Support Vector Machine

Hung Le

University of Victoria

February 12, 2019

Binary Classification

You are given a set of n data points $\mathcal{D} = \{(\mathbf{x}_1, y_1), \mathbf{x}_2, y_y), \ldots, (\mathbf{x}_n, y_n)\}$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Find a classifier $f(.) : \mathbb{R}^d \rightarrow \{-1, 1\}$ such that:

$$f(\mathbf{x}_i) = \begin{cases} 1 & \text{if } y_i = 1 \\ -1 & \text{if } y_i = -1 \end{cases}$$

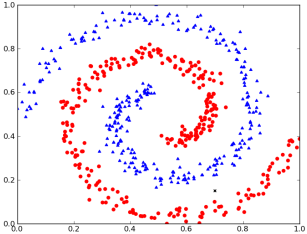


Figure: Sprial data[1]

Applications

- Spam email classification:
  - Each data point is $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ is a vector representation of i-th email, and $y_i = 1/-1$ indicates the email is spam/non-spam.
- Testing disease: determine a person as a certain disease or not.
- Weather forecasting: tomorrow is rainy or not.

Support Vector Machine

You are given a set of n data points $\mathcal{D} = \{(\mathbf{x}_1, y_1), \mathbf{x}_2, y_y), \ldots, (\mathbf{x}_n, y_n)\}$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Find a separating hyperplane $w^T\mathbf{x} + b = 0$ such that:

- $w^T\mathbf{x}_i + b > 0$ if $y_i = 1$
- $w^T\mathbf{x}_i + b < 0$ if $y_i = -1$

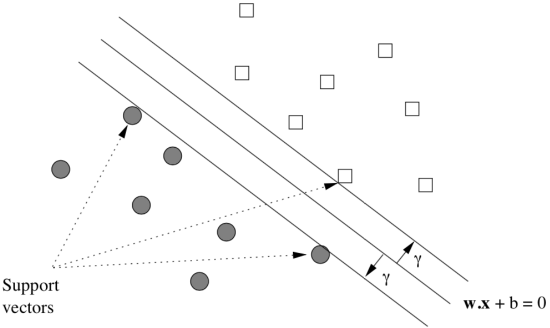
Support Vector Machine

You are given a set of n data points $\mathcal{D} = \{(\mathbf{x}_1, y_1), \mathbf{x}_2, y_y), \ldots, (\mathbf{x}_n, y_n)\}$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Find a separating hyperplane $w^T\mathbf{x} + b = 0$ such that:

- $w^T\mathbf{x}_i + b > 0$ if $y_i = 1$
- $w^T\mathbf{x}_i + b < 0$ if $y_i = -1$

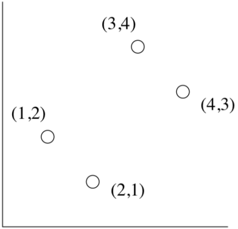
We assume that our data is linearly separable, i.e, there exists such a separating hyperplane.

Support Vector Machine - An Example

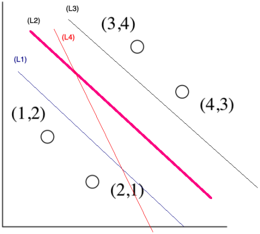


Support Vector Machine - A Toy Example

Given four points $(\begin{bmatrix}1\\2\end{bmatrix}, -1), (\begin{bmatrix}2\\1\end{bmatrix}, -1), (\begin{bmatrix}3\\4\end{bmatrix}, 1), (\begin{bmatrix}4\\3\end{bmatrix}, 1)$. Find a separating line $w_1 \cdot x_1 + w_2 \cdot x_2 + b = 0$ for these points.



Support Vector Machine - A Toy Example



There are several possible lines:

$$\begin{aligned} &(L_1): x_1 + x_2 - 4 = 0 \qquad && (L_2): x_1 + x_2 - 5 = 0 \\ &(L_3): x_1 + x_2 - 6 = 0 \qquad && (L_4): x_1 + 2x_2 - 6 = 0 \end{aligned} \tag{1}$$

Which line should we choose? In theory, any line is acceptable.

SVM separating principle

Choose a line that that maximizes the margin of the point set to the separating line.

SVM separating principle

Choose a line that that maximizes the margin of the point set to the separating line.

Margin of a separating line (L) w.r.t the point set $\mathcal{D}$ is the minimum distance of the point set to the line:

$$\gamma(L) = \min_{(\mathbf{x}_i, y_i) \in \mathcal{D}} d(\mathbf{x}_i, L) \tag{2}$$

SVM separating principle

Choose a line that that maximizes the margin of the point set to the separating line.

Margin of a separating line (L) w.r.t the point set $\mathcal{D}$ is the minimum distance of the point set to the line:

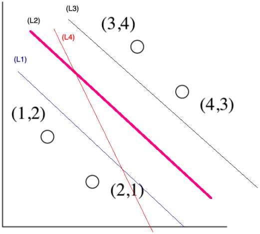
$$\gamma(L) = \min_{(\mathbf{x}_i, y_i) \in \mathcal{D}} d(\mathbf{x}_i, L) \tag{2}$$

Recall, distance from a point $\mathbf{x}_0 \in \mathrm{R}^d$ to the line $(L) : \mathbf{w}^T\mathbf{x} + b = 0$ is:

$$d(\mathbf{x}_0, L) = \frac{|\mathbf{w}^T\mathbf{x}_0 + b|}{||\mathbf{w}||_2} \tag{3}$$

where $||\mathbf{w}||_2 = \sqrt{\sum_{i=1}^{d} w[i]^2}$

Back to our Toy Example



$$(L_1): x_1 + x_2 - 4 = 0 \qquad (L_2): x_1 + x_2 - 5 = 0$$
$$(L_3): x_1 + x_2 - 6 = 0 \qquad (L_4): x_1 + 2x_2 - 6 = 0 \tag{4}$$

SVM will choose $(L2)$ with $\gamma(L_2) = \sqrt{2}$ (see the board calculation)

Support Vector Machine

You are given a set of n data points $\mathcal{D} = \{(\mathbf{x}_1, y_1), \mathbf{x}_2, y_y), \ldots, (\mathbf{x}_n, y_n)\}$ where each $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Find a separating hyperplane $(L) : w^T\mathbf{x} + b = 0$ such that:

- $w^T\mathbf{x}_i + b > 0$ if $y_i = 1$
- $w^T\mathbf{x}_i + b < 0$ if $y_i = -1$

and the margin $\gamma(L)$ is maximum among all possible separating hyperplanes.

Points $\mathbf{x}_j$ that have $d(L, \mathbf{x}_i) = \gamma(L)$ are called support vectors.

Support Vector Machine

The problem is equivalent to:

Find $\mathbf{w}, b$ that:

$$\text{maximize}(\min_i \frac{|\mathbf{w}^T\mathbf{x}_i + b|}{||w||_2}) \tag{5}$$

Support Vector Machine

The problem is equivalent to:

Find $\mathbf{w}, b$ that:

$$\text{maximize}(\min_i \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{||w||_2}) \tag{5}$$

Observation

If $(\mathbf{w}, b)$ defines a valid SVM hyperplane, then $(c \cdot \mathbf{w}, c \cdot b)$ also defines a valid SVM hyperplane.

Support Vector Machine

The problem is equivalent to:

Find $\mathbf{w}, b$ that:

$$\text{maximize}(\min_i \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{||w||_2}) \tag{5}$$

Observation

If $(\mathbf{w}, b)$ defines a valid SVM hyperplane, then $(c \cdot \mathbf{w}, c \cdot b)$ also defines a valid SVM hyperplane.

Thus, we can assume that:

- $\mathbf{w}^T \mathbf{x}_j + b = 1$ for all support vectors $\mathbf{x}_j$ of 1-class.
- $\mathbf{w}^T \mathbf{x}_j + b = -1$ for all support vectors $\mathbf{x}_j$ of (-1)-class.

Support Vector Machine

The problem becomes (see the board calculation):

Find $\mathbf{w}, b$ that :

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}||\mathbf{w}||_2^2 \\ \text{subject to} \quad & y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \forall i \end{aligned} \tag{6}$$

Regularization Variant of SVM

Transform the constrained optimization problem from SVM to:

Find $\mathbf{w}, b$ that :

$$\text{minimize} \qquad f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||_2^2 + C(\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))) \quad (7)$$

where C is a chosen positive number.

Regularization Variant of SVM

Transform the constrained optimization problem from SVM to:

Find $\mathbf{w}, b$ that :

$$\text{minimize} \qquad f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||_2^2 + C(\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))) \quad (7)$$

where C is a chosen positive number.

- When C is sufficiently big, we force the optimization algorithm returning $(\mathbf{w}, b)$ such that $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$ for all i. This is only possible when the data is linearly separable.

Regularization Variant of SVM

Transform the constrained optimization problem from SVM to:

Find $\mathbf{w}, b$ that :

$$\text{minimize} \qquad f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||_2^2 + C(\sum_{i=1}^{n} \max(0, 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))) \quad (7)$$

where C is a chosen positive number.

- When C is sufficiently big, we force the optimization algorithm returning $(\mathbf{w}, b)$ such that $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$ for all i. This is only possible when the data is linearly separable.
- When C is chosen appropriately, the optimization problem 7 has a regularizing effect.
  - We accept mis-classified points, but most other points are far away from the hyperplane.
  - The problem is well-defined even if the data is NOT linearly separable.

C is called the regularization parameter.

Optimization by SGD

Let

$$L_i(\mathbf{w}, b) = \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)) \tag{8}$$

Optimization by SGD

Let

$$L_i(\mathbf{w}, b) = \max(0, 1 - y_i(\mathbf{w}^T \mathbf{x}_i + b)) \tag{8}$$

$L_i(\mathbf{w}, b)$ is called a *hinge function* and its value is called a hinge loss.

Optimization by SGD

Let

$$L_i(\mathbf{w}, b) = \max(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)) \tag{8}$$

$L_i(\mathbf{w}, b)$ is called a *hinge function* and its value is called a hinge loss.
We have:

$$\frac{\partial L_i(\mathbf{w}, b)}{\partial w[j]} = \begin{cases} -y_i x_i[j] & \text{if } y_i(\mathbf{w}^T x + b) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

and

$$\frac{\partial L_i(\mathbf{w}, b)}{\partial b} = \begin{cases} -y_i & \text{if } y_i(\mathbf{w}^T x + b) < 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Optimization by SGD

Since:

$$f(\mathbf{w}, b) = \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{n} L_i(\mathbf{w}) \tag{11}$$

we have:

$$\frac{\partial f(\mathbf{w}, b)}{\partial w[j]} = w[j] + C\sum_{i=1}^{n} \frac{\partial L_i(\mathbf{w}, b)}{\partial w[j]} \tag{12}$$

and

$$\frac{\partial f(\mathbf{w}, b)}{\partial b} = C\sum_{i=1}^{n} \frac{\partial L_i(\mathbf{w}, b)}{\partial b} \tag{13}$$