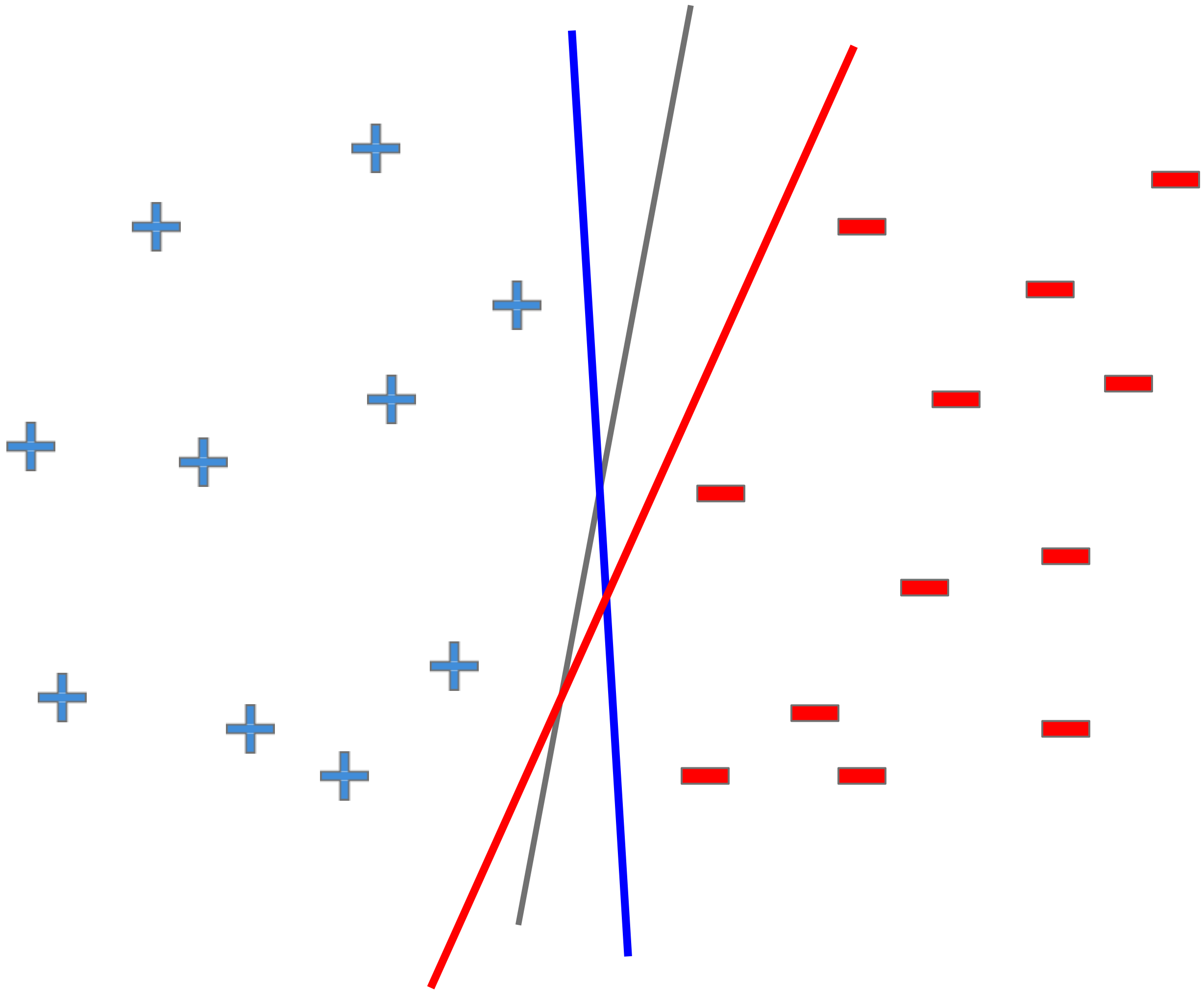
机器学习

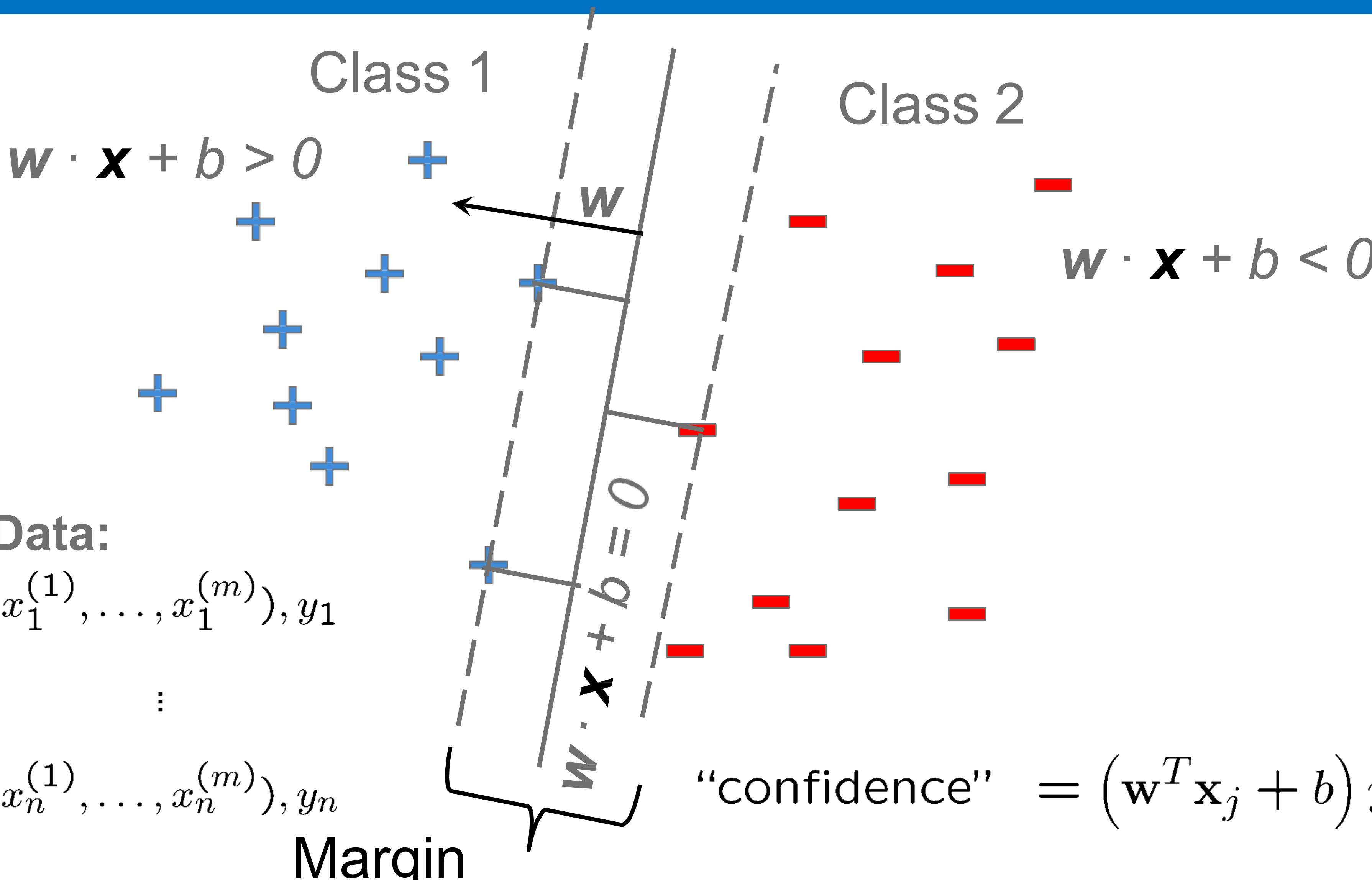
实验课题1-Support Vector Machines

Linear classifiers
which line is better?



Which decision boundary is better?

Pick the one with the largest margin!



Data:

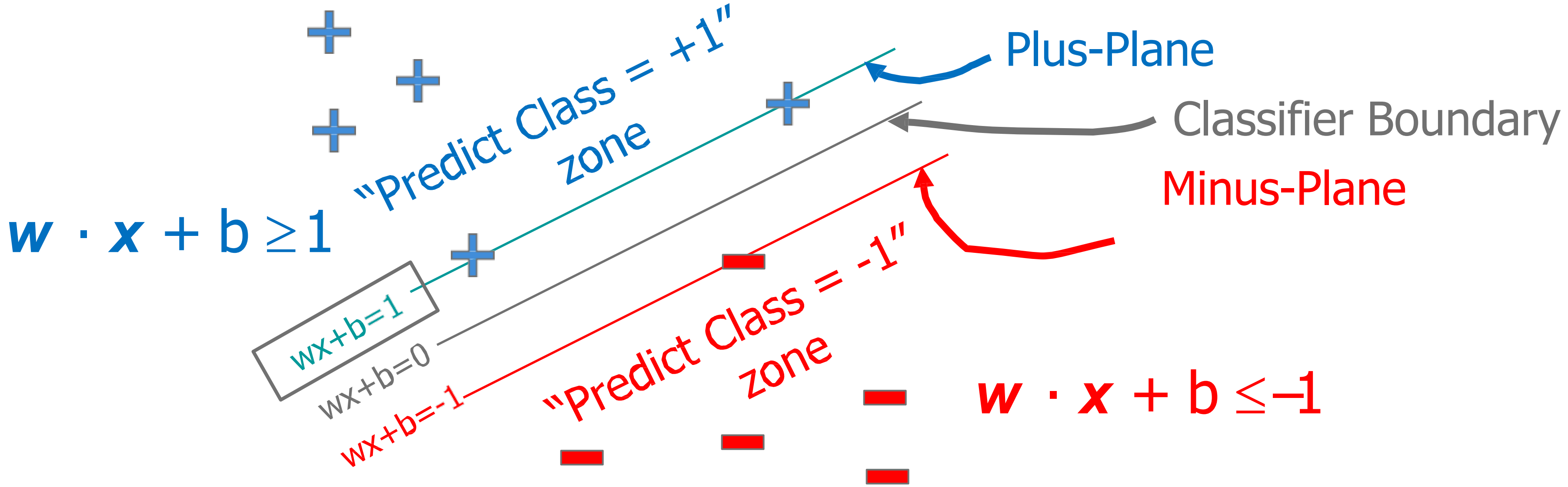
$$(x_1^{(1)}, \ldots, x_1^{(m)}), y_1$$

$$\vdots$$

$$(x_n^{(1)}, \ldots, x_n^{(m)}), y_n$$

$$\text{“confidence”} = \left(\mathbf{w}^T \mathbf{x}_j + b\right) y_j$$

Scaling



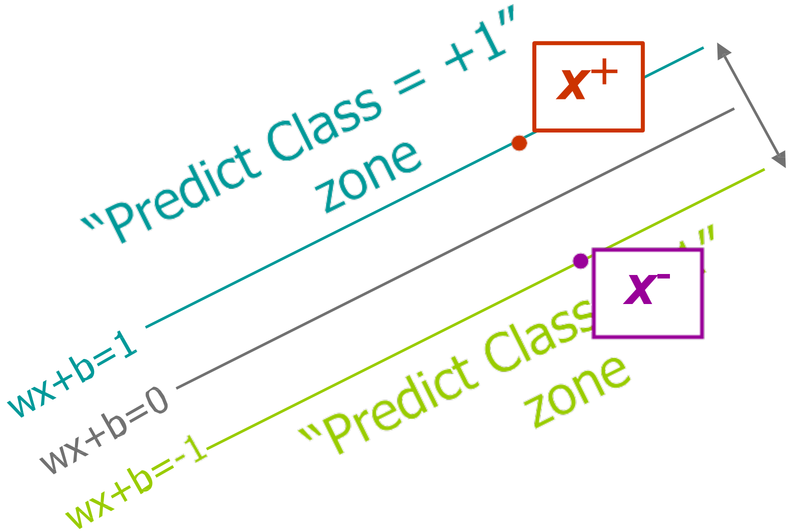
Classification rule:

Classify as..	+1	if	$\boldsymbol{w} \cdot \boldsymbol{x} + b \geq 1$
	-1	if	$\boldsymbol{w} \cdot \boldsymbol{x} + b \leq -1$
	Universe explodes	if	$-1 < \boldsymbol{w} \cdot \boldsymbol{x} + b < 1$

How large is the margin of this classifier?

Goal: Find the maximum margin classifier

Computing the margin width



M = Margin Width $= \dfrac{2}{\sqrt{\mathbf{w} \cdot \mathbf{w}}}$

Let $\boldsymbol{x}^+$ and $\boldsymbol{x}^-$ be such that

- $\boldsymbol{w} \cdot \boldsymbol{x}^+ + b = +1$
- $\boldsymbol{w} \cdot \boldsymbol{x}^- + b = -1$
- $\boldsymbol{x}^+ = \boldsymbol{x}^- + \lambda \boldsymbol{w}$
- $|\boldsymbol{x}^+ - \boldsymbol{x}^-| = M$

Maximize M $\equiv$ minimize **w·w !**

Observations

We can assume b=0

Classify as..	+1	if	$\boldsymbol{w} \cdot \boldsymbol{x} + b \geq 1$
	−1	if	$\boldsymbol{w} \cdot \boldsymbol{x} + b \leq -1$
	Universe explodes	if	$-1 < \boldsymbol{w} \cdot \boldsymbol{x} + b < 1$

This is the same as $y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \geq 1, \; \forall i = 1, \ldots, n$

The Primal SVM

- Given $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ training data set.
- Assume that D is **linearly separable**.

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathbb{R}^m} \frac{1}{2} \|\mathbf{w}\|^2$$

$$\text{subject to } y_i \langle \mathbf{x}_i, \mathbf{w} \rangle \geq 1, \ \forall i = 1, \ldots, n$$

Prediction: $f_{\hat{\mathbf{w}}}(\mathbf{x}) = \text{sign}(\langle \hat{\mathbf{w}}, \mathbf{x} \rangle)$

This is a QP problem (m-dimensional)
(Quadratic cost function, linear constraints)

The Primal SVM

- Given $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ training data set.
- Assume that D is **linearly separable**.

$$\text{argmin: } \frac{1}{2}|\mathrm{w}|^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i \mathrm{w}^T x_i)$$

The cost function is often referred to as **the primal optimization problem**, which seeks to minimize the sum of hinge losses over all training examples, subject to a constraint that the magnitude of the weight vector is bounded.

The Primal SVM

- Given $D = \{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$ training data set.
- Assume that D is **linearly separable**.

When we assume b≠0

$$\text{argmin: } \frac{1}{2}|\mathrm{w}|^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(\mathrm{w}^T x_i + b))$$

The cost function is often referred to as **the primal optimization problem**, which seeks to minimize the sum of hinge losses over all training samples, subject to a constraint that the magnitude of the weight vector is bounded.

Implementation--Training

1. Any programming language can be used.
2. Write the function of the gradient computation with respect to **w** and b (if you assume b $\neq$ 0). (manually derive the gradients on papers).
3. Write the function of computing the loss of the objective function over all training samples.
4. Construct a loop structure to compute gradients, update weights (**w**, b) according to the gradients, and update the loss in each iteration.

When we assume b≠0

argmin: $\frac{1}{2}|\mathrm{w}|^2 + C\sum_{i=1}^{n}\max(0, 1 - y_i(\mathrm{w}^T x_i + b))$

Algorithm: gradient descent

Initialize $\mathbf{w} = [0, \ldots, 0]$

For $t = 1, \ldots, T$:

$$\mathbf{w} \leftarrow \mathbf{w} - \underbrace{\eta}_{\text{step size}} \underbrace{\nabla_{\mathbf{w}} \text{TrainLoss}(\mathbf{w})}_{\text{gradient}}$$

Implementation--Testing

1. Load the trained model (load the parameters of **w** and b learned).
2. Load the testing data
3. Input each testing sample into the below prediction function to make a prediction.
4. Report the accuracy by comparing the prediction outputs and the true labels.

Prediction:

$$f_{\widehat{\mathbf{W}}}(x) = \mathrm{sign}\left(\langle \widehat{\mathbf{W}}, \mathbf{x} \rangle\right)$$

$$f_{\widehat{\mathbf{W}}}(x) = \mathrm{sign}\left(\langle \widehat{\mathbf{W}}, \mathbf{x} \rangle + b\right)$$