Introduction to Machine Learning

Lecture 12 Classification - Support Vector Machine

Hongteng Xu

中国人民大学 RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院 Gaoling School of Artificial Intelligence

Outline

Review

- Classification, definition, evaluation
- Linear classifiers (Naïve Bayes classifier, Linear discriminant analysis, Logistic regression)

Outline

Review

- Classification, definition, evaluation
- Linear classifiers (Naïve Bayes classifier, Linear discriminant analysis, Logistic regression)

Today

- Support-vector machine (SVM)

Outline

Review

- Classification, definition, evaluation
- Linear classifiers (Naïve Bayes classifier, Linear discriminant analysis, Logistic regression)

Today

- Support-vector machine (SVM)
- Take it easy. Most of the following content can be found at Wikipedia:)
- https://en.wikipedia.org/wiki/Support-vector_machine

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)}$$

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0)$$

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c$$

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c \tag{1}$$

(What is T for Naïve Bayes?)

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c \tag{1}$$

(What is T for Naïve Bayes?)

Logistic regression

$$p(y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta})}$$

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c \tag{1}$$

(What is T for Naïve Bayes?)

Logistic regression

$$p(y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta})} > 0.5$$

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c \tag{1}$$

(What is T for Naïve Bayes?)

Logistic regression

$$p(y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta})} > 0.5 \quad \Rightarrow \quad \boldsymbol{x}^T\boldsymbol{\beta} > 0$$

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c \tag{1}$$

(What is T for Naïve Bayes?)

Logistic regression

$$p(y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta})} > 0.5 \quad \Rightarrow \quad \boldsymbol{x}^T\boldsymbol{\beta} > 0 \tag{2}$$

Essentially, they aim at finding a boundary/hyperplane to separate the samples of different classes.

Revisit The Principle of Typical Classifiers

Linear determinant analysis

$$\log \frac{p(\boldsymbol{x}|y=1)}{p(\boldsymbol{x}|y=0)} > T(=0) \quad \Rightarrow \quad \boldsymbol{w}^T\boldsymbol{x} > c \tag{1}$$

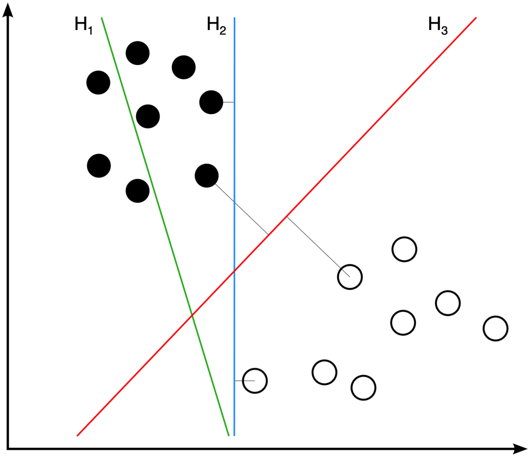
(What is T for Naïve Bayes?)

Logistic regression

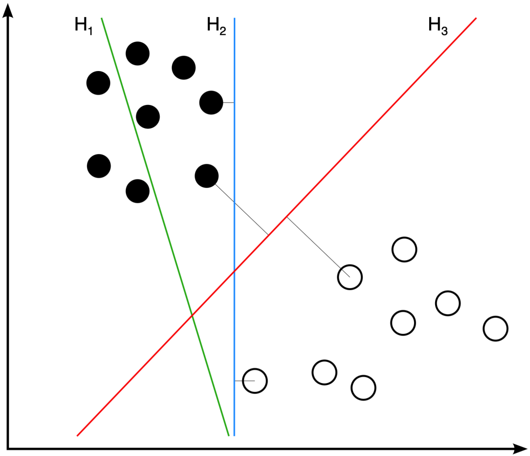
$$p(y=1|\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x}^T\boldsymbol{\beta})} > 0.5 \quad \Rightarrow \quad \boldsymbol{x}^T\boldsymbol{\beta} > 0 \tag{2}$$

Essentially, they aim at finding a boundary/hyperplane to separate the samples of different classes. (Recall the analytic algebra you learned in your high school.)

Ambiguity of Decision Boundary/Hyperplane



Ambiguity of Decision Boundary/Hyperplane



How to find the best decision boundary? What is the criterion?

Support-Vector Machine (SVM)

Motivation:

- Find a hyperplane $\boldsymbol{x}^T\boldsymbol{w} = b$ with maximum-margin/largest separation,

Support-Vector Machine (SVM)

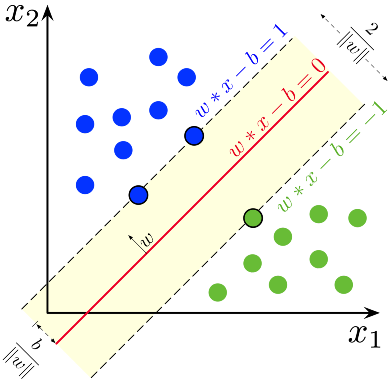
Motivation:

- Find a hyperplane $\boldsymbol{x}^T\boldsymbol{w} = b$ with maximum-margin/largest separation, such that the discriminative power is maximized, error risk is minimized.

Support-Vector Machine (SVM)

Motivation:

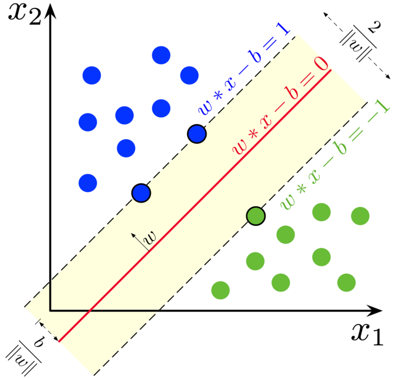
- Find a hyperplane $\boldsymbol{x}^T\boldsymbol{w} = b$ with maximum-margin/largest separation, such that the discriminative power is maximized, error risk is minimized.



Support-Vector Machine (SVM)

Principle:

- The desired hyperplane has the largest distance to the nearest training-data point of any class (so-called functional margin)



Implementation of SVM (Linearly Separable 2-Class)

- A set of training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, and each $y_n \in \{-1, 1\}$

Implementation of SVM (Linearly Separable 2-Class)

- A set of training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, and each $y_n \in \{-1, 1\}$
- Any hyperplane is the set of the points $\boldsymbol{x}$ satisfying $\boldsymbol{w}^T\boldsymbol{x} - b = 0$.

Implementation of SVM (Linearly Separable 2-Class)

- A set of training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$, and each $y_n \in \{-1, 1\}$
- Any hyperplane is the set of the points $\boldsymbol{x}$ satisfying $\boldsymbol{w}^T\boldsymbol{x} - b = 0$.
- $\boldsymbol{w}$: the normal vector
- $\frac{b}{\|w\|_2}$: the offset of the hyperplane shifting from the original along the normal vector. (Derive Them)

Implementation of SVM (Linearly Separable 2-Class)

- A set of training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, and each $y_n \in \{-1, 1\}$
- Any hyperplane is the set of the points $\boldsymbol{x}$ satisfying $\boldsymbol{w}^T\boldsymbol{x} - b = 0$.
- $\boldsymbol{w}$: the normal vector
- $\frac{b}{\|w\|_2}$: the offset of the hyperplane shifting from the original along the normal vector. (Derive Them)
- Shift the hyperplane along two opposite directions and define classification criteria (margins):

$$\boldsymbol{w}^T\boldsymbol{x} - b = 1 \quad \Rightarrow \quad \text{Any } \boldsymbol{x} \text{ on or above it is labeled by } 1$$

$$\boldsymbol{w}^T\boldsymbol{x} - b = -1 \quad \Rightarrow \quad \text{Any } \boldsymbol{x} \text{ on or above it is labeled by } -1.$$

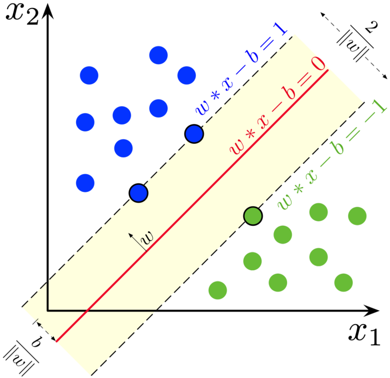
Implementation of SVM (Linearly Separable 2-Class)

- A set of training data $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, and each $y_n \in \{-1, 1\}$
- Any hyperplane is the set of the points $\boldsymbol{x}$ satisfying $\boldsymbol{w}^T\boldsymbol{x} - b = 0$.
- $\boldsymbol{w}$: the normal vector
- $\frac{b}{\|w\|_2}$: the offset of the hyperplane shifting from the original along the normal vector. (Derive Them)
- Shift the hyperplane along two opposite directions and define classification criteria (margins):

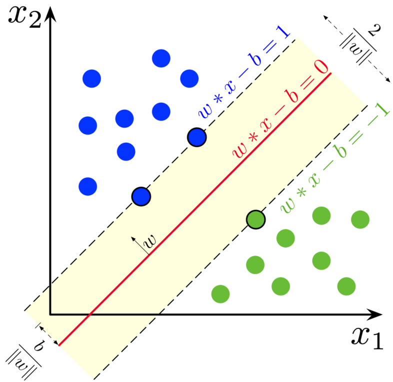
$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{x} - b = 1 \quad &\Rightarrow \quad \text{Any } \boldsymbol{x} \text{ on or above it is labeled by } 1 \\ \boldsymbol{w}^T\boldsymbol{x} - b = -1 \quad &\Rightarrow \quad \text{Any } \boldsymbol{x} \text{ on or above it is labeled by } -1. \end{aligned} \tag{3}$$

- $\frac{2}{\|\boldsymbol{w}\|_2}$: The distance between these two margins. (Derive It)

Implementation of SVM (Linearly Separable 2-Class)



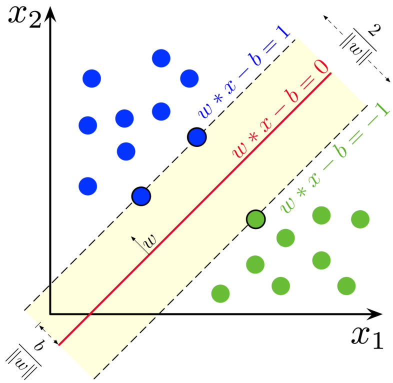
Implementation of SVM (Linearly Separable 2-Class)



For each $\boldsymbol{x}_n, y_n$, the data point $\boldsymbol{x}_n$ must lie on the correct side of the margin.

$$\boldsymbol{w}^T\boldsymbol{x}_n - b \begin{cases} \geq 1 & \text{if } y_n = 1, \\ \leq -1 & \text{if } y_n = -1. \end{cases}$$

Implementation of SVM (Linearly Separable 2-Class)



For each $\boldsymbol{x}_n, y_n$, the data point $\boldsymbol{x}_n$ must lie on the correct side of the margin.

$$\boldsymbol{w}^T\boldsymbol{x}_n - b \begin{cases} \geq 1 & \text{if } y_n = 1, \\ \leq -1 & \text{if } y_n = -1. \end{cases} \quad \Rightarrow \quad y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1, \ \forall n = 1, .., N. \tag{4}$$

Implementation of SVM (Linearly Separable 2-Class)

Principle:

- Maximize the distance between margins, i.e., max $\frac{2}{\|\boldsymbol{w}\|_2}$

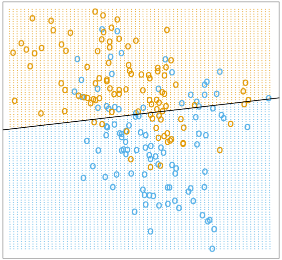
Implementation of SVM (Linearly Separable 2-Class)

Principle:

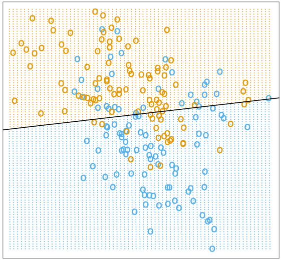
- Maximize the distance between margins, i.e., max $\frac{2}{\|\boldsymbol{w}\|_2}$
- Equivalently, the problem becomes

$$
\begin{aligned}
&\min_{w,b} \|\boldsymbol{w}\|_2 \\
&s.t.\ y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1,\ \forall\, n = 1, ..., N
\end{aligned}
\tag{5}
$$

How To Deal with Fuzzy Decision Boundary/Hyperplane?



How To Deal with Fuzzy Decision Boundary/Hyperplane?

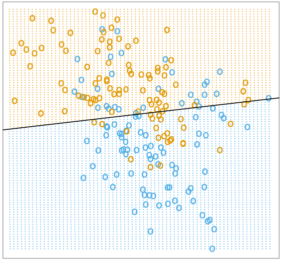


Soft-margin SVM: for noisy linearly-separable classes

$$\min_{w,b,\boldsymbol{\xi}} \lambda \|\boldsymbol{w}\|_2^2 + \frac{1}{N} \sum_{n=1}^{N} \xi_n$$

$$s.t.\ y_n(\boldsymbol{w}^T \boldsymbol{x}_n - b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \ \forall\, n = 1, \ldots, N$$

How To Deal with Fuzzy Decision Boundary/Hyperplane?



Soft-margin SVM: for noisy linearly-separable classes

$$\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n$$
$$s.t.\ y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \ \forall\, n = 1, ..., N \tag{6}$$

where $\xi_n = \max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}$, i.e., the smallest nonnegative number satisfying $y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n$.

Solve Soft-margin SVM (Primal Problem)

- The Primal Problem of Soft-margin SVM:

$$\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n$$

$$s.t.\ y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \ \forall\, \boldsymbol{n} = 1, ..., N$$

Solve Soft-margin SVM (Primal Problem)

- The Primal Problem of Soft-margin SVM:

$$\begin{aligned} &\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n \\ &s.t.\ y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \ \forall\, n = 1, ..., N \end{aligned} \tag{7}$$

- Plug the definition $\xi_n = \max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}$ into (7):

$$\min_{w,b}\ \underbrace{\lambda\|\boldsymbol{w}\|_2^2}_{\text{Tikhonov reg.}} + \frac{1}{N}\sum_{n=1}^{N} \underbrace{\max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}}_{\text{Hinge loss}}$$

Solve Soft-margin SVM (Primal Problem)

- The Primal Problem of Soft-margin SVM:

$$\begin{aligned} &\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n \\ &s.t.\ y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \ \forall\, n = 1, ..., N \end{aligned} \tag{7}$$

- Plug the definition $\xi_n = \max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}$ into (7):

$$\min_{w,b}\ \underbrace{\lambda\|\boldsymbol{w}\|_2^2}_{\text{Tikhonov reg.}} + \frac{1}{N}\sum_{n=1}^{N} \underbrace{\max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}}_{\text{Hinge loss}} \tag{8}$$

- The objective function is convex (but non-smooth) w.r.t. $\boldsymbol{w}$ and b, so GD or SGD can be applied.

Solve Soft-margin SVM (Primal Problem)

- The Primal Problem of Soft-margin SVM:

$$
\begin{aligned}
&\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n \\
&s.t.\ y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,\ \xi_n \geq 0,\ \ \forall\, n = 1, ..., N
\end{aligned}
\tag{7}
$$

- Plug the definition $\xi_n = \max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}$ into (7):

$$
\min_{w,b} \underbrace{\lambda\|\boldsymbol{w}\|_2^2}_{\text{Tikhonov reg.}} + \frac{1}{N}\sum_{n=1}^{N} \underbrace{\max\{0, 1 - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)\}}_{\text{Hinge loss}}
\tag{8}
$$

- The objective function is convex (but non-smooth) w.r.t. $\boldsymbol{w}$ and b, so GD or SGD can be applied.
- Derive the gradient of Hinge loss

Solve Soft-margin SVM (Dual Problem)

Because the Primal problem is convex optimization, we can solve it equivalently via solving its dual problem.

- Recall the Primal Problem of Soft-margin SVM:

$$\min_{w,b,\boldsymbol{\xi}} \lambda \|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n$$

$$s.t.\ \underbrace{y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n}_{\text{Constraint 1}},\ \underbrace{\xi_n \geq 0}_{\text{Constraint 2}},\ \forall\, n = 1, \ldots, N$$

Solve Soft-margin SVM (Dual Problem)

Because the Primal problem is convex optimization, we can solve it equivalently via solving its dual problem.

- Recall the Primal Problem of Soft-margin SVM:

$$
\begin{aligned}
&\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n \\
&s.t.\ \underbrace{y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,}_{\text{Constraint 1}}\ \ \underbrace{\xi_n \geq 0}_{\text{Constraint 2}},\ \forall\, n = 1, ..., N
\end{aligned}
\tag{9}
$$

- The Lagrangian dual of (9): Introduce dual variables $\{c_n\}_{n=1}^N$ for the first N constraints, and $\{\tau_n\}_{n=1}^N$ for the second N constraints

$$
\max_{\{c_n,\tau_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^{N} \left(\frac{1}{N}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) - \tau_n\xi_n \right) \right)
$$

$$
s.t.\ c_n \geq 0,\ \tau_n \geq 0,\ \forall n = 1, ..., N
$$

Solve Soft-margin SVM (Dual Problem)

Because the Primal problem is convex optimization, we can solve it equivalently via solving its dual problem.

- Recall the Primal Problem of Soft-margin SVM:

$$
\begin{aligned}
&\min_{w,b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{N}\sum_{n=1}^{N} \xi_n \\
&s.t.\ \underbrace{y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b) \geq 1 - \xi_n,}_{\text{Constraint 1}}\ \underbrace{\xi_n \geq 0}_{\text{Constraint 2}},\ \forall\, n = 1, ..., N
\end{aligned} \tag{9}
$$

- The Lagrangian dual of (9): Introduce dual variables $\{c_n\}_{n=1}^N$ for the first N constraints, and $\{\tau_n\}_{n=1}^N$ for the second N constraints

$$
\max_{\{c_n,\tau_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^{N} \left(\frac{1}{N}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) - \tau_n\xi_n \right) \right) \tag{10}
$$

$s.t.\ c_n \geq 0,\ \tau_n \geq 0,\ \forall n = 1, ..., N$

- Looks terrible... but simple actually;)

Solve Soft-margin SVM (Dual Problem)

Original dual problem:

$$\max_{\{c_n,\tau_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \sum\nolimits_{n=1}^N \left(\frac{1}{N}\xi_n + c_n(1-\xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) - \tau_n\xi_n \right) \right) \tag{11}$$

$$s.t.\ c_n \geq 0,\ \tau_n \geq 0,\ \forall n = 1, ..., N$$

- **Firstly, we can ignore $\{\tau_n\}_{n=1}^N$ because $\tau_n^* = 0$ (Why?)**

Solve Soft-margin SVM (Dual Problem)

Original dual problem:

$$\max_{\{c_n,\tau_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \lambda\|\boldsymbol{w}\|_2^2 + \sum\nolimits_{n=1}^N \left(\frac{1}{N}\xi_n + c_n(1-\xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) - \tau_n\xi_n \right) \right) \tag{11}$$

$$s.t.\ c_n \geq 0,\ \tau_n \geq 0,\ \forall n = 1, ..., N$$

- **Firstly, we can ignore $\{\tau_n\}_{n=1}^N$ because $\tau_n^* = 0$ (Why?)**
  - $\xi_n \geq 0$ and $\tau_n \geq 0$
  - Objective function: $\max_{\tau_n \geq 0} \min_{\xi_n \geq 0} -\tau_n\xi_n = 0$ when $\tau_n^* = 0$

Solve Soft-margin SVM (Dual Problem)

Original dual problem:

$$\max_{\{c_n,\tau_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \lambda \|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{N}\xi_n + {\color{red}c_n}(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) - {\color{blue}\tau_n}\xi_n \right) \right) \tag{11}$$

$$s.t.\ c_n \geq 0,\ \tau_n \geq 0,\ \forall n = 1, ..., N$$

- **Firstly, we can ignore $\{\tau_n\}_{n=1}^N$ because $\tau_n^* = 0$ (Why?)**
  - $\xi_n \geq 0$ and $\tau_n \geq 0$
  - Objective function: $\max_{\tau_n \geq 0} \min_{\xi_n \geq 0} -\tau_n\xi_n = 0$ when $\tau_n^* = 0$
- So, rewrite (11) as

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \lambda \|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{N}\xi_n + {\color{red}c_n}(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right) \right) \tag{12}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 1):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \lambda \|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{N} \xi_n + c_n (1 - \xi_n - y_n (\boldsymbol{w}^T \boldsymbol{x}_n - b)) \right) \right) \tag{13}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Secondly, the problem can be re-scale to**

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \frac{1}{2} \|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{2N\lambda} \xi_n + c_n (1 - \xi_n - y_n (\boldsymbol{w}^T \boldsymbol{x}_n - b)) \right) \right) \tag{14}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

(Why?)

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^{N} \left(\frac{1}{2N\lambda}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{2N\lambda}\xi_n + c_n(1-\xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w}^* = \sum_{n=1}^N c_n y_n \boldsymbol{x}_n \quad \text{Relation between opt. primal and dual}$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{2N\lambda}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w}^* = \sum_{n=1}^N c_n y_n \boldsymbol{x}_n \quad \text{Relation between opt. primal and dual}$$

$$\frac{\mathrm{d}L}{\mathrm{d}b} = 0$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\underbrace{\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{2N\lambda}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w}^* = \sum_{n=1}^N c_n y_n \boldsymbol{x}_n \quad \text{Relation between opt. primal and dual}$$

$$\frac{\mathrm{d}L}{\mathrm{d}b} = 0 \quad \Rightarrow \quad \sum_{n=1}^N c_n y_n = 0 \quad \text{Equality constraints of dual variables}$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum\nolimits_{n=1}^{N} \left(\frac{1}{2N\lambda}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w}^* = \sum\nolimits_{n=1}^{N} c_n y_n \boldsymbol{x}_n \quad \text{Relation between opt. primal and dual}$$

$$\frac{\mathrm{d}L}{\mathrm{d}b} = 0 \quad \Rightarrow \quad \sum\nolimits_{n=1}^{N} c_n y_n = 0 \quad \text{Equality constraints of dual variables}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\xi_n}$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{2N\lambda}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, \dots, N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w}^* = \sum_{n=1}^N c_n y_n \boldsymbol{x}_n \quad \text{Relation between opt. primal and dual}$$

$$\frac{\mathrm{d}L}{\mathrm{d}b} = 0 \quad \Rightarrow \quad \sum_{n=1}^N c_n y_n = 0 \quad \text{Equality constraints of dual variables}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\xi_n} = \frac{1}{2N\lambda} - c_n$$

Solve Soft-margin SVM (Dual Problem)

Modified dual problem (Version 2):

$$\max_{\{c_n\}_{n=1}^N} \left(\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \underbrace{\frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{n=1}^N \left(\frac{1}{2N\lambda}\xi_n + c_n(1 - \xi_n - y_n(\boldsymbol{w}^T\boldsymbol{x}_n - b)) \right)}_{L(\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{c})} \right) \tag{15}$$

$$s.t.\ c_n \geq 0,\ \forall n = 1, ..., N$$

- **Thirdly, consider the optimality condition of inner problem:**

$$\frac{\partial L}{\partial \boldsymbol{w}} = 0 \quad \Rightarrow \quad \boldsymbol{w}^* = \sum_{n=1}^N c_n y_n \boldsymbol{x}_n \quad \text{Relation between opt. primal and dual}$$

$$\frac{\mathrm{d}L}{\mathrm{d}b} = 0 \quad \Rightarrow \quad \sum_{n=1}^N c_n y_n = 0 \quad \text{Equality constraints of dual variables} \tag{16}$$

$$\frac{\mathrm{d}L}{\mathrm{d}\xi_n} = \frac{1}{2N\lambda} - c_n \geq 0 \quad \text{(Inequality constraints of dual variables, why?)}$$

Solve Soft-margin SVM (Dual Problem)

- **Finally, plugging the relation and the constraints into (15), we obtain the final dual problem:**

$$
\max_{\{c_n\}_{n=1}^N} \sum_{n=1}^{N} c_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n c_n (\boldsymbol{x}_n^T \boldsymbol{x}_m) y_m c_m
$$

$$
s.t. \sum_{n=1}^{N} c_n y_n = 0, \quad 0 \leq c_n \leq \frac{1}{2N\lambda}, \; \forall n = 1, ..., N.
$$

Solve Soft-margin SVM (Dual Problem)

- **Finally, plugging the relation and the constraints into (15), we obtain the final dual problem:**

$$
\begin{aligned}
&\max_{\{c_n\}_{n=1}^N} \sum_{n=1}^N c_n - \frac{1}{2}\sum_{n=1}^N \sum_{m=1}^N y_n c_n (\boldsymbol{x}_n^T \boldsymbol{x}_m) y_m c_m \\
&s.t. \ \sum_{n=1}^N c_n y_n = 0, \quad 0 \leq c_n \leq \frac{1}{2N\lambda}, \ \forall n = 1, ..., N.
\end{aligned}
\tag{17}
$$

- Derive it and write it in a matrix form.

Solve Soft-margin SVM (Dual Problem)

- **Finally, plugging the relation and the constraints into (15), we obtain the final dual problem:**

$$\begin{aligned}&\max_{\{c_n\}_{n=1}^N} \sum_{n=1}^N c_n - \frac{1}{2}\sum_{n=1}^N \sum_{m=1}^N y_n c_n (\boldsymbol{x}_n^T \boldsymbol{x}_m) y_m c_m \\ &s.t. \sum_{n=1}^N c_n y_n = 0, \quad 0 \le c_n \le \frac{1}{2N\lambda}, \ \forall n = 1, \ldots, N.\end{aligned} \tag{17}$$

- Derive it and write it in a matrix form.
- Why are the terms of ξ_n's ignored?

Solve Soft-margin SVM (Dual Problem)

- **Finally, plugging the relation and the constraints into (15), we obtain the final dual problem:**

$$
\begin{aligned}
&\max_{\{c_n\}_{n=1}^N} \sum_{n=1}^N c_n - \frac{1}{2}\sum_{n=1}^N \sum_{m=1}^N y_n c_n (\boldsymbol{x}_n^T \boldsymbol{x}_m) y_m c_m \\
&s.t. \sum_{n=1}^N c_n y_n = 0, \quad 0 \le c_n \le \frac{1}{2N\lambda}, \ \forall n = 1, \ldots, N.
\end{aligned} \tag{17}
$$

- Derive it and write it in a matrix form.
- Why are the terms of ξ_n's ignored?
- This problem can be solved by GD or Coordinate Gradient Descent.

Revisit Dual SVM as Kernel Methods

- Key relation:

$$\boldsymbol{w}^* = \sum_{n=1}^{N} c_n y_n \boldsymbol{x}_n$$

Revisit Dual SVM as Kernel Methods

- Key relation:

$$\boldsymbol{w}^* = \sum_{n=1}^{N} c_n y_n \boldsymbol{x}_n \tag{18}$$

- The dual SVM can be treated as applying a linear kernel function:

$$\begin{aligned} &\max_{\{c_n\}_{n=1}^N} \sum_{n=1}^{N} c_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n c_n \big(\underbrace{\boldsymbol{x}_n^T \boldsymbol{x}_m}_{K(\boldsymbol{x}_n, \boldsymbol{x}_m)} \big) y_m c_m \\ &s.t. \sum_{n=1}^{N} c_n y_n = 0, \quad 0 \le c_n \le \frac{1}{2N\lambda}, \ \forall n = 1, ..., N. \end{aligned} \tag{19}$$

Revisit Dual SVM as Kernel Methods

- Kernel SVM = Applying Linear Dual SVM in the feature space $\mathcal{F}$ defined by the kernel function:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{F}}, \quad \phi : \mathcal{X} \mapsto \mathcal{F}.$$

Revisit Dual SVM as Kernel Methods

- Kernel SVM = Applying Linear Dual SVM in the feature space $\mathcal{F}$ defined by the kernel function:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{F}}, \quad \phi : \mathcal{X} \mapsto \mathcal{F}. \tag{20}$$

- Accordingly, the key relation becomes

$$\boldsymbol{w}^* = \sum_{n=1}^{N} c_n y_n \phi(\boldsymbol{x}_n) \in \mathcal{F}$$

Revisit Dual SVM as Kernel Methods

- Kernel SVM = Applying Linear Dual SVM in the feature space $\mathcal{F}$ defined by the kernel function:

$$K(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle_{\mathcal{F}}, \quad \phi : \mathcal{X} \mapsto \mathcal{F}. \tag{20}$$

- Accordingly, the key relation becomes

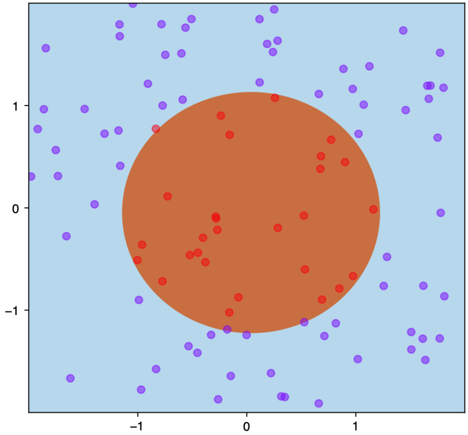
$$\boldsymbol{w}^* = \sum_{n=1}^{N} c_n y_n \phi(\boldsymbol{x}_n) \in \mathcal{F} \tag{21}$$

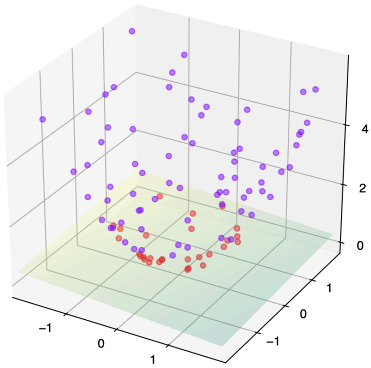
- Kernel SVM:

$$\begin{aligned} &\max_{\{c_n\}_{n=1}^N} \sum_{n=1}^{N} c_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} y_n c_n \underbrace{\langle \phi(\boldsymbol{x}_n), \phi(\boldsymbol{x}_m) \rangle_{\mathcal{F}}}_{K(\boldsymbol{x}_n, \boldsymbol{x}_m)} y_m c_m \\ &s.t. \sum_{n=1}^{N} c_n y_n = 0, \quad 0 \le c_n \le \frac{1}{2N\lambda}, \ \forall n = 1, \ldots, N. \end{aligned} \tag{22}$$

Similar to other kernel method, we don't have to find ϕ explicitly — just define K.

Revisit Dual SVM as Kernel Methods





Extensions of SVM

Multi-class SVM

- Like LDA, “one against the rest” or pairwise classification

Extensions of SVM

Multi-class SVM

- Like LDA, “one against the rest” or pairwise classification

Support-Vector Regression (SVR)

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|_2^2 \\ &s.t.\ \ |y_n - \boldsymbol{w}^T\boldsymbol{x}_n - b| \le \epsilon \end{aligned} \tag{23}$$

In Summary

- Support-vector machine (SVM)
- Kernelized SVM

Next...

- Information theory in Statistic ML
- Decision tree model

HW 4: DDL June 3, 2022

Python Programming

1. Lab # 9 (4 Pts)
2. Lab # 10 (4 Pts)

Questions for Tech Report (6 Pts, $\leq$ 3 Pages)

1. Given $\boldsymbol{X} \in \mathbb{R}^{N \times D}$ and $\boldsymbol{Y} \in \mathbb{R}^{N \times C}$, where each row of $\boldsymbol{X}$ is a D-dimensional feature, and each row of $\boldsymbol{Y}$ is a one-hot vector indicating one of the C classes. **Can we solve classification as regression?** e.g.,

$$\min_{\boldsymbol{W}} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{W}\|_F^2 \tag{24}$$

   If it does not work well in general, when it works? (Hint: Find the answer in ESL) (3 Pts)

2. Derive from (15) to (17) in details, and demonstrate that the terms related to ξ_n's are ignorable (Hint: consider the objective function in (15)). (3 Pts)