VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

Support Vector Machine

(SVM)

Nguyen Thanh Hy
Le Tin Nghia
Tran Nhu Cam Nguyen

December 8, 2023

Contents

1. Introduction
2. SVM

2.1 The linearly separable case
- Linear SVM
- Separating hyperplane
- Solve Problem
- Results
- Noise
- Penalty
- Solve soft-margin

2.2 Non-linear SVM
- Difficulty
- Kernel function

2.3 Summary

3. References
4. Demo

Introduction

- **Support Vector Machines (SVM)** was proposed by Vapnik and his colleagues in 1970s. Then it became famous and popular in 1990s.
- Originally, SVM is a method for linear classification. It finds a hyperplane (also called linear classifier) to separate the two classes of data.
- For non-linear classification for which no hyperplane separates well the data, kernel functions will be used.
  - Kernel functions play the role to transform the data into another space, in which the data is linearly separable.

Introduction

- Sometimes, we call linear SVM when no kernel function is used. (in fact, linear SVM uses a linear kernel)
- SVM has a strong theory that supports its performance.
- It can work well with very high dimensional problems.
- It is now one of the most popular and strong methods.
- For text categorization, linear SVM performs very well.

The linearly separable case

- Problem representation:
  - Training data D = $\{(x_1, y_1), (x_2, y_2), ..., (x_r, y_r)\}$ with r instances.
  - Each x_i is a vector in an n-dimensional space. Each dimension represents an attribute.
  - Bold characters denote vectors.
  - y_i is a class label in $\{-1; 1\}$. ‘1’ is positive class, ‘-1’ is negative class.
- **Linear separability assumption**: there exists a hyperplane (of linear form) that well separates the two classes

Linear SVM

UIT-HCM

- SVM finds a hyperplane of the form:

$$f(x) = \langle w \cdot x \rangle + b \tag{1}$$

  - w is the weight vector; b is a real number (bias).
  - $\langle w \cdot x \rangle$ and $\langle w, x \rangle$ denote the inner product of two vectors.

- Such that for each x_i:

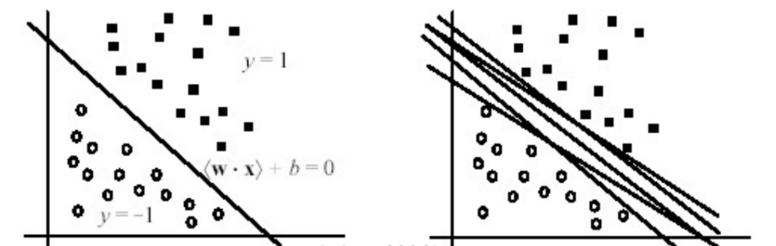
$$y_i = \begin{cases} 1 & \text{if } \langle w \cdot x_i \rangle + b \geq 0, \\ -1 & \text{if } \langle w \cdot x_i \rangle + b \leq 0. \end{cases} \tag{2}$$

Separating hyperplane

- The hyperplane (H0) which separates the positive from negative class is of the form:

$$\langle w \cdot x_i \rangle + b = 0 \tag{3}$$

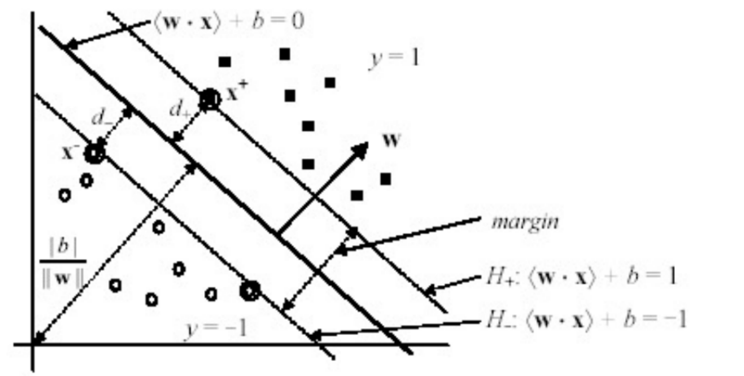
- It is also known as the decision boundary/surface.
- But there might be infinitely many separating hyperplanes. Which one should we choose?



[Liu, 2006]

Hyperplane with max margin

- SVM selects the hyperplane with max margin
- It is proven that the max-margin hyperplane has minimal errors among all possible hyperplanes.



[Liu, 2006]

Marginal hyperplanes

- Assume that the two classes in our data can be separated clearly by a hyperplane.
- Denote (x^+,1) in possitive class and (x^-,-1) in negative class which are closest to the separating hyperplane H0 ($\langle w \cdot x \rangle$ + b = 0)
- We define two parallel marginal hyperplanes as follows:
  - H_+ crosses x^+ and is parallel with H_0: ($\langle w \cdot x \rangle$ + b = 1)
  - H_+ crosses x^+ and is parallel with H_0: ($\langle w \cdot x \rangle$ + b = -1)
  - No data point lies between these two marginal hyperplanes, and satisfying:
    $\langle w \cdot x_i \rangle + b \geq 1, \quad \text{if } y_i = 1$
    $\langle w \cdot x_i \rangle + b \leq -1, \quad \text{if } y_i = -1$

The margin

- Margin is defined as the distance between the two marginal hyperplanes.
  - Denote d_+ the distance from H_0 to H_+.
  - Denote d_- the distance from H_0 to H_-.
  - $(d_+ + d_-)$ is the margin.
- As a result the margin is:

$$\text{Margin} = d_+ + d_- = \frac{2}{||w||} \tag{4}$$

- This learning principle can be formulated as the following quadratic optimization problem:
  - Find w and b that maximize "Margin".
  - And satisfy the below conditions for any training data x_i:

    $\langle w \cdot x_i \rangle + b \geq 1, \quad \text{if } y_i = 1$
    $\langle w \cdot x_i \rangle + b \leq -1, \quad \text{if } y_i = -1$

The margin

- Learning SVM is equivalent to the following minimization problem:
  - Minimize:
$$\frac{\langle w \cdot w \rangle}{2} \tag{5}$$
  - Conditioned on:
$$y_i(\langle w \cdot x_i \rangle + b) \geq 1 \quad \forall i = 1...r \tag{6}$$
- This is a constrained optimization problem.
- The Lagrange function for problem [5] is:

$$L(w, b, \alpha) = \frac{1}{2}\langle w \cdot w \rangle - \sum_{i=1}^{r} [\alpha_i y_i (\langle w \cdot x_i \rangle + b) - 1] \tag{7}$$

Solve Problem

- Solving [5] is equivalent to the following minimax problem:

$$\arg\min_{w,b}\max_{\alpha\geq 0} L(w,b,\alpha) \tag{8}$$

- The primal problem [8] can be derived by solving:

$$\arg\max_{\alpha\geq 0} L(w,b,\alpha) \tag{9}$$

$$\arg\min_{w,b} L(w,b,\alpha) \tag{10}$$

Solve Problem

- It is known that the optimal solution to [5] will satisfy some conditions which is called the Karush-Kuhn-Tucker(KKT) conditions.

$$\alpha_i \geq 0 \tag{11}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{r} \alpha_i y_i x_i = 0 \tag{12}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{r} \alpha_i y_i = 0 \tag{13}$$

$$y_i(\langle w \cdot x_i \rangle + b) \geq 1 \quad \forall i = 1...r \tag{14}$$

$$\alpha_i(y_i(\langle w \cdot x_i \rangle + b) - 1) = 0 \tag{15}$$

Results

- The last equation [15] comes from a nice result from the duality theory.
  - For any $\alpha_i > 0$, point x_i is on a boundary hyperplane (H_+ or H_-).
  - Such a boundary point is named as a **support vector**.
  - A non-support vector will correspond to $\alpha_i = 0$.
- After we solve the original problem or the dual problem, we can derive the function:

$$f(x) = \langle w^* \cdot x \rangle + b^* = \sum_{x_i} \alpha_i y_i \langle x_i \cdot x \rangle + b^* = 0 \tag{16}$$

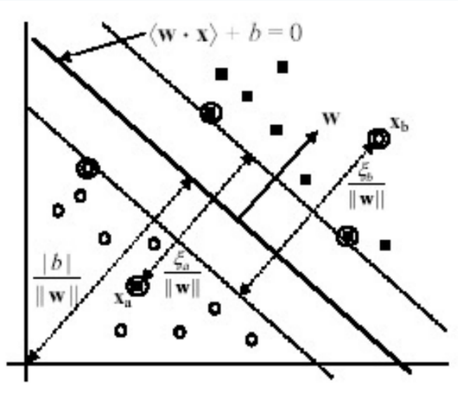
- For a new instance z, we compute:

$$sign(\langle w^* \cdot z \rangle + b^*) = sign(\sum_{x_i} \alpha_i y_i \langle x_i \cdot z \rangle + b^*) \tag{17}$$

  - If the result is 1, z will be assigned to the positive class; otherwise z will be assigned to the negative class.

Noise

- What if the two classes are not linearly separable?
- Noisy points x_a and x_b are mis-placed.



Noise

- To work with noises/errors, we need to relax the constraints about margin by using some slack variables ξ_i (*geq* 0):
  $\langle w \cdot x_i \rangle + b \geq 1 - \xi_i, \quad \text{if } y_i = 1$
  $\langle w \cdot x_i \rangle + b \leq -1 + \xi_i, \quad \text{if } y_i = -1$
  - For a noisy/erronous point x_i, we have: x_i >1
  - Otherwise ξ_i = 0.
- Therefore, we have the following constraints for the cases of linear inseparability:
  $y_i(\langle w \cdot x_i \rangle + b) \geq 1 - \xi_i \quad \forall i = 1...r$
  $\xi_i \geq 0 \quad \forall i = 1...r$

Penalty

- A penalty term will be used so that learning is to minimize:

$$\frac{\langle w \cdot w \rangle}{2} + C \sum_{i=1}^{r} \xi_i^k \quad (18)$$

  - Where C (>0) is the penalty constant.
  - The greater C, the heavier the penalty on noises/errors.
- k = 1 is often used in practice, due to simplicity for solving the optimization problem.
- Conditioned on:
  $y_i(\langle w \cdot x_i \rangle + b) \geq 1 - \xi_i \quad \forall i = 1...r$
  $\xi_i \geq 0 \quad \forall i = 1...r$
- This problem is called **Soft-margin SVM**.

Solve soft-margin

- The same as before, we use KKT for this problem

$$\alpha_i \geq 0 \tag{19}$$

$$\xi_i \geq 0 \tag{20}$$

$$\mu_i \geq 0 \tag{21}$$

$$y_i(\langle w \cdot x_i \rangle + b) - 1 + \xi_i \geq 0 \quad \forall i = 1 \ldots r \tag{22}$$

$$\alpha_i(y_i(\langle w \cdot x_i \rangle + b) - 1 + \xi_i) = 0 \tag{23}$$

$$\mu_i \xi_i = 0 \tag{24}$$

Solve soft-margin

- From [19] to [24] we conclude that:

$$\begin{array}{|llll|}
\hline
\text{If} \quad \alpha_{\mathrm{i}} = 0 & \text{then} \quad y_i(\langle \mathbf{w} \cdot \mathbf{x}_{\mathbf{i}} \rangle + b) \geq 1, & \text{and} \quad \xi_{\mathrm{i}} = 0 \\
\text{If} \quad 0 < \alpha_{\mathrm{i}} < C & \text{then} \quad y_i(\langle \mathbf{w} \cdot \mathbf{x}_{\mathbf{i}} \rangle + b) = 1, & \text{and} \quad \xi_{\mathrm{i}} = 0 \\
\text{If} \quad \alpha_{\mathrm{i}} = C & \text{then} \quad y_i(\langle \mathbf{w} \cdot \mathbf{x}_{\mathbf{i}} \rangle + b) < 1, & \text{and} \quad \xi_{\mathrm{i}} > 0 \\
\hline
\end{array}$$

- Points making $\alpha_i \neq 0$ we merge to SV (Support vector group).
- The classifier can be write by linear combination SV.
- Hence the optimal classifier is a very sparse combination of the training data.

Solve soft-margin

- Classify function:

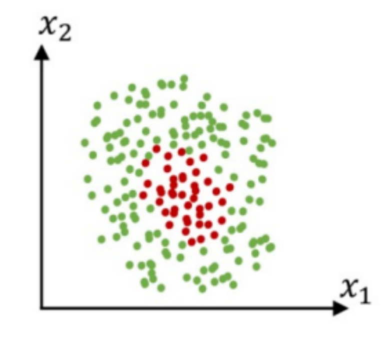
$$f(x) = \langle w^* \cdot x \rangle + b^* = \sum_{x_i \in SV} \alpha_i y_i \langle x_i \cdot x \rangle + b^* = 0 \tag{25}$$

- For a new instance z, we compute:

$$sign(\langle w^* \cdot z \rangle + b^*) = sign(\sum_{x_i \in SV} \alpha_i y_i \langle x_i \cdot z \rangle + b^*) \tag{26}$$

- Note: it is important to choose a good value of C, since it significantly affects performance of SVM.
    - We often use a validation set to choose a value for C.

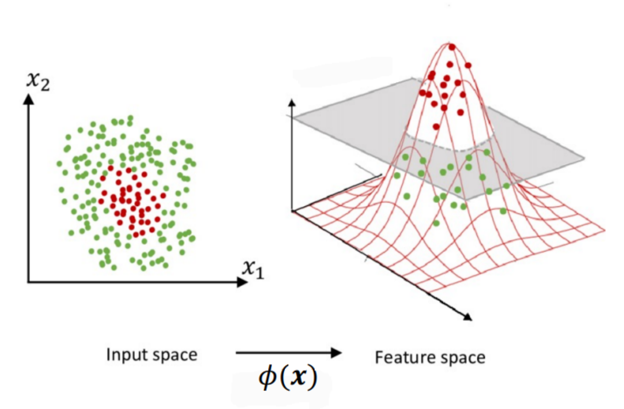
Non-linear SVM



- **Idea of Non-linear SVM:**
  - **Step 1**: transform the input into another space, which often has higher dimensions, so that the projection of data is linearly separable.
  - **Step 2**: use linear SVM in the new space

Non-linear SVM

- **Input space**: initial representation of data
- **Feature space**: the new space after the transformation



Difficulty

- How to find the mapping?
  - An intractable problem
- The curse of dimensionality
  - As the dimensionality increases, the volume of the space increases so fast that the available data become sparse.
  - This sparsity is problematic.
  - Increasing the dimensionality will require significantly more training data.

Kernel function

- An explicit form of a transformation is not necessary
- Both require only the inner product $\langle \phi(x), \phi(z) \rangle$ and we call that is kernel function.
- Polynomial: K(x, z) = $\langle x, z \rangle^d$
- Example: We choose d = 2, x = (x_1, x_2), z = (z_1, z_2).

$$\begin{aligned}
\langle x, z \rangle^2 &= (x_1 z_1 + x_2 z_2)^2 \\
&= x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\
&= \langle (x_1^2, x_2^2, \sqrt{2} x_1 x_2), (z_1^2, z_2^2, \sqrt{2} z_1 z_2) \rangle \\
&= \langle \phi(x), \phi(z) \rangle = K(x, z)
\end{aligned}$$

- Where $\phi(x) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$.

Kernel functions: popular choices

- Polynomial:

$$K(x,z) = (\langle x \cdot z \rangle + \theta)^d \quad \theta \in \mathbb{R}, d \in \mathbb{N} \tag{27}$$

- Gaussian radial basis function (RBF):

$$K(x,z) = e^{-\frac{||x-z||^2}{2\sigma}} \quad \sigma > 0 \tag{28}$$

- Sigmoid:

$$K(x,z) = tanh(\beta\langle x \cdot z \rangle - \lambda) = \frac{1}{e^{-\beta\langle x \cdot z \rangle - \lambda}} \quad \beta, \lambda \in \mathbb{R} \tag{29}$$

- What conditions ensure a kernel function?
  - **Mercer's theorem**
  - Instead we find kernel function, we find **kernel matrix** (only applicable in finite training data sets).

Summary

- SVM works with real-value attributes
    - Any nominal attribute need to be transformed into a real one.
- The learning formulation of SVM focuses on 2 classes
    - How about a classification problem with > 2 classes?
    - One-vs-the-rest, one-vs-one: a multiclass problem can be solved by reducing to many different problems with 2 classes
- The decision function is simple, but may be hard to interpret
    - It is more serious if we use some kernel functions

References

- B. Liu. Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data. Springer, 2006.
- C. J. C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 2(2): 121–167, 1998.
- Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20.3 (1995): 273–297.
- Than Quang Khoat (HUST). Machine Learning and Data Mining Course 2021.

Demo

- Colaboratory (Click here to be redirected to our source code on Google Colaboratory.)

VIET NAM NATIONAL UNIVERSITY HCMC
UNIVERSITY OF INFORMATION TECHNOLOGY
VNUHCM - UIT

Thank you for your attention

Nguyen Thanh Hy
Le Tin Nghia
Tran Nhu Cam Nguyen

December 8, 2023