CSC 4020 Fundamentals of Machine Learning: Support Vector Machine I

Baoyuan Wu
School of Data Science, CUHK-SZ

large

margin

February 24, 2021

Outline

Classification

Binary classification:

Classification

Binary classification:

- Given training data set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, and $\boldsymbol{x}_i \in \mathbb{R}^n$, $y_i \in \{-1, +1\}$

Classification

Binary classification:

- Given training data set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, and $\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$
- We adopt the sign hypothesis function $y = \text{sign}(f_{\boldsymbol{w}}(\boldsymbol{x})) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x})$

Classification

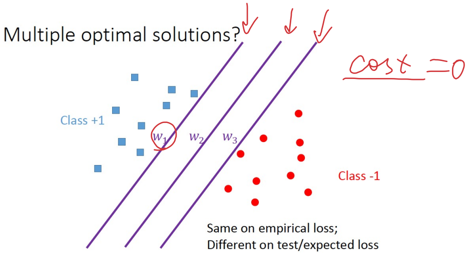
Binary classification:

- Given training data set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, and $\boldsymbol{x}_i \in \mathbb{R}^n, y_i \in \{-1, +1\}$
- We adopt the sign hypothesis function $y = \text{sign}(f_{\boldsymbol{w}}(\boldsymbol{x})) = \text{sign}(\boldsymbol{w}^\top \boldsymbol{x})$
- Then, we require that
  - If $y_i = +1$, then $\boldsymbol{w}^\top \boldsymbol{x} > 0$
  - If $y_i = -1$, then $\boldsymbol{w}^\top \boldsymbol{x} < 0$

$\text{sign}(w^T x) = 1$

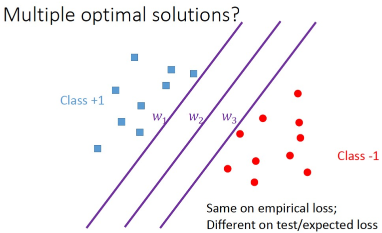
-1

Classification



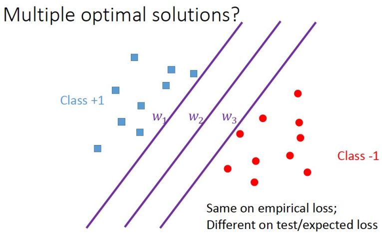
- There could be multiple decision boundaries to perfectly separate the above data.

Classification



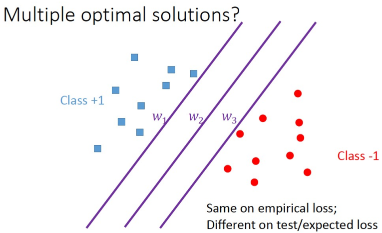
- There could be multiple decision boundaries to perfectly separate the above data. Why?

Classification



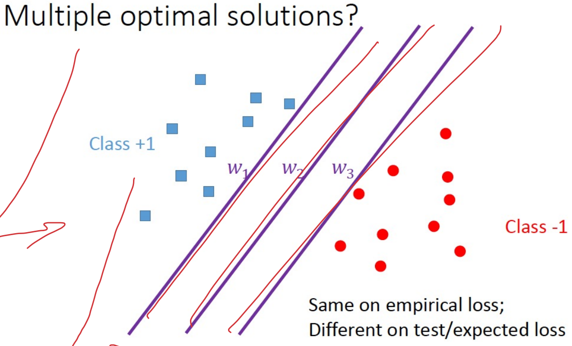
- There could be multiple decision boundaries to perfectly separate the above data. Why?
- For standard logistic regression, the objective function (*i.e.*, cross entropy loss) is convex, rather than strongly/strictly convex. Consequently, there are multiple values of parameters that can perfectly fit the training data.

Classification



- There could be multiple decision boundaries to perfectly separate the above data. Why?
- For standard logistic regression, the objective function (*i.e.*, cross entropy loss) is convex, rather than strongly/strictly convex. Consequently, there are multiple values of parameters that can perfectly fit the training data.
- For regularized logistic regression, the objective function (*i.e.*, cross entropy loss + $\lambda \cdot \ell_2$ regularization) is strictly convex, which has the unique optimal solution. However, it depends on the trade-off hyper-parameter λ. For sure you can use cross-validation to use a suitable λ, but is there any more elegant approach?

Classification

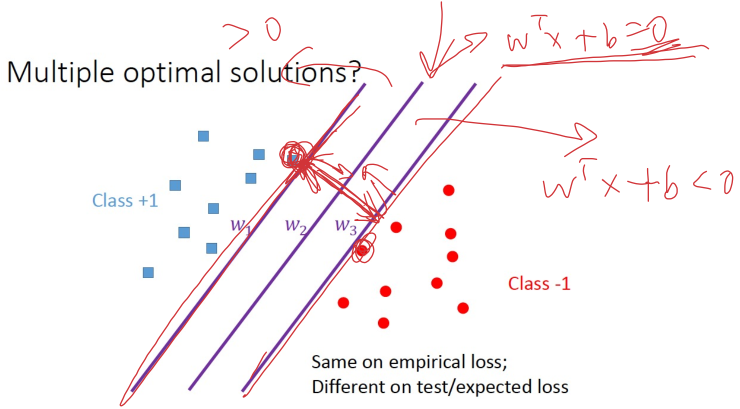


- Just following your intuition, which decision boundary do you prefer?

$$J(\theta) = \left[CE(\theta) + \lambda \|\theta\|^2 \right]$$

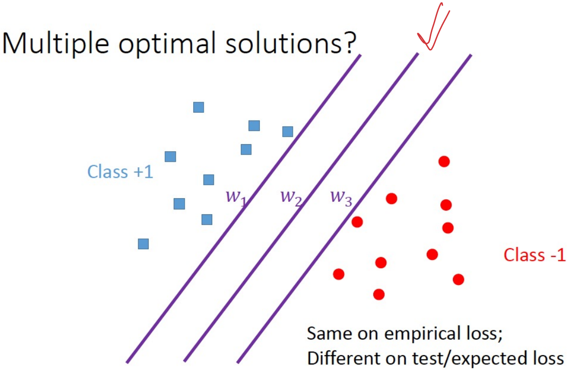
$$[H + \lambda I > 0]$$

Classification



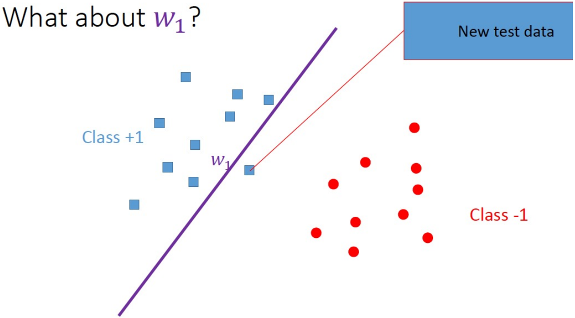
- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.

Classification



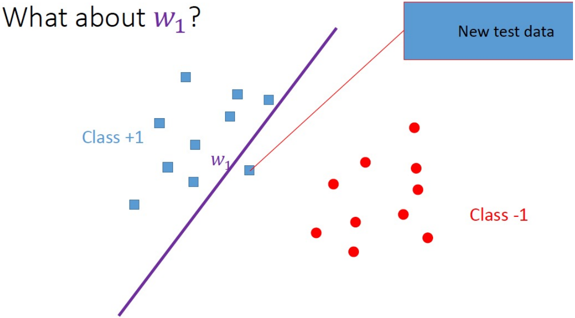
- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.
- How to model such intuition?

Classification



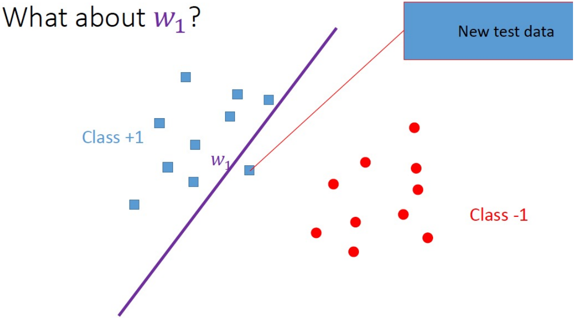
- Just following your intuition, which decision boundary do you prefer?

Classification



- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.

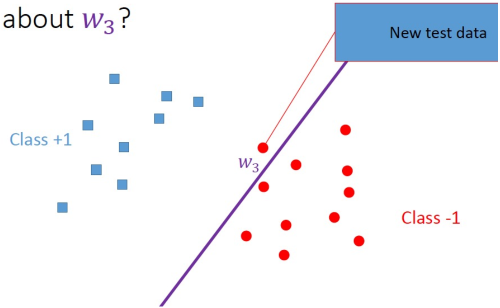
Classification



- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.
- How to model such intuition?

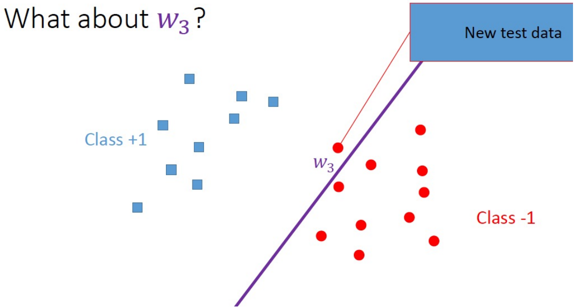
Classification

What about w_3?



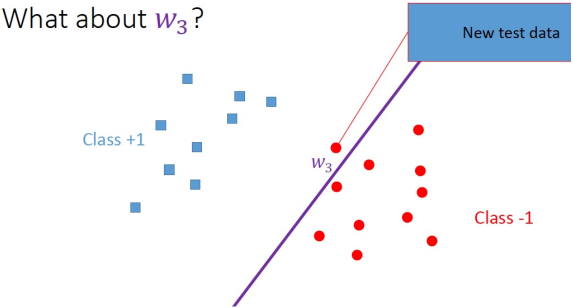
- Just following your intuition, which decision boundary do you prefer?

Classification



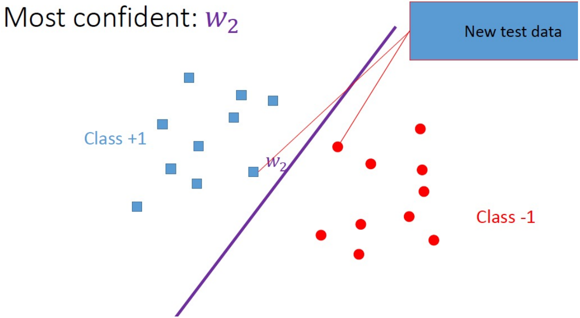
- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.

Classification



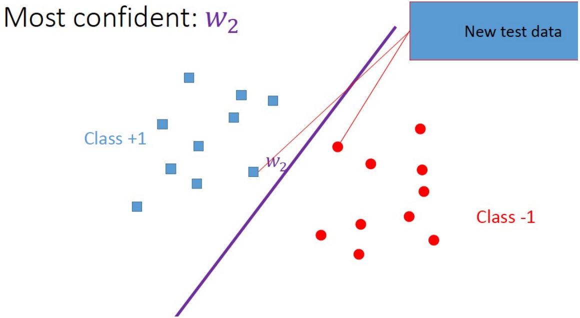
- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.
- How to model such intuition?

Classification



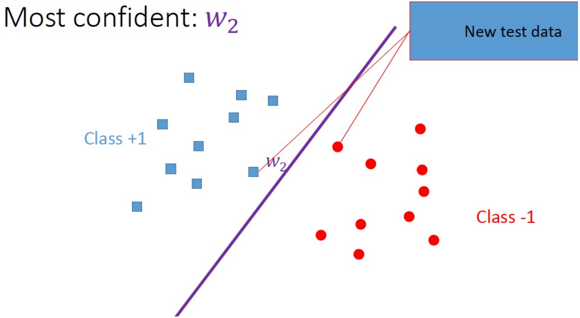
- Just following your intuition, which decision boundary do you prefer?

Classification



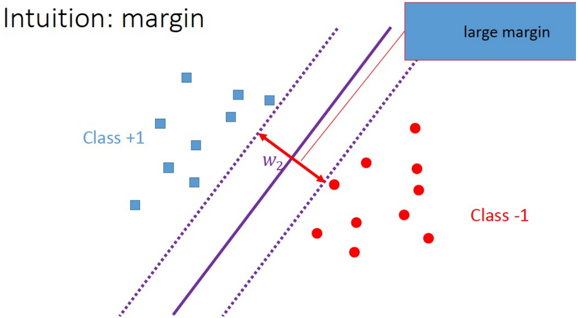
- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.

Classification



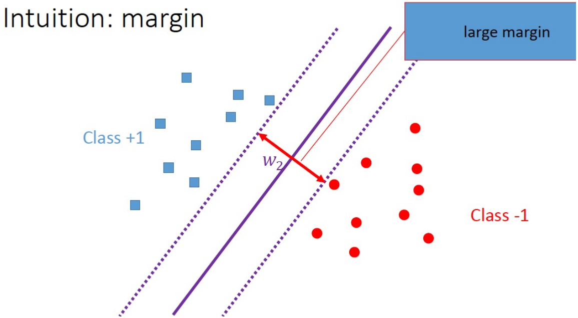
- Just following your intuition, which decision boundary do you prefer?
- The middle one (*i.e.*, $\boldsymbol{w}_2^\top \boldsymbol{x} = 0$) seems better, as it is far from data of both positive and negative classes.
- How to model such intuition?

Large margin intuition



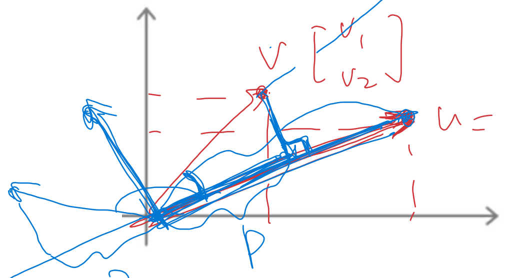
- We introduce the concept **margin**: the distance from the closest point of positive and negative classes to the decision boundary

Large margin intuition



- We introduce the concept **margin**: the distance from the closest point of positive and negative classes to the decision boundary
- The intuition is to choose the decision boundary with large margin, which is called **large margin classifier**, also called **support vector machine** (SVM)

Mathematics behind large margin classification



$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

$u = (u_1; u_2)$

$\|u\|^2 = \sqrt{u_1^2 + u_2^2}$

Inner vector product:

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$

$u^T v = u_1 v_1 + u_2 v_2 = 0$

$= \underbrace{p}_{>0} \cdot \|u\|$

Mathematics behind large margin classification

Inner vector product:

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$
- $\|\boldsymbol{\mu}\| = \sqrt{\mu_1^2 + \mu_2^2}$, the length of $\boldsymbol{\mu}$

Mathematics behind large margin classification

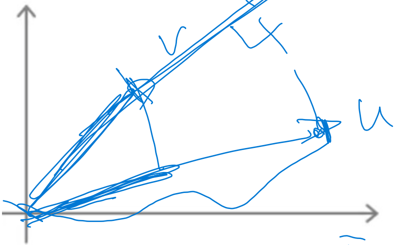
Inner vector product:

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$, $\boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$
- $\|\boldsymbol{\mu}\| = \sqrt{\mu_1^2 + \mu_2^2}$, the length of $\boldsymbol{\mu}$
- $\boldsymbol{\mu}^\top \boldsymbol{\nu} = \mu_1\nu_1 + \mu_2\nu_2$.

Mathematics behind large margin classification

Inner vector product:

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$
- $\|\boldsymbol{\mu}\| = \sqrt{\mu_1^2 + \mu_2^2}$, the length of $\boldsymbol{\mu}$
- $\boldsymbol{\mu}^\top \boldsymbol{\nu} = \mu_1 \nu_1 + \mu_2 \nu_2$. How to represent it in the above plot?



$u^T v = p_1 \cdot \|u\|$

$= p_2 \cdot \|v\|$

Inner vector product:

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$
- $\|\boldsymbol{\mu}\| = \sqrt{\mu_1^2 + \mu_2^2}$, the length of $\boldsymbol{\mu}$
- $\boldsymbol{\mu}^\top \boldsymbol{\nu} = \mu_1 \nu_1 + \mu_2 \nu_2$. How to represent it in the above plot?
- $\boldsymbol{\mu}^\top \boldsymbol{\nu} = p \cdot \|\boldsymbol{\mu}\|$, where p is the length of projection of $\boldsymbol{\nu}$ on $\boldsymbol{\mu}$

Mathematics behind large margin classification

Inner vector product:

- $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\nu} = \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$
- $\|\boldsymbol{\mu}\| = \sqrt{\mu_1^2 + \mu_2^2}$, the length of $\boldsymbol{\mu}$
- $\boldsymbol{\mu}^\top \boldsymbol{\nu} = \mu_1\nu_1 + \mu_2\nu_2$. How to represent it in the above plot?
- $\boldsymbol{\mu}^\top \boldsymbol{\nu} = p \cdot \|\boldsymbol{\mu}\|$, where p is the length of projection of $\boldsymbol{\nu}$ on $\boldsymbol{\mu}$
- Note that if the angle between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ is larger than $90°$, then $p < 0$

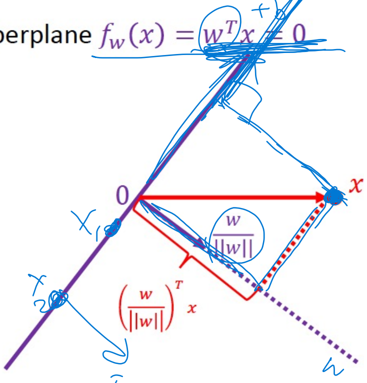
Mathematics behind large margin classification

- Lemma 1: x has distance $\frac{|f_w(x)|}{||w||}$ to the hyperplane $f_w(x) = w^T x = 0$

Proof:

- w is orthogonal to the hyperplane
- The unit direction is $\frac{w}{||w||}$
- The projection of x is $\left(\frac{w}{||w||}\right)^T x = \frac{f_w(x)}{||w||}$

$$\frac{w^T x}{||w||} = p \cdot ||w||$$



$w^T x_1 = 0$

$w^T x_2 = 0$

$w^T (x_1 - x_2) = 0$

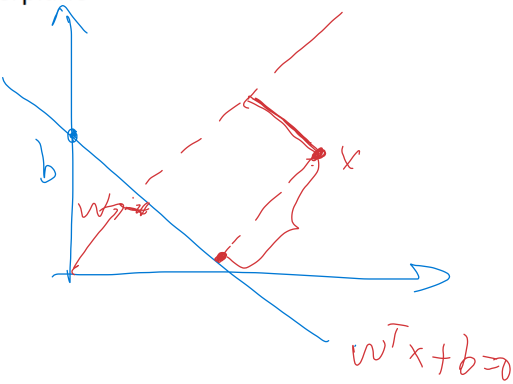
Mathematics behind large margin classification

- Claim 1: w is orthogonal to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Proof:

- pick any x_1 and x_2 on the hyperplane
- $w^T x_1 + b = 0$
- $w^T x_2 + b = 0$

- So $w^T (x_1 - x_2) = 0$

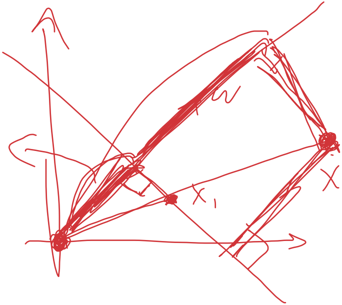


Mathematics behind large margin classification

- Claim 2: 0 has distance $\frac{-b}{||w||}$ to the hyperplane $w^T x + b = 0$

Proof:

- pick any x_1 the hyperplane
- Project x_1 to the unit direction $\frac{w}{||w||}$ to get the distance
- $\left(\frac{w}{||w||}\right)^T x_1 = \frac{-b}{||w||}$ since $w^T x_1 + b = 0$



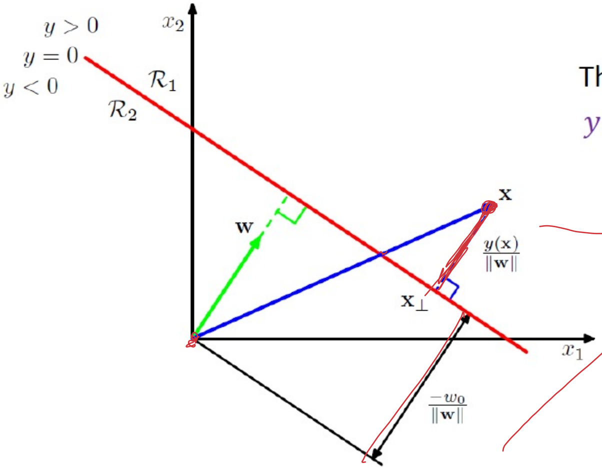
Mathematics behind large margin classification

- Lemma 2: x has distance $\frac{|f_{w,b}(x)|}{||w||}$ to the hyperplane $f_{w,b}(x) = w^T x + b = 0$

Proof:

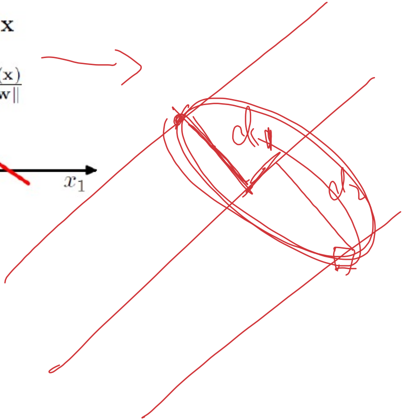
- Let $x = x_{\perp} + r\frac{w}{||w||}$, then $|r|$ is the distance
- Multiply both sides by w^T and add b
- Left hand side: $w^T x + b = f_{w,b}(x)$
- Right hand side: $w^T x_{\perp} + r\frac{w^T w}{||w||} + b = 0 + r||w||$

Mathematics behind large margin classification

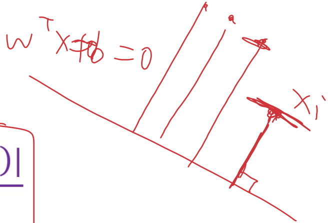


The notation here is:

$$y(x) = w^T x + w_0$$



Large margin classification



- Margin over all training data points:

$$\gamma = \min_i \boxed{\frac{|f_{w,b}(x_i)|}{||w||}}$$

- Since only want correct $f_{w,b}$, and recall $y_i \in \{+1, -1\}$, we have

$$\gamma = \min_i \frac{y_i f_{w,b}(x_i)}{||w||}$$

- If $f_{w,b}$ incorrect on some x_i, the margin is negative

Large margin classification

- Maximize margin over all training data points:

$$\max_{w,b} \gamma = \max_{w,b} \left[\min_i \frac{y_i f_{w,b}(x_i)}{||w||} \right] = \max_{w,b} \min_i \frac{y_i (w^T x_i + b)}{||w||}$$

- A bit complicated …

min max

$$\max_{w,b} \frac{1}{||w||}$$

$$\min ||w||^2$$

Large margin classification

- Observation: when (w, b) scaled by a factor c, the margin unchanged

$$\frac{y_i(cw^T x_i + cb)}{||cw||} = \frac{y_i(w^T x_i + b)}{||w||}$$

$= c||w||$

- Let's consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where x_{i^*} is the point closest to the hyperplane

Large margin classification

- Let's consider a fixed scale such that

$$y_{i^*}(w^T x_{i^*} + b) = 1$$

where x_{i^*} is the point closet to the hyperplane

- Now we have for all data

$$y_i(w^T x_i + b) \geq 1$$

and at least for one i the equality holds

- Then the margin is $\frac{1}{||w||}$

Large margin classification

- Optimization simplified to

$$\min_{w,b} \frac{1}{2} ||w||^2$$

$$y_i(w^T x_i + b) \geq 1, \forall i$$

Alternative view of logistic regression

- Hypothesis function:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x})} = g(z)$$

  where $z = \boldsymbol{w}^\top \boldsymbol{x}$

Alternative view of logistic regression

- Hypothesis function:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x})} = g(z)$$

  where $z = \boldsymbol{w}^\top \boldsymbol{x}$

- If $y = 1$, we want $h_{\boldsymbol{w}}(\boldsymbol{x}) \approx 1$, *i.e.*, $\boldsymbol{w}^\top \boldsymbol{x} \gg 0$

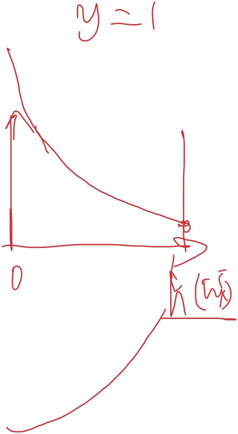
Alternative view of logistic regression

- Hypothesis function:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^\top \boldsymbol{x})} = g(z)$$

where $z = \boldsymbol{w}^\top \boldsymbol{x}$

- If $y = 1$, we want $h_{\boldsymbol{w}}(\boldsymbol{x}) \approx 1$, *i.e.*, $\boldsymbol{w}^\top \boldsymbol{x} \gg 0$
- If $y = -1$, we want $h_{\boldsymbol{w}}(\boldsymbol{x}) \approx 0$, *i.e.*, $\boldsymbol{w}^\top \boldsymbol{x} \ll 0$



Alternative view of logistic regression

- Hypothesis function:

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{w}^\top \boldsymbol{x})} = g(z)$$

  where $z = \boldsymbol{w}^\top \boldsymbol{x}$
- If $y = 1$, we want $h_{\boldsymbol{w}}(\boldsymbol{x}) \approx 1$, *i.e.*, $\boldsymbol{w}^\top \boldsymbol{x} \gg 0$
- If $y = -1$, we want $h_{\boldsymbol{w}}(\boldsymbol{x}) \approx 0$, *i.e.*, $\boldsymbol{w}^\top \boldsymbol{x} \ll 0$
  $(-y)$
- Objective function of logistic regression

$$J(\boldsymbol{w}) = -\delta_{y=1} \log(h_{\boldsymbol{w}}(\boldsymbol{x})) - \delta_{y=-1} \log(1 - h_{\boldsymbol{w}}(\boldsymbol{x})), \tag{1}$$

  where $\delta_a = 1$ if a is true, otherwise 0.

Objective of SVM

Cost_1 Cost_{-1}

- Objective function of logistic regression

$$\frac{1}{m}\sum_{i}^{m}\left[\delta_{y^{(i)}=1}\left(-\log(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))\right)+\delta_{y^{(i)}=-1}\left(-\log(1-h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))\right)\right]+\frac{\lambda}{2m}\sum_{j=1}^{n}w_j^2$$

Objective of SVM

- Objective function of logistic regression

$$\frac{1}{m}\sum_{i}^{m}\Big[\delta_{y^{(i)}=1}\big(-\log(h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))\big)+\delta_{y^{(i)}=-1}\big(-\log(1-h_{\boldsymbol{w}}(\boldsymbol{x}^{(i)}))\big)\Big]+\frac{\lambda}{2m}\sum_{j=1}^{n}w_j^2$$

- Objective function of **support vector machine**

$$\frac{1}{m}\sum_{i}^{m}\Big[\delta_{y^{(i)}=1}\text{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)+\delta_{y^{(i)}=-1}\text{cost}_{-1}(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)\Big]+\frac{\lambda}{2m}\sum_{j=1}^{n}w_j^2$$

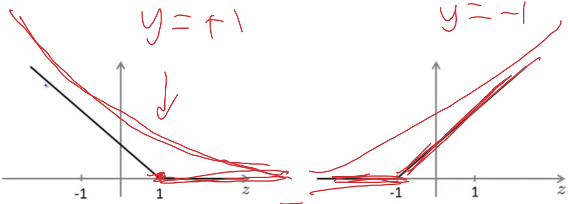
$$\equiv C\sum_{i}^{m}\Big[\delta_{y^{(i)}=1}\text{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)+\delta_{y^{(i)}=-1}\text{cost}_{-1}(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)\Big]+\frac{1}{2}\sum_{j=1}^{n}w_j^2$$

Objective of SVM

Objective of SVM

- Objective function of support vector machine

$$C\sum_{i}^{m}\left[\delta_{y^{(i)}=1}\text{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)+\delta_{y^{(i)}=-1}\text{cost}_{-1}(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)\right]+\frac{1}{2}\sum_{j=1}^{n}w_j^2$$



$z = w^T x$

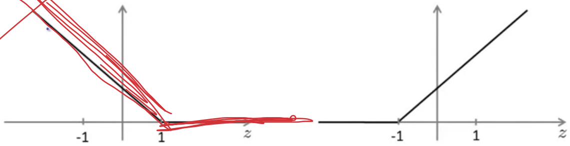
$h(w^T x)$

Objective of SVM

- Objective function of support vector machine

$$C\sum_{i}^{m}\left[\delta_{y^{(i)}=1}\boxed{\mathrm{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)}+\delta_{y^{(i)}=-1}\mathrm{cost}_{-1}(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)\right]+\frac{1}{2}\sum_{j=1}^{n}w_j^2$$

$y = 1$



- If $y=+1$, we require that $\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b\geq 1$. In other words, $\mathrm{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)=0$ if $\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b\geq 1$

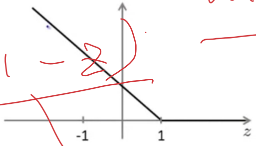
Objective of SVM

- Objective function of support vector machine

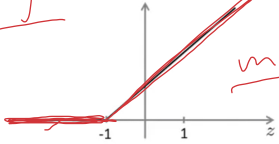
$$C \sum_{i}^{m} \left[\delta_{y^{(i)}=1} \text{cost}_1(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) + \delta_{y^{(i)}=-1} \text{cost}_{-1}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) \right] + \frac{1}{2} \sum_{j=1}^{n} w_j^2$$

y = 1 hinge loss y = -1

max(0, 1 - z)



max(0, 1+z)



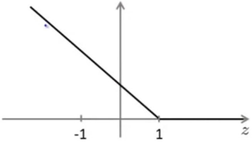
- If $y = +1$, we require that $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \geq 1$. In other words, $\text{cost}_1(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) = 0$ if $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \geq 1$
- If $y = +1$, we require that $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \leq -1$. In other words, $\text{cost}_{-1}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) = 0$ if $\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \leq -1$

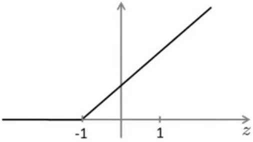
max(0, 1 - y · z)

Objective of SVM

- Objective function of support vector machine

$$C\sum_{i}^{m}\left[\delta_{y^{(i)}=1}\text{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)+\delta_{y^{(i)}=-1}\text{cost}_{-1}(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)\right]+\frac{1}{2}\sum_{j=1}^{n}w_j^2$$





- If $y=+1$, we require that $\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b\geq 1$. In other words, $\text{cost}_1(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)=0$ if $\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b\geq 1$
- If $y=+1$, we require that $\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b\leq -1$. In other words, $\text{cost}_{-1}(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)=0$ if $\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b\leq -1$
- **Hinge loss:**

$$\max(0,1-y(\boldsymbol{w}^\top\boldsymbol{x}^{(i)}+b)) \tag{2}$$

Mathematics behind large margin classification

- However, hinge loss is non-smooth. We transform the objective function of support vector machine to the following

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{j=1}^{n} w_j^2 \tag{3}$$

$$s.t.\ \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \geq 1, \text{if } y^{(i)} = 1;\ \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b < -1, \text{if } y^{(i)} = -1.$$

Mathematics behind large margin classification

- However, hinge loss is non-smooth. We transform the objective function of support vector machine to the following

$$\begin{aligned} & \min_{\boldsymbol{w}} \frac{1}{2}\sum_{j=1}^{n} w_j^2 \\ & s.t.\ \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \geq 1, \text{if } y^{(i)} = 1;\ \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b < -1, \text{if } y^{(i)} = -1. \end{aligned} \tag{3}$$

- It can be simplified as follows

$$\begin{aligned} & \min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ & s.t.\ y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) \geq 1, \forall i \end{aligned} \tag{4}$$

Mathematics behind large margin classification

- However, hinge loss is non-smooth. We transform the objective function of support vector machine to the following

$$\min_{\boldsymbol{w}} \frac{1}{2}\sum_{j=1}^{n} w_j^2 \tag{3}$$

$$s.t.\ \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b \geq 1, \text{if } y^{(i)} = 1;\ \boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b < -1, \text{if } y^{(i)} = -1.$$

- It can be simplified as follows

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{4}$$

$$s.t.\ y^{(i)}(\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b) \geq 1, \forall i$$

- Utilizing $p = \frac{\boldsymbol{w}^\top \boldsymbol{x}+b}{\|\boldsymbol{w}\|}$, we have

$$\boldsymbol{w}^\top \boldsymbol{x}^{(i)} + b = p^{(i)} \cdot \|\boldsymbol{w}\| \tag{5}$$

Mathematics behind large margin classification

- The objective function of support vector machine is transformed to

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &s.t.\ y^{(i)} \cdot p^{(i)} \cdot \|\boldsymbol{w}\| \geq 1, \forall i \end{aligned} \tag{6}$$

where $p^{(i)}$ indicates the projection length of $\boldsymbol{x}^{(i)}$ on $\boldsymbol{w}$.

Mathematics behind large margin classification

- The objective function of support vector machine is transformed to

$$\begin{aligned} &\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &s.t.\ y^{(i)} \cdot p^{(i)} \cdot \|\boldsymbol{w}\| \geq 1, \forall i \end{aligned} \tag{6}$$

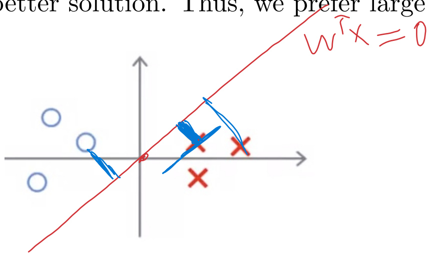
  where $p^{(i)}$ indicates the projection length of $\boldsymbol{x}^{(i)}$ on $\boldsymbol{w}$.
- Let's see the following two decision boundaries (plot below)

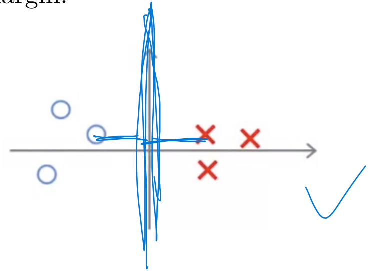
Mathematics behind large margin classification

- The objective function of support vector machine is transformed to

$$\begin{aligned} \min_{\boldsymbol{w}} \ & \frac{1}{2}\|\boldsymbol{w}\|^2 \\ s.t. \ & y^{(i)} \cdot p^{(i)} \cdot \|\boldsymbol{w}\| \geq 1, \forall i \end{aligned} \tag{6}$$

  where $p^{(i)}$ indicates the projection length of $\boldsymbol{x}^{(i)}$ on $\boldsymbol{w}$.
- Let's see the following two decision boundaries (plot below)
- If the projection length p is larger, then $\|\boldsymbol{w}\|$ could be smaller, leading to better solution. Thus, we prefer large margin.





Reading material

Reading materials:

- Andrew Ng's note on SVM:
  `https://see.stanford.edu/materials/aimlcs229/cs229-notes3.pdf`
- Chapter 7.1 of Bishop's book