# Machine Learning

CSE 142

Xin (Eric) Wang

Friday, October 29 , 2021

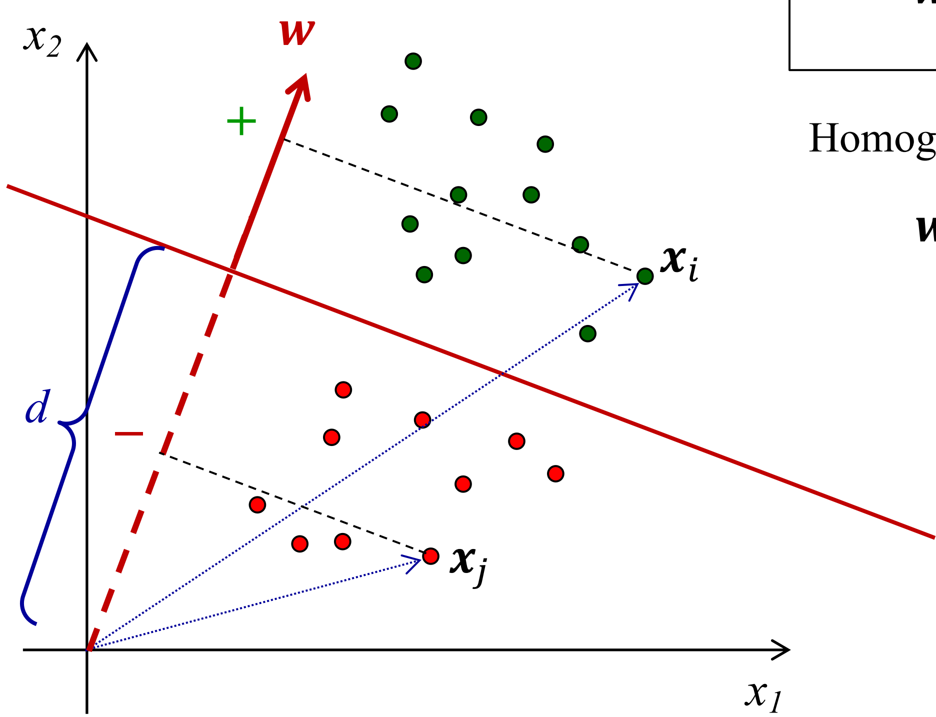
Today

- Linear learning models (cont.)
  - Classifier margin
  - *SVM

# Notes

- Our tutor Winson is running a poll of changing the time of his office hours on Piazza to suit you the best
  - Ends next Monday
- Machine learning in practice, guest lecture by Linjie Li, a researcher at Microsoft, Nov. 3 (next Wednesday)
- HW2 extended by two days, now due on Nov. 5 (next Friday)
  - HW3 will also be released then
  - No further extension (but you can use your four late days)

# Classifier geometry – $\boldsymbol{w}$ and $t$




Non-homogeneous:

$$\boldsymbol{w}^T\boldsymbol{x} - t = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - t > 0$$

Homogeneous:

$$\boldsymbol{w}^T\boldsymbol{x} = \begin{bmatrix} w_1 & w_2 & -t \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ 1 \end{bmatrix} > 0$$

Is $\boldsymbol{w}$ a unit vector?

*Doesn't have to be*

What's the relationship between $\boldsymbol{w}$ and $t$ ?

$(w, t) \equiv (kw, kt)$

$$\begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - t = 0 \qquad \begin{bmatrix} 2w_1 & 2w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 2t = 0$$

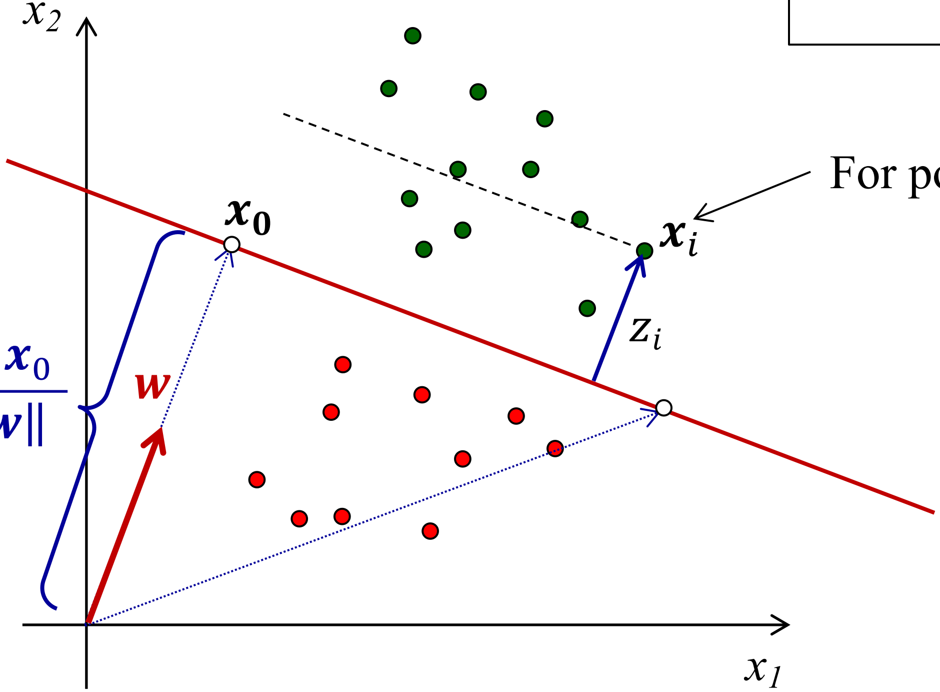
These describe the same line

# Classifier geometry – $\boldsymbol{w}$ and $\boldsymbol{t}$

Non-homogeneous:

$$\boldsymbol{w}^T\boldsymbol{x} - t = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - t > 0$$

$$\frac{t}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T\boldsymbol{x}_0}{\|\boldsymbol{w}\|}$$




For point $\boldsymbol{x_i}$:

Distance: $d_i - d = \dfrac{\boldsymbol{w}^T\boldsymbol{x_i}}{\|\boldsymbol{w}\|} - \dfrac{t}{\|\boldsymbol{w}\|}$

Margin: $z_i = \dfrac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i - t)}{\|\boldsymbol{w}\|}$

$\boldsymbol{w}^T\boldsymbol{x} = t$

Note:

Regular classifier: the margin is the distance from the decision boundary

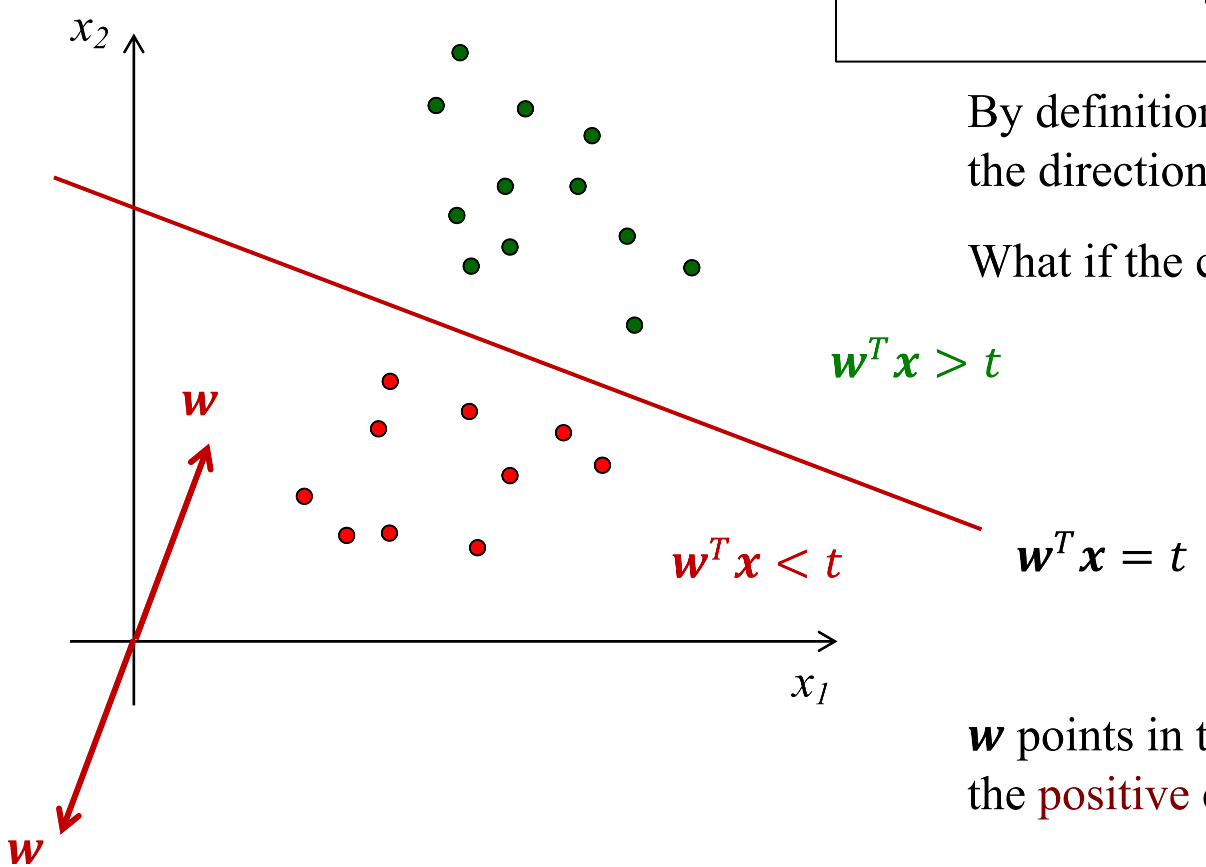
Scoring classifier: the margin is the score

In both cases, value is positive for correctly classified, negative for incorrectly classified

# Classifier geometry – $\boldsymbol{w}$ and $t$

Non-homogeneous:

$$\boldsymbol{w}^T\boldsymbol{x} - t = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - t > 0$$




By definition, the vector $\boldsymbol{w}$ points in the direction of the positive class.

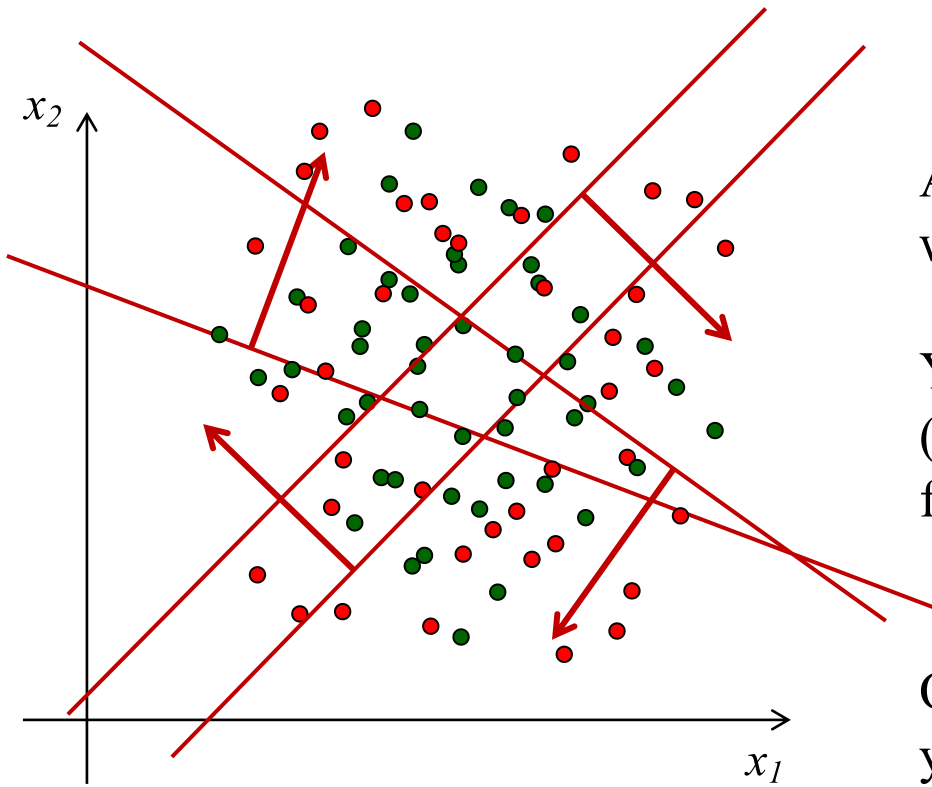
What if the classes here are swapped?

$$\boldsymbol{w} \leftarrow (-\boldsymbol{w})$$

$$t \leftarrow (-t)$$

$\boldsymbol{w}$ points in the (relative) direction of the positive class

# Classifier geometry – $\boldsymbol{w}$ and $t$




An appropriate choices of $\boldsymbol{w}$ and $t$ will achieve any decision line

You can start with the line equation (and direction of positive class) and figure out $\boldsymbol{w}$ and $t$

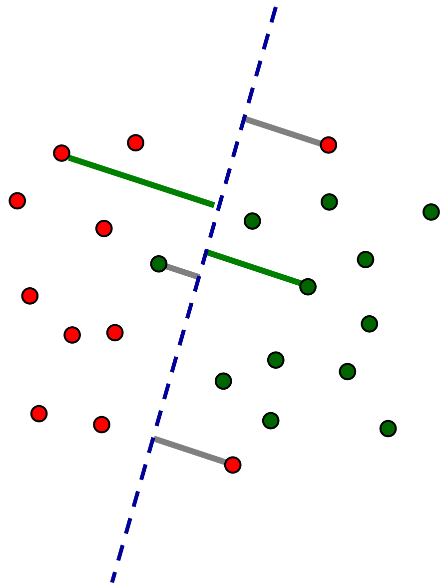
Or, alternative, if given $\boldsymbol{w}$ and $t$, you can figure out the classification line

Non-homogeneous:

$$\boldsymbol{w}^T\boldsymbol{x} - t = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - t > 0$$

# Classifier margin

- The margin ($z$) of a sample is its distance from the classification boundary
  - Positive if it's correctly classified
  - Negative if it's incorrectly classified



Perceptron margin for point $\boldsymbol{x}$:

$$\boldsymbol{z}(\boldsymbol{x}) = \frac{y(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} - t)}{||\boldsymbol{w}||} = \frac{m}{||\boldsymbol{w}||}$$

Non-homogeneous representation

Note: $m$ is not the margin; it's the result of plugging $\boldsymbol{x}_i$ into $y(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} - t)$

# Margin, distance, and *m*

- Note that on p. 211 in the textbook (beginning of Section 7.3), the discussion of *m*, margin, and distance is a bit confusing.

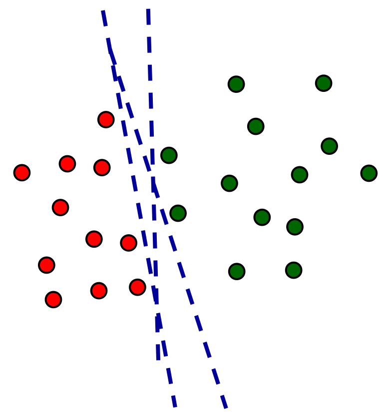
- In this class, we use the following terminology:

$$\text{Nonhomogeneous}\begin{cases} m = y(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} - t) & \text{(not a measure of distance)} \\[2ex] \text{margin: } \boldsymbol{z}(\boldsymbol{x}) = \dfrac{y(\boldsymbol{wTx} - t)}{||\boldsymbol{w}||} = \dfrac{m}{||\boldsymbol{w}||} & \text{(a measure of distance)} \end{cases}$$

(I think the book is relatively consistent with this terminology elsewhere)

# Classifiers and margins

- The class margin (on the training set) is the minimum margin of the data points for that class
- The classifier margin is the sum of the class margins
- There are an infinite number of linear classifiers that can perfectly separate linearly separable data



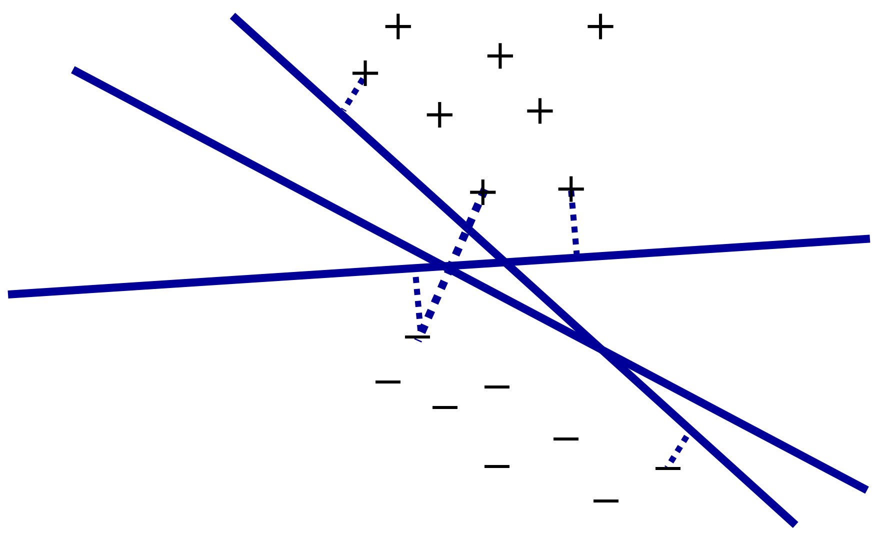
Infinitely many lines can separate the two classes

- But which (of all these) is the best linear classifier?
  - Perhaps the one that maximizes the classifier margin

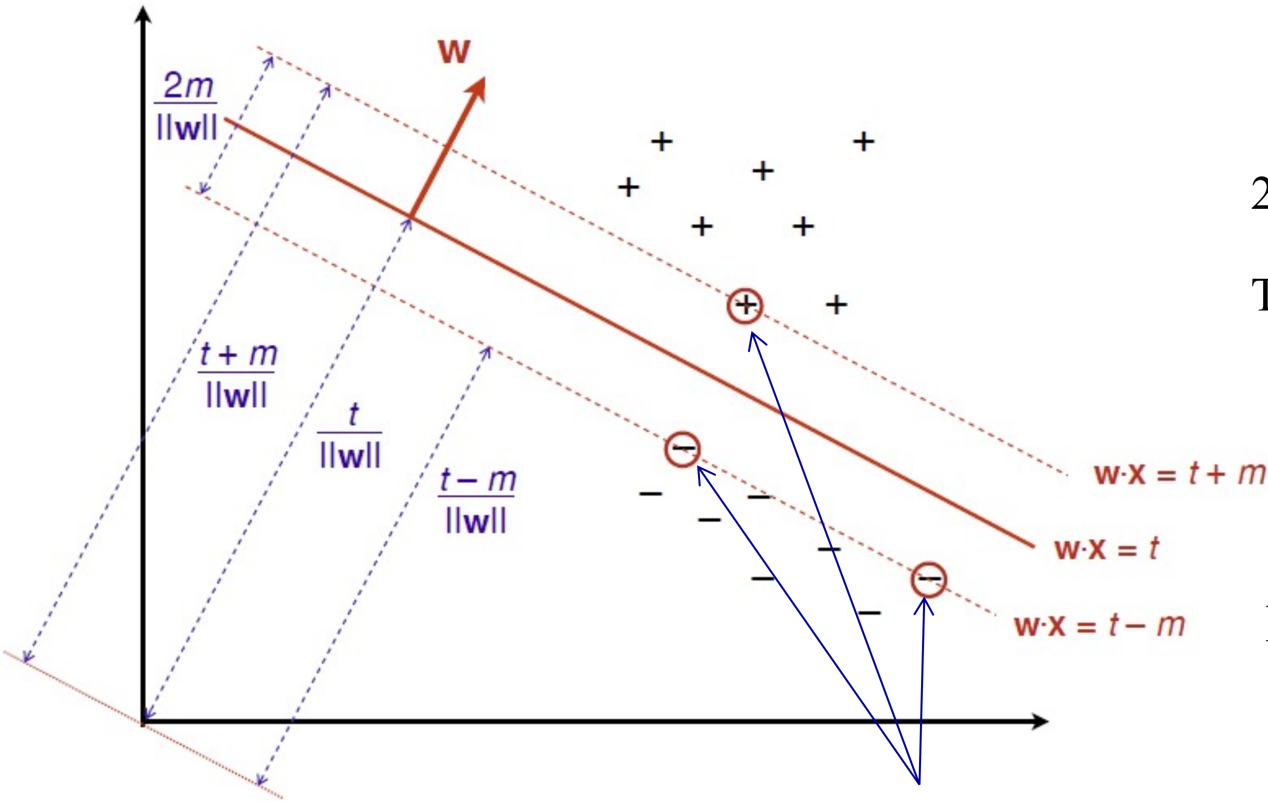
# Classifier margin

Let's look at the margins for a given training set and decision boundary:

Choose a classification line that maximizes the classifier margin (i.e., the sum of the class margins)



# Classifier margin




Choose the decision boundary $\boldsymbol{w}^T\boldsymbol{x} - t = 0$ that:

1. Maximizes the sum of the minimum positive class margin and the minimum negative class margin
2. Makes them equal

Thus, maximize $\frac{2m}{||\boldsymbol{w}||}$

$$\text{Margin}(\boldsymbol{x}) = \frac{y(\boldsymbol{w}^T\boldsymbol{x} - t)}{||\boldsymbol{w}||} = \frac{m}{||\boldsymbol{w}||}$$

These training samples nearest to the decision boundary define the classifier margin and are called support vectors

# Support vector machine (SVM)

- A support vector machine (SVM) is a linear classifier whose decision boundary is a linear combination of the support vectors (training samples at the margins)
- In an SVM, we find classifier parameters $(\boldsymbol{w}, t)$ that maximize the classifier margin
- Since $m = y(\boldsymbol{w}^T\boldsymbol{x} - t)$ and we wish to maximize the margin $\frac{m}{||\boldsymbol{w}||}$, we can instead fix $m = 1$ and minimize $||w||$
  - Provided that none of the training points fall inside the margin
- This leads to a constrained optimization problem:

$$\mathbf{w}^*, t^* = \underset{\mathbf{w},t}{\arg\min} \frac{1}{2}||\mathbf{w}||^2 \qquad \text{subject to } y_i(\mathbf{w}\cdot\mathbf{x}_i - t) \geq 1, 1 \leq i \leq n$$

# Quiz: margin, support vectors, perceptron, SVM

**Algorithm** Perceptron($D, \eta$) – train a perceptron for linear classification.

**Input** : labelled training data $D$ in homogeneous coordinates; learning rate $\eta$.
**Output** : weight vector $\mathbf{w}$ defining classifier $\hat{y} = \text{sign}(\mathbf{w} \cdot \mathbf{x})$.

```
w ← 0 ;                    // Other initialisations of the weight vector are possible
converged ← false;
while converged = false do
    converged ← true;
    for i = 1 to |D| do
        if y_i w·x_i ≤ 0           // i.e., ŷ_i ≠ y_i
        then
            w ← w + η y_i x_i ;
            converged ← false;     // We changed w so haven't converged yet
        end
    end
end
```