the objective from  $r$  to  $r^2$ .
- Define  $\mathbf{w}''$  and  $b''$  by rescaling the constraint:

$$y_n\left(\frac{\mathbf{w}'^T}{\|\mathbf{w}'\|} \mathbf{x}_n + b\right) \geq r \iff y_n(\mathbf{w}''^T \mathbf{x}_n + b'') \geq 1, \quad \mathbf{w}'' = \frac{\mathbf{w}'}{\|\mathbf{w}'\| r} \text{ and } b'' = \frac{b}{r}$$

- Note that  $\|\mathbf{w}''\| = \frac{1}{r}$
- Thus, we have the following reformulated problem:

$$\begin{aligned} \max_{\mathbf{w}'', b''} \quad & \frac{1}{\|\mathbf{w}''\|^2} \\ \text{subject to} \quad & y_n(\mathbf{w}''^T \mathbf{x}_n + b'') \geq 1, \text{ for all } n = 1, \dots, N, \end{aligned}$$

=

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{subject to} \quad & y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1, \text{ for all } n = 1, \dots, N, \end{aligned}$$

L12(2)

April 3, 2021 13 / 28

- Given the training dataset  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$  and a hyperplane  $\mathbf{w}^T \mathbf{x} + b = 0$ , what is the constraint that all data points are  $\frac{r}{\|\mathbf{w}\|}$ -away from the hyperplane?

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq \frac{r}{\|\mathbf{w}\|}$$

- **Formulation 1.** Note that  $r$  and  $\|\mathbf{w}\|$  are scaled together, so if we fix  $\|\mathbf{w}\| = 1$ , then

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq r.$$

And, maximize  $r$ .

- **Formulation 2.** If we fix  $r = 1$ , then

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1.$$

And, minimize  $\|\mathbf{w}\|$ 

L12(2)

April 3, 2021 14 / 28

- (1) Story and Separating Hyperplanes
- (2) Primal SVM: Hard SVM
- (3) **Primal SVM: Soft SVM**
- (4) Dual SVM
- (5) Kernels
- (6) Numerical Solution

L12(3)

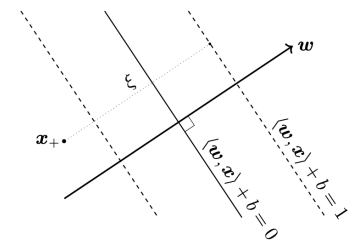
April 3, 2021 15 / 28

- Now we allow some classification errors, because it's not linearly separable.
- Introduce a slack variable that quantifies how much errors will be allowed in my optimization problem

- $\xi = (\xi_n : n = 1, \dots, N)$
- $\xi_n$ : slack for the  $n$ -th sample  $(\mathbf{x}_n, y_n)$

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n \\ \text{subject to} \quad & y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \\ & \xi_n \geq 0, \quad \text{for all } n \end{aligned}$$

- $C$ : Trade-off between width and slack



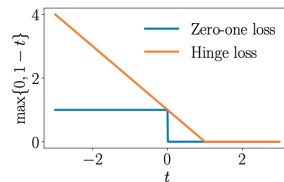
L12(3)

April 3, 2021 16 / 28

- From the perspective of empirical risk minimization
- Loss function design
  - zero-one loss**  $1(f(x_n) \neq y_n)$ : # of mismatches between the prediction and the label  
 $\implies$  combinatorial optimization (typically NP-hard)
  - hinge loss**

$$\ell(t) = \max(0, 1 - t), \text{ where } t = yf(\mathbf{x}) = y(\mathbf{w}^T \mathbf{x} + b)$$

- If  $\mathbf{x}$  is really at the correct side,  $t \geq 1$   
 $\rightarrow \ell(t) = 0$
- If  $\mathbf{x}$  is at the correct side, but too close to the boundary,  $0 < t < 1$   
 $\rightarrow 0 < \ell(t) = 1 - t < 1$
- If  $\mathbf{x}$  is at the wrong side,  $t < 0$   
 $\rightarrow 1 < \ell(t) = 1 - t$



L12(3)

April 3, 2021 17 / 28

$$\min_{\mathbf{w}, b} (\text{regularizer} + \text{loss}) = \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \max\{0, 1 - y(\mathbf{w}^T \mathbf{x} + b)\}$$

- $\frac{1}{2} \|\mathbf{w}\|^2$ : L2-regularizer (margin maximization = regularization)
- $C$ : regularization parameter, which moves from the regularization term to the loss term
- Why this loss function view = geometric view?  
 $\min_t \max(0, 1 - t) \iff \min_{\xi, t} \xi, \text{ subject to } \xi \geq 0, \xi \geq 1 - t$

L12(3)

April 3, 2021 18 / 28

- (1) Story and Separating Hyperplanes
- (2) Primal SVM: Hard SVM
- (3) Primal SVM: Soft SVM
- (4) **Dual SVM**
- (5) Kernels
- (6) Numerical Solution

L12(4)

April 3, 2021 19 / 28

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n$$

subject to  $y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 - \xi_n, \xi_n \geq 0, \text{ for all } n$

- The above primal problem is a convex optimization problem.
- Let's apply Lagrange multipliers, find another formulation, and see what other nice properties are shown **L7(2), L7(4)**
- Convert the problem into " $\leq$ " constraints, so as to apply **min-min-max** rule

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n, \text{ s.t. } -y_n(\mathbf{w}^T \mathbf{x}_n + b) \leq -1 + \xi_n, -\xi_n \leq 0, \text{ for all } n$$

L12(4)

April 3, 2021 20 / 28

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n, \text{ s.t. } -y_n(\mathbf{w}^T \mathbf{x}_n + b) \leq -1 + \xi_n, \quad -\xi_n \leq 0, \quad \text{for all } n$$

- Lagrangian with multipliers  $\alpha_n \geq 0$  and  $\gamma_n \geq 0$

$$\mathcal{L}(\mathbf{w}, b, \xi, \alpha, \gamma) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{n=1}^N \xi_n - \sum_{n=1}^N \alpha_n [y_n(\mathbf{w}^T \mathbf{x}_n + b) - 1 + \xi_n] - \sum_{n=1}^N \gamma_n \xi_n$$

- Dual function:  $\mathcal{D}(\alpha, \gamma) = \inf_{\mathbf{w}, b, \xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \gamma)$  for which the followings should be met:

$$(D1) \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w}^T - \sum_{n=1}^N \alpha_n y_n \mathbf{x}_n^T = 0, \quad (D2) \frac{\partial \mathcal{L}}{\partial b} = \sum_{n=1}^N \alpha_n y_n = 0, \quad (D3) \frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \gamma_n = 0$$

L12(4)

April 3, 2021 21 / 28

- Dual function  $\mathcal{D}(\alpha, \gamma) = \inf_{\mathbf{w}, b, \xi} \mathcal{L}(\mathbf{w}, b, \xi, \alpha, \gamma)$  with (D1) is given by:

$$\begin{aligned} \mathcal{D}(\alpha, \gamma) &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^N y_i \alpha_i \left\langle \sum_{j=1}^N y_j \alpha_j \mathbf{x}_j, \mathbf{x}_i \right\rangle - b \sum_{i=1}^N y_i \alpha_i \\ &\quad + \sum_{i=1}^N \alpha_i + \sum_{i=1}^N (C - \alpha_i - \gamma_i) \xi_i \end{aligned}$$

- From (D2) and (D3), the above is simplified into:

$$\mathcal{D}(\alpha, \gamma) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^N \alpha_i$$

- $\alpha_i, \gamma_i \geq 0$  and  $C - \alpha_i - \gamma_i = 0 \implies 0 \leq \alpha_i \leq C$

L12(4)

April 3, 2021 22 / 28

- (Lagrangian) Dual Problem: maximize  $\mathcal{D}(\alpha, \gamma)$

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N y_i y_j \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{i=1}^N \alpha_i \\ \text{subject to} \quad & \sum_{i=1}^N y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \quad \forall i = 1, \dots, N \end{aligned}$$

- Primal SVM: the number of parameters scales as the number of features ( $D$ )
- Dual SVM
  - the number of parameters scales as the number of training data ( $N$ )
  - only depends on the inner products of individual training data points  $\langle \mathbf{x}_i, \mathbf{x}_j \rangle \rightarrow$  allow the application of kernel

L12(4)

April 3, 2021 23 / 28

- (1) Story and Separating Hyperplanes
- (2) Primal SVM: Hard SVM
- (3) Primal SVM: Soft SVM
- (4) Dual SVM
- (5) Kernels
- (6) Numerical Solution

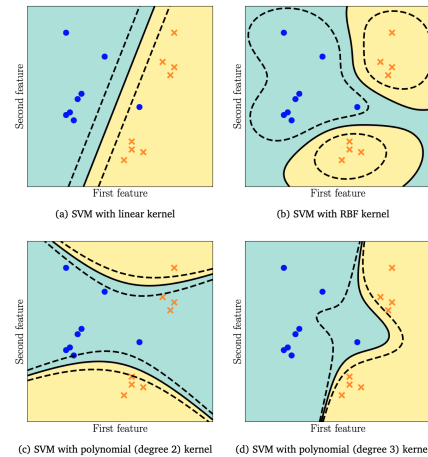
L12(5)

April 3, 2021 24 / 28

- Modularity: Using the feature transformation  $\phi(\mathbf{x})$ , dual SVMs can be modularized

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \implies \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

- Similarity function  $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ ,  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$
- Kernel matrix, Gram matrix: must be symmetric and positive semidefinite
- Examples: polynomial kernel, Gaussian radial basis function, rational quadratic kernel



L12(5)

April 3, 2021 25 / 28

L12(5)

April 3, 2021 26 / 28

Questions?

1)

L12(5)

April 3, 2021 27 / 28

L12(5)

April 3, 2021 28 / 28