Lecture 12: Classification with Support Vector Machines

Yi, Yung (이융)

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

April 2, 2021

Warm-Up

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

https://youtu.be/Lpr__X8zuE8

Roadmap

(1) Story and Separating Hyperplanes

(2) Primal SVM: Hard SVM

(3) Primal SVM: Soft SVM

(4) Dual SVM

(5) Kernels

(6) Numerical Solution

Roadmap

(1) Story and Separating Hyperplanes

(2) Primal SVM: Hard SVM

(3) Primal SVM: Soft SVM

(4) Dual SVM

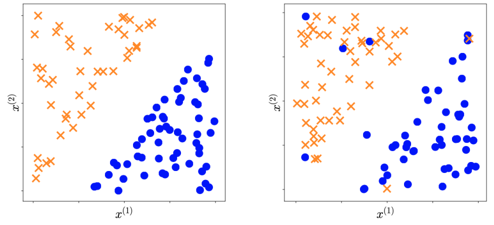
(5) Kernels

(6) Numerical Solution

Storyline

- (Binary) classification vs. regression
- A Classification predictor $f : \mathbb{R}^D \mapsto \{+1, -1\}$, where D is the dimension of features.
- Suppervised learning as in the regression with a given dataset $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$, where our task is to learn the model parameters which produces the smallest classification errors.
- SVM
  - Geometric way of thinking about supvervised 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 vs. Soft SVM



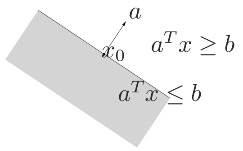
(a) Linearly separable data, with a large margin

(b) Non-linearly separable data

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

Separating Hyperplane

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

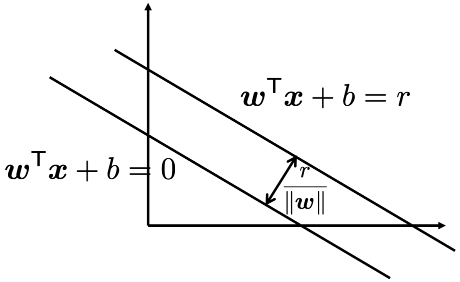


- In our problem, we consider the hyperplane $\mathbf{w}^\mathsf{T} \mathbf{x} + b = 0$, where $\mathbf{w}$ and b are the parameters of the model.
- Classification logic

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

Distance bertween Two Hyperplanes

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


$\boldsymbol{w}^\mathsf{T} \boldsymbol{x} + b = r$
$\boldsymbol{w}^\mathsf{T} \boldsymbol{x} + b = 0$
$\frac{r}{\|\boldsymbol{w}\|}$

[1] Shortested distance between two hyperplanes.

Roadmap

KAIST EE

(1) Story and Separating Hyperplanes

(2) Primal SVM: Hard SVM

(3) Primal SVM: Soft SVM

(4) Dual SVM

(5) Kernels

(6) Numerical Solution

Hard Support Vector Machine

KAIST EE

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

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

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

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

Hard SVM: Formulation 1

KAIST EE

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

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

Formulation 2 (1)

KAIST EE

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

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

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

Formulation 2 (2)

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

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

$=$

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

Understanding Formulation 2 Intuitively

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

$$y_n\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_n + b\right) \geq \frac{r}{\|\boldsymbol{w}\|}$$

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

$$y_n\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_n + b\right) \geq r.$$

  And, maximize r.

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

$$y_n\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_n + b\right) \geq 1.$$

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

Roadmap

(1) Story and Separating Hyperplanes

(2) Primal SVM: Hard SVM

(3) Primal SVM: Soft SVM

(4) Dual SVM

(5) Kernels

(6) Numerical Solution

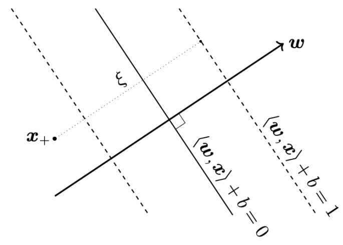
Soft SVM: Geometric View

- 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

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

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

- C: Trade-off between width and slack

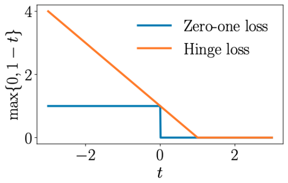

$\boldsymbol{w}$
ξ
$\boldsymbol{x}_+$
$\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b = 0$
$\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b = 1$

Soft SVM: Loss Function View (1)

- From the perspective of empirical risk minimizaiton
- 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

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

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



Soft SVM: Loss Function View (2)

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

- $\frac{1}{2}\|\boldsymbol{w}\|^2$: 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?

$$\min_{t} \max(0, 1 - t) \iff \min_{\xi, t} \xi, \ \text{subject to } \xi \geq 0, \ \xi \geq 1 - t$$

Roadmap

(1) Story and Separating Hyperplanes

(2) Primal SVM: Hard SVM

(3) Primal SVM: Soft SVM

(4) Dual SVM

(5) Kernels

(6) Numerical Solution

Dual SVM: Idea

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

- 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

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

Applying Lagrange Multipliers (1)

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

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

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

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

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

Applying Lagrange Multipliers (2)

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

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

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

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

- $\alpha_i, \gamma_i \ge 0$ and $C - \alpha_i - \gamma_i = 0 \implies 0 \le \alpha_i \le C$

Dual SVM

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

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j \langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \sum_{i=1}^{N}\alpha_i$$
$$\text{subject to} \quad \sum_{i=1}^{N} y_i\alpha_i = 0, \quad 0 \le \alpha_i \le C,\ \forall i = 1, \ldots, N$$

- 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 $\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle$ → allow the application of kernel

Roadmap

(1) Story and Separating Hyperplanes

(2) Primal SVM: Hard SVM

(3) Primal SVM: Soft SVM

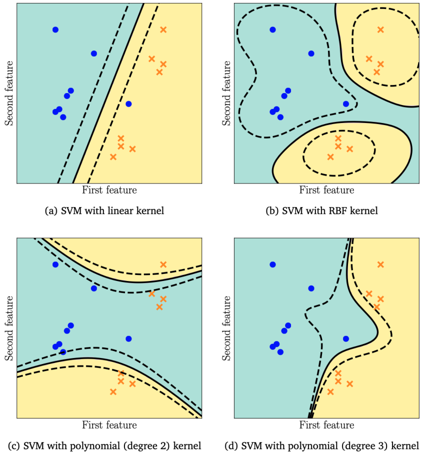
(4) Dual SVM

(5) Kernels

(6) Numerical Solution

Kernel

- Modularity: Using the feature transformation $\phi(\boldsymbol{x})$, dual SVMs can be modularized
$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \implies \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$$
- Similarity function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$, $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$
- Kernel matrix, Gram matrix: must be symmetric and positive semidifinite
- Examples: polynomial kernel, Gaussian radial basis function, rational quadratic kernel



(a) SVM with linear kernel

(b) SVM with RBF kernel

(c) SVM with polynomial (degree 2) kernel

(d) SVM with polynomial (degree 3) kernel

Numerical Solution

-

Questions?

Review Questions

1)