50.007
Machine Learning

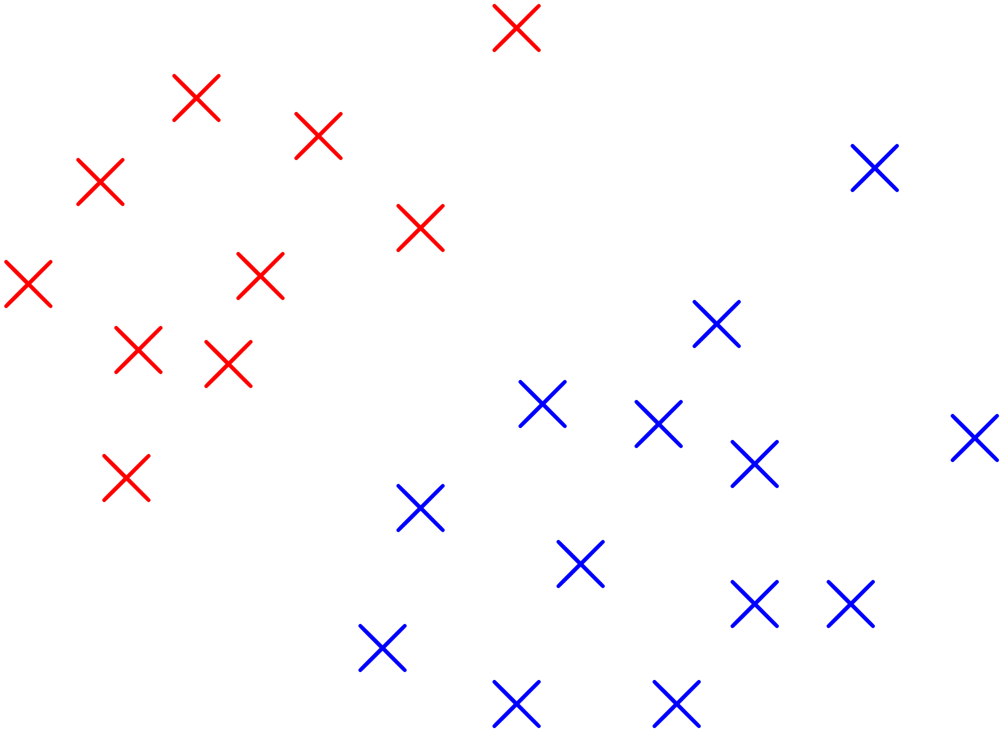
Lu, Wei

SINGAPORE UNIVERSITY OF
TECHNOLOGY AND DESIGN

Logistic Regression

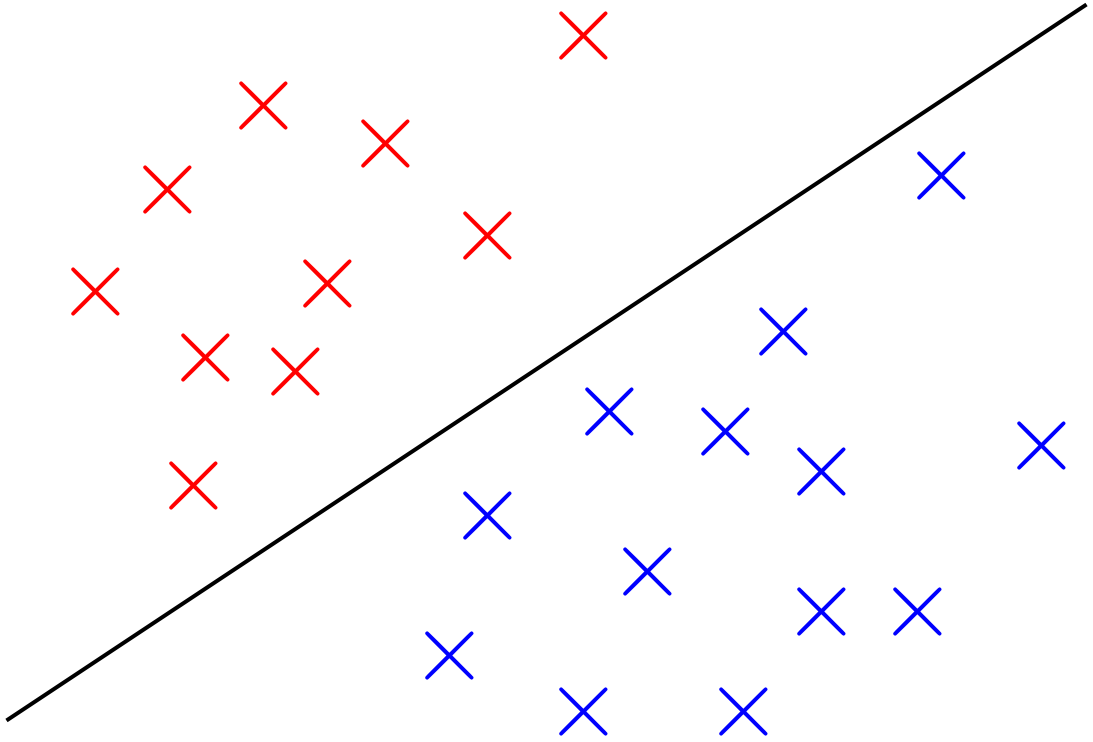
Linear Classification

Linearly Separable



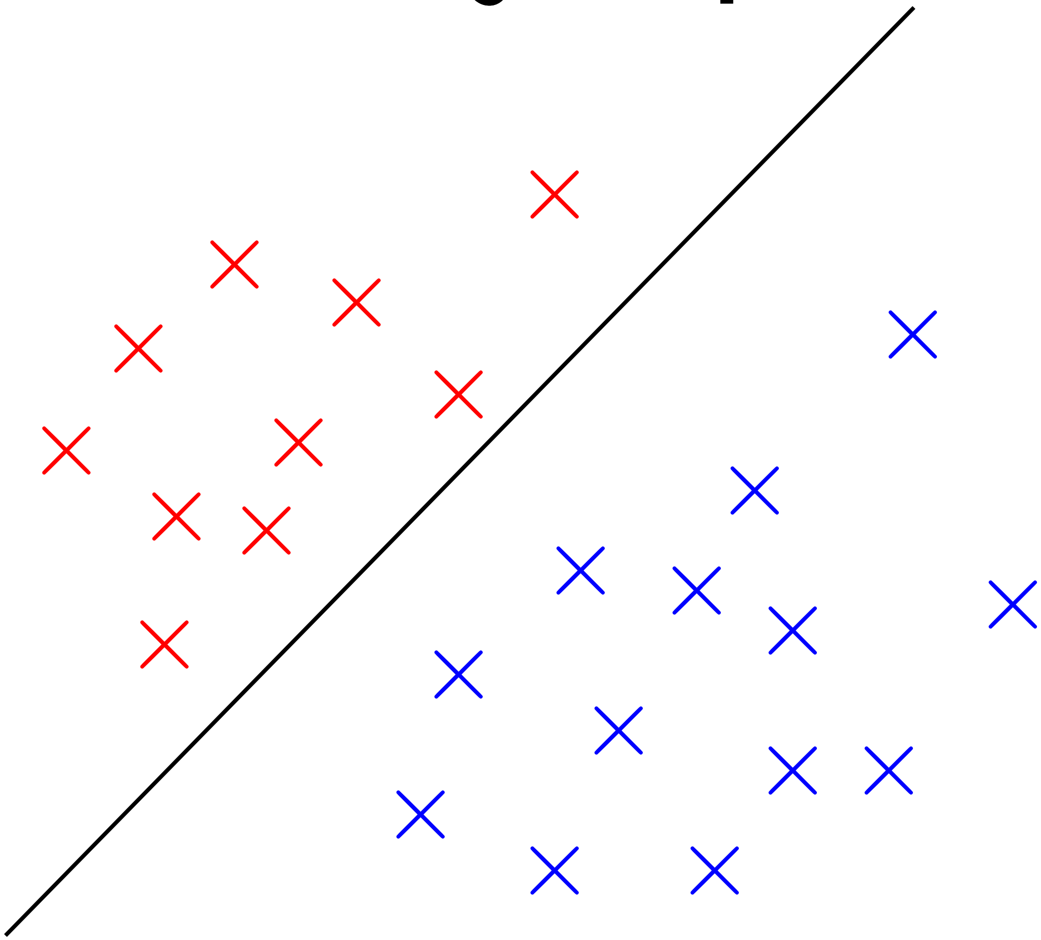
Linear Classification

Linearly Separable



Linear Classification

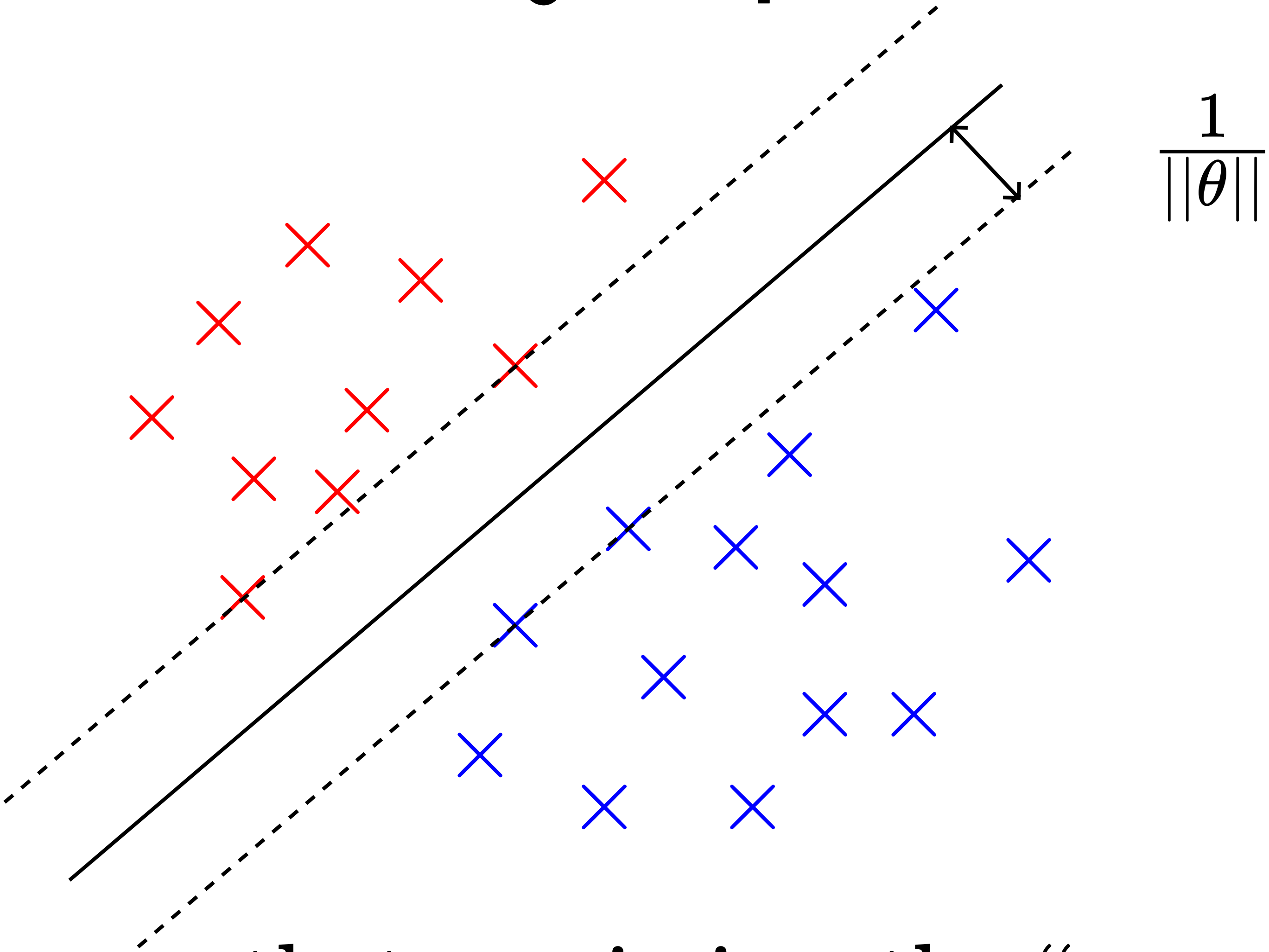
Linearly Separable



Which is the “best” hyperplane?

Linear Classification

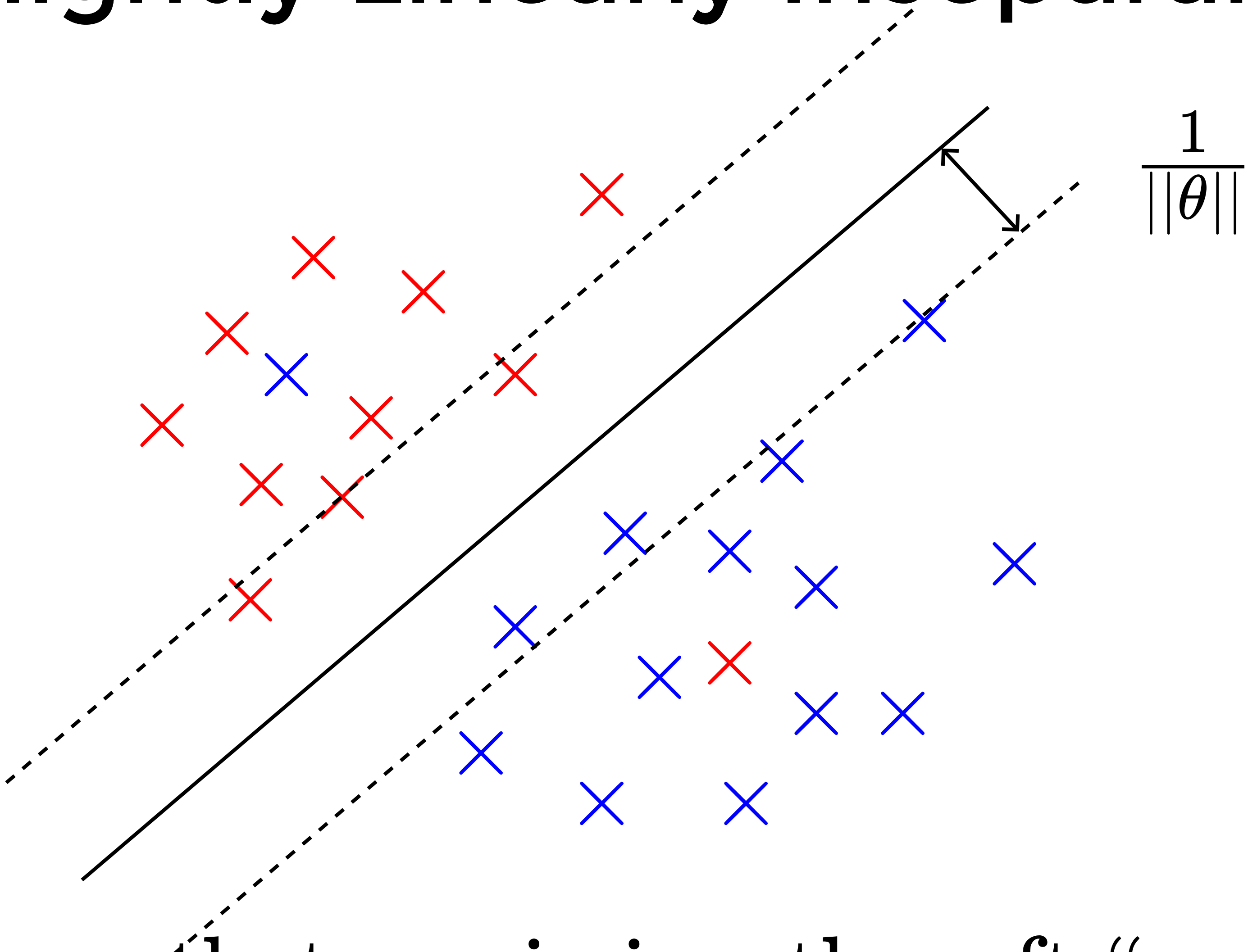
Linearly Separable



The one that maximizes the “margin”!

Linear Classification

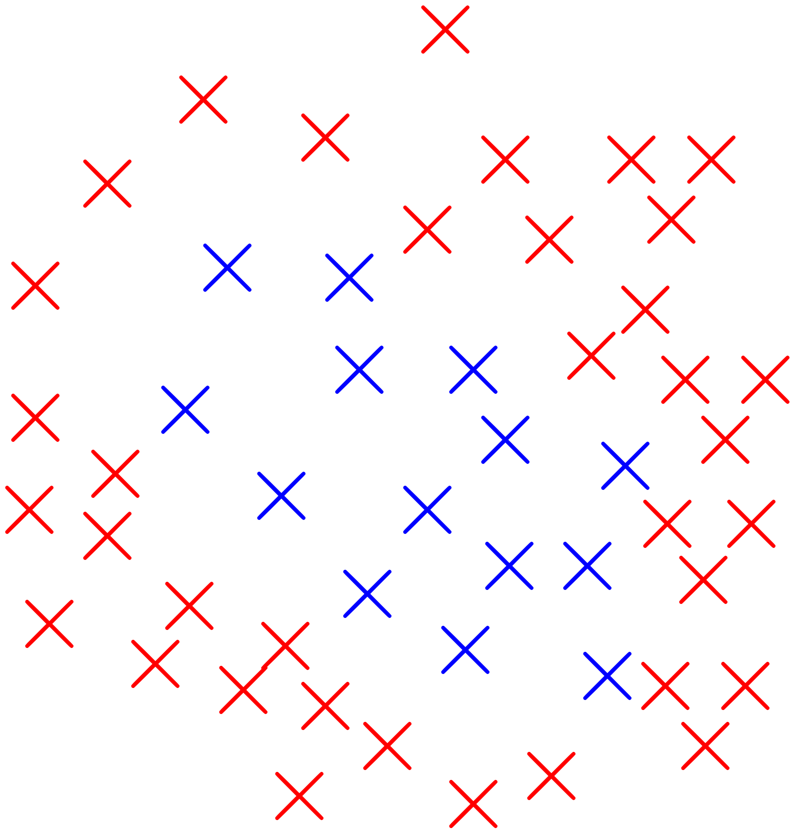
Slightly Linearly Inseparable



The one that maximizes the soft “margin”!

Linear Classification

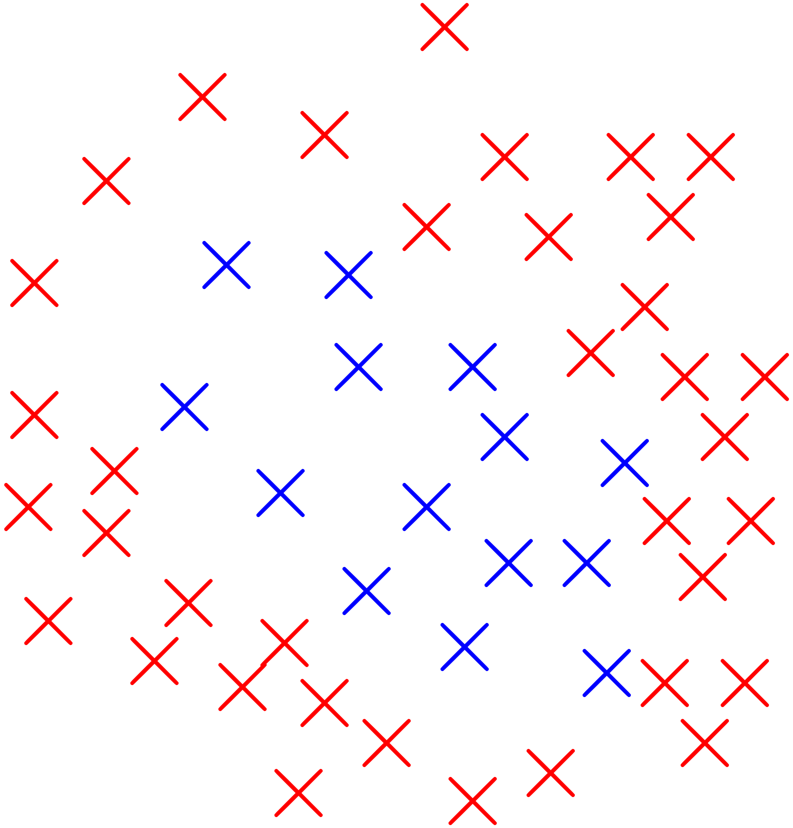
Severely Linearly Inseparable



We will have to use the kernel trick!

Linear Classification

Severely Linearly Inseparable

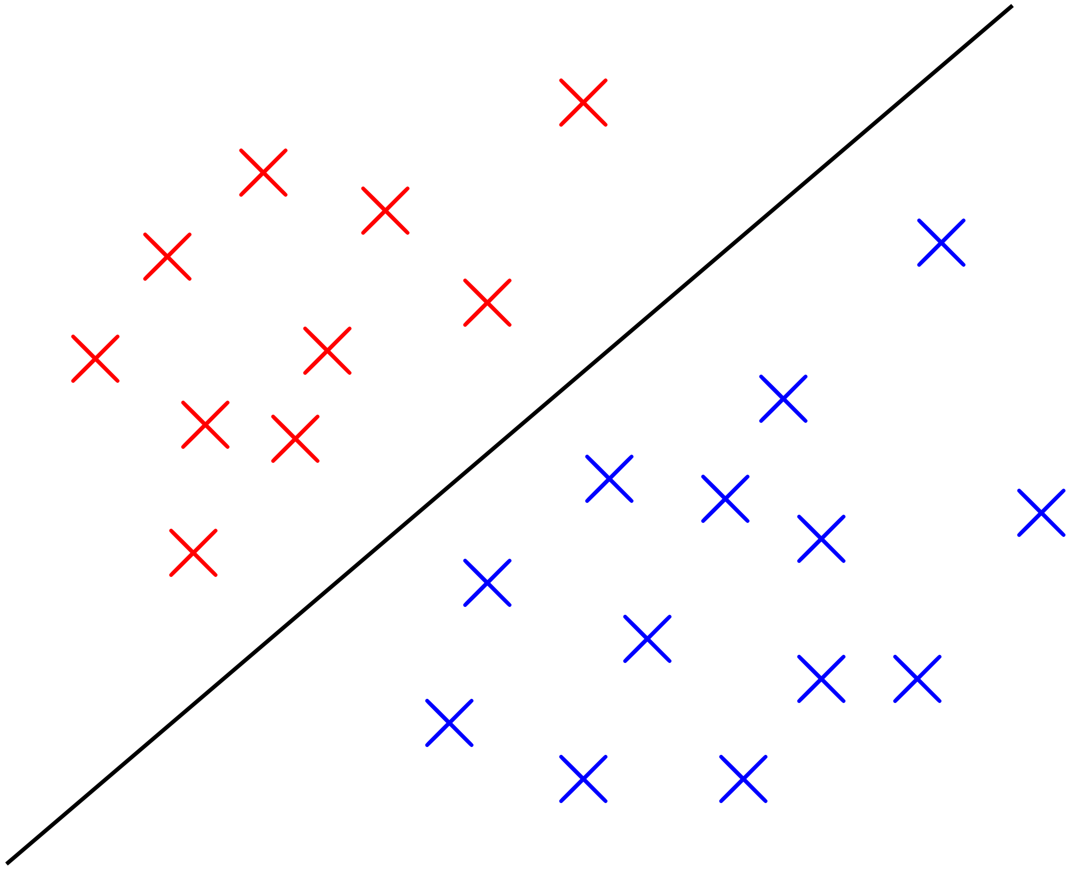


Map the data into a new space before applying linear SVM

We will have to use the kernel trick!

Linear Classification

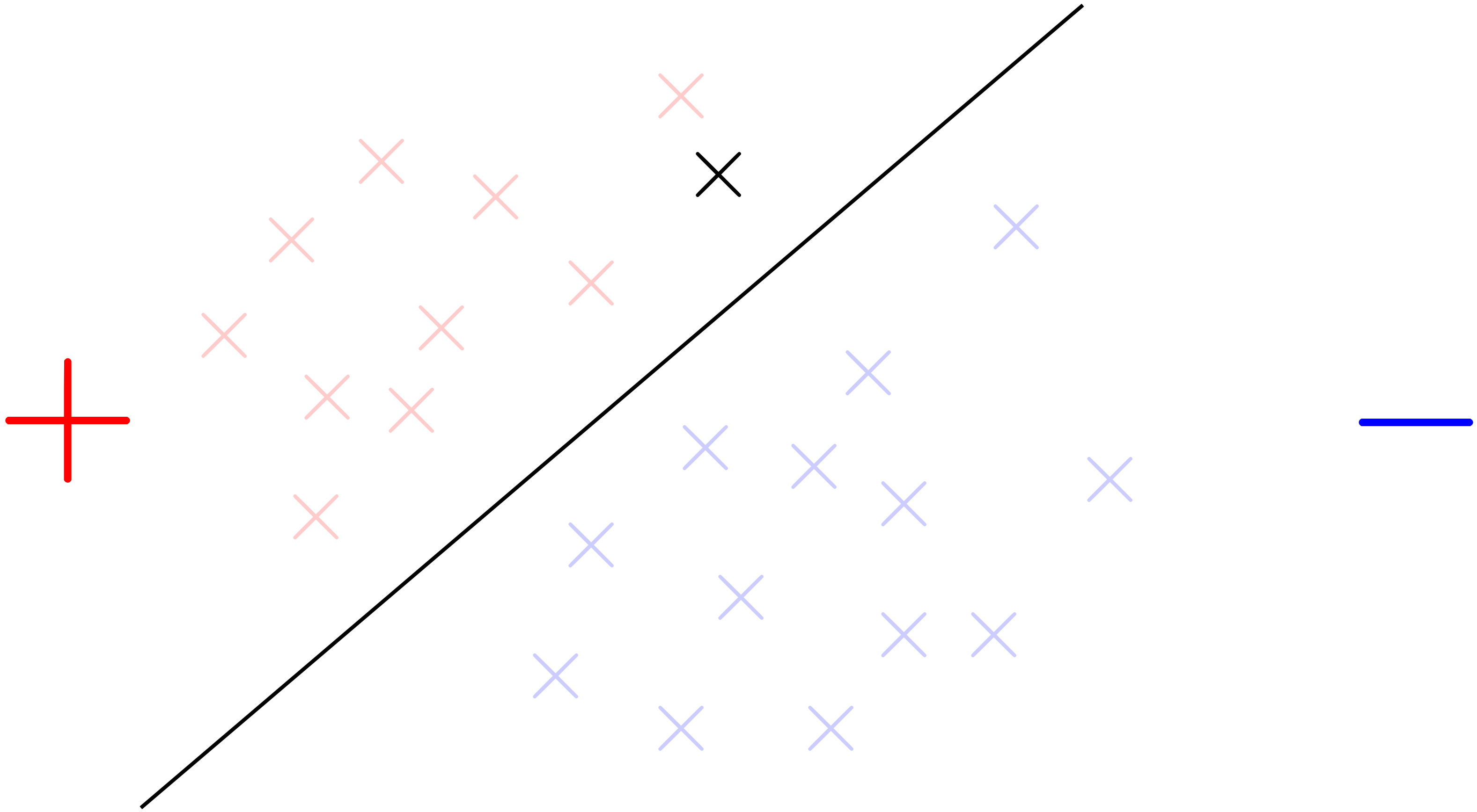
Classifier Evaluation



Assume we are done with training.

Linear Classification

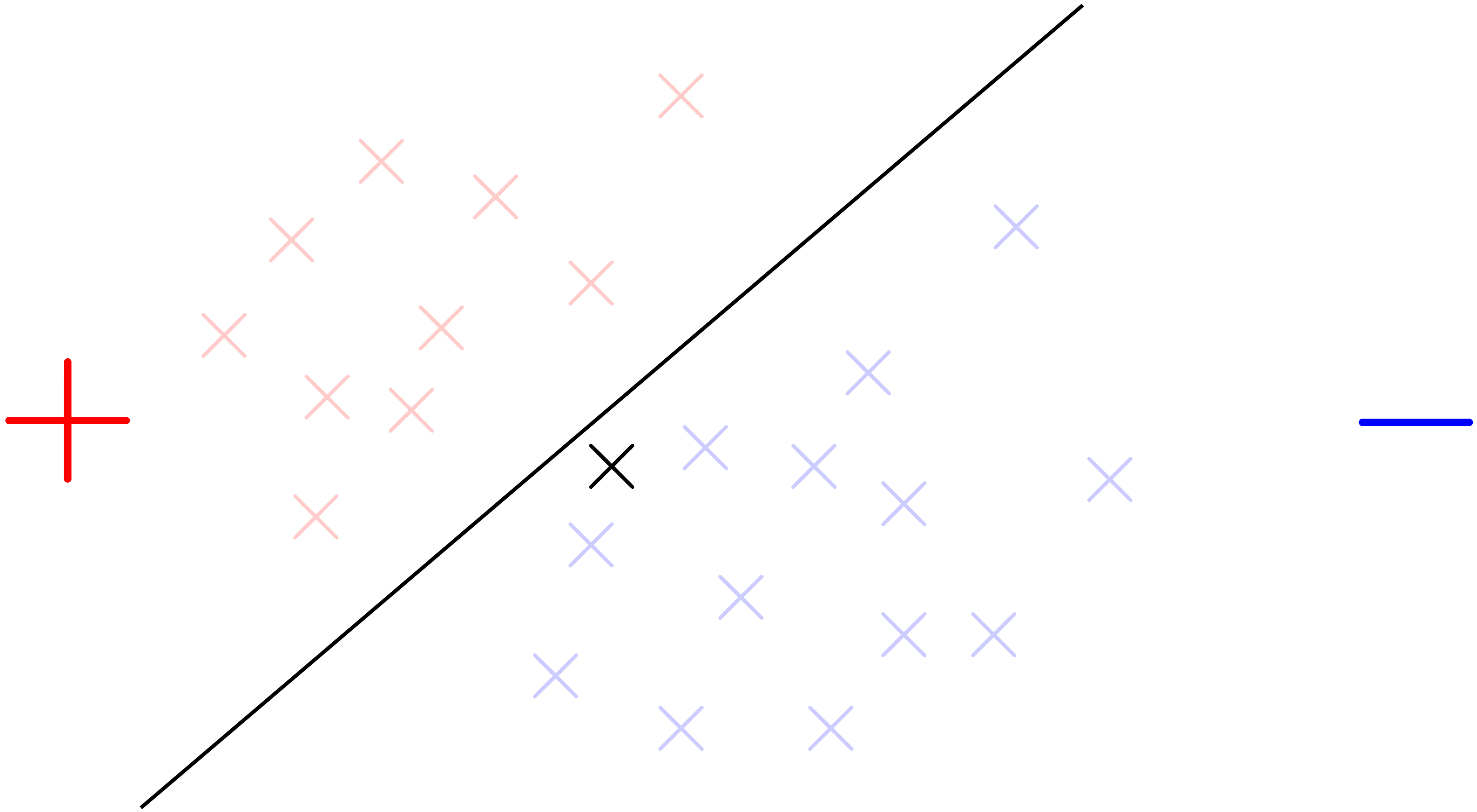
Classifier Evaluation



What should be the label for this new point?

Linear Classification

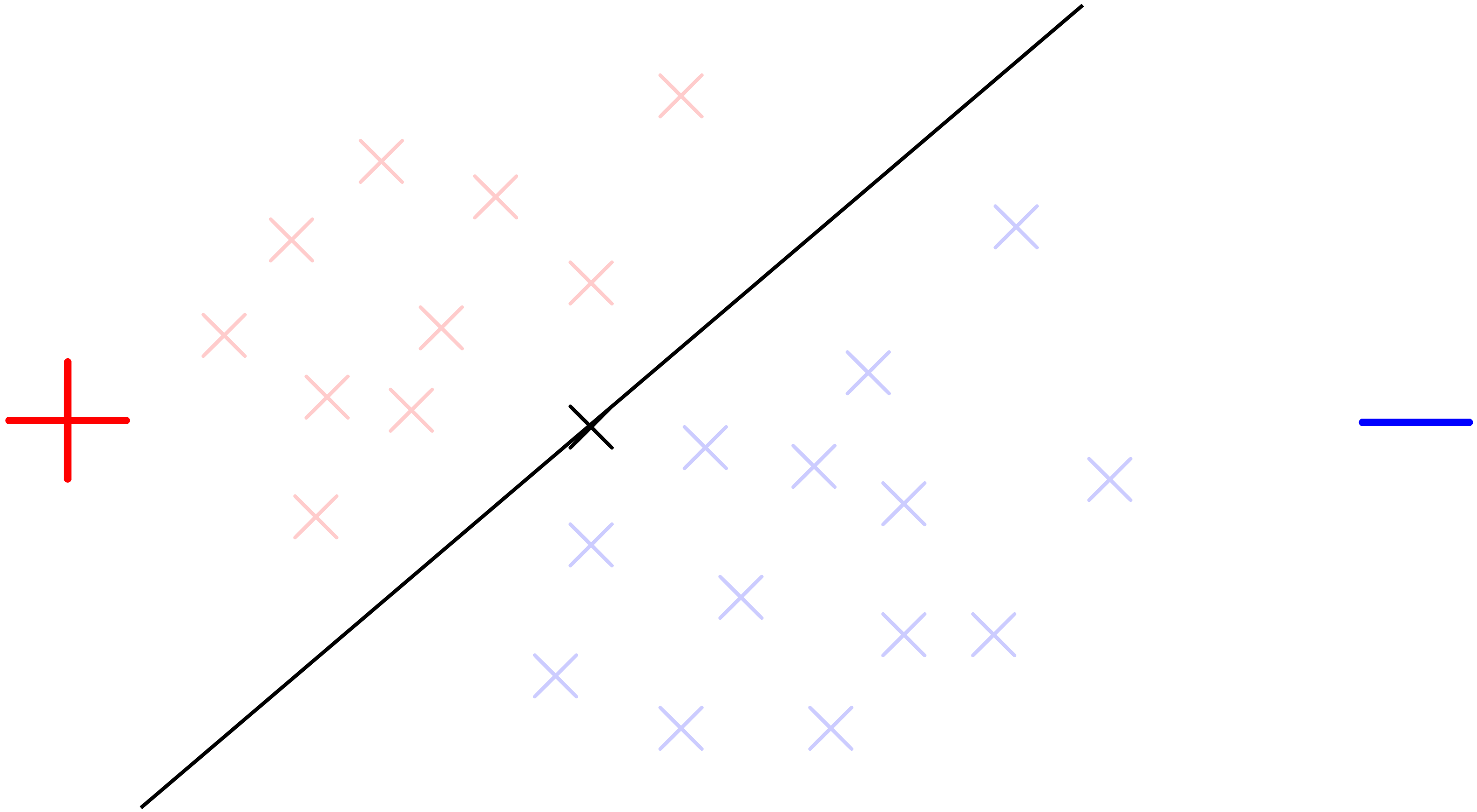
Classifier Evaluation



What should be the label for this new point?

Linear Classification

Classifier Evaluation



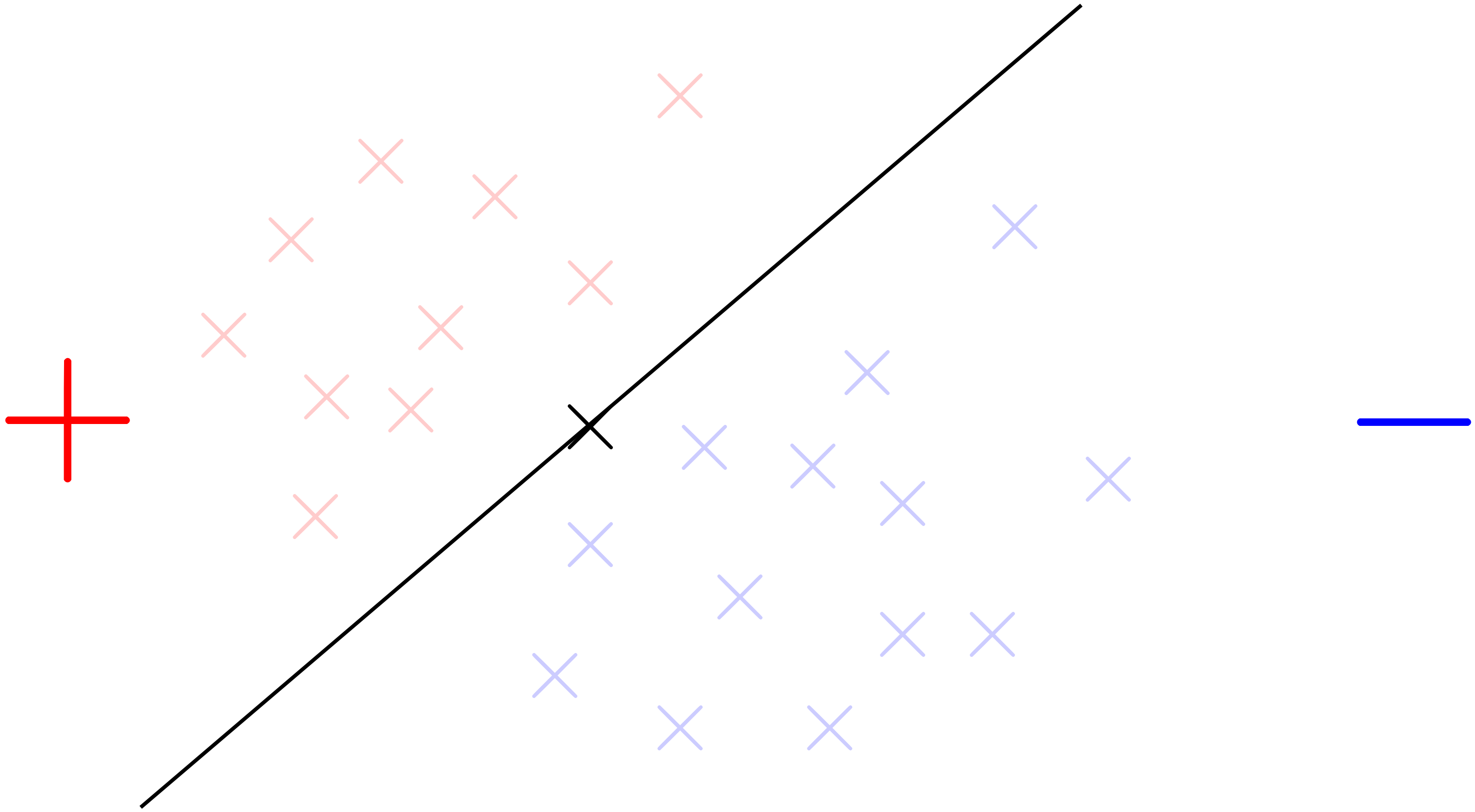
Ok, then what about this point?

Question

Is it possible to introduce the notion of confidence/probability score into the model?

Linear Classification

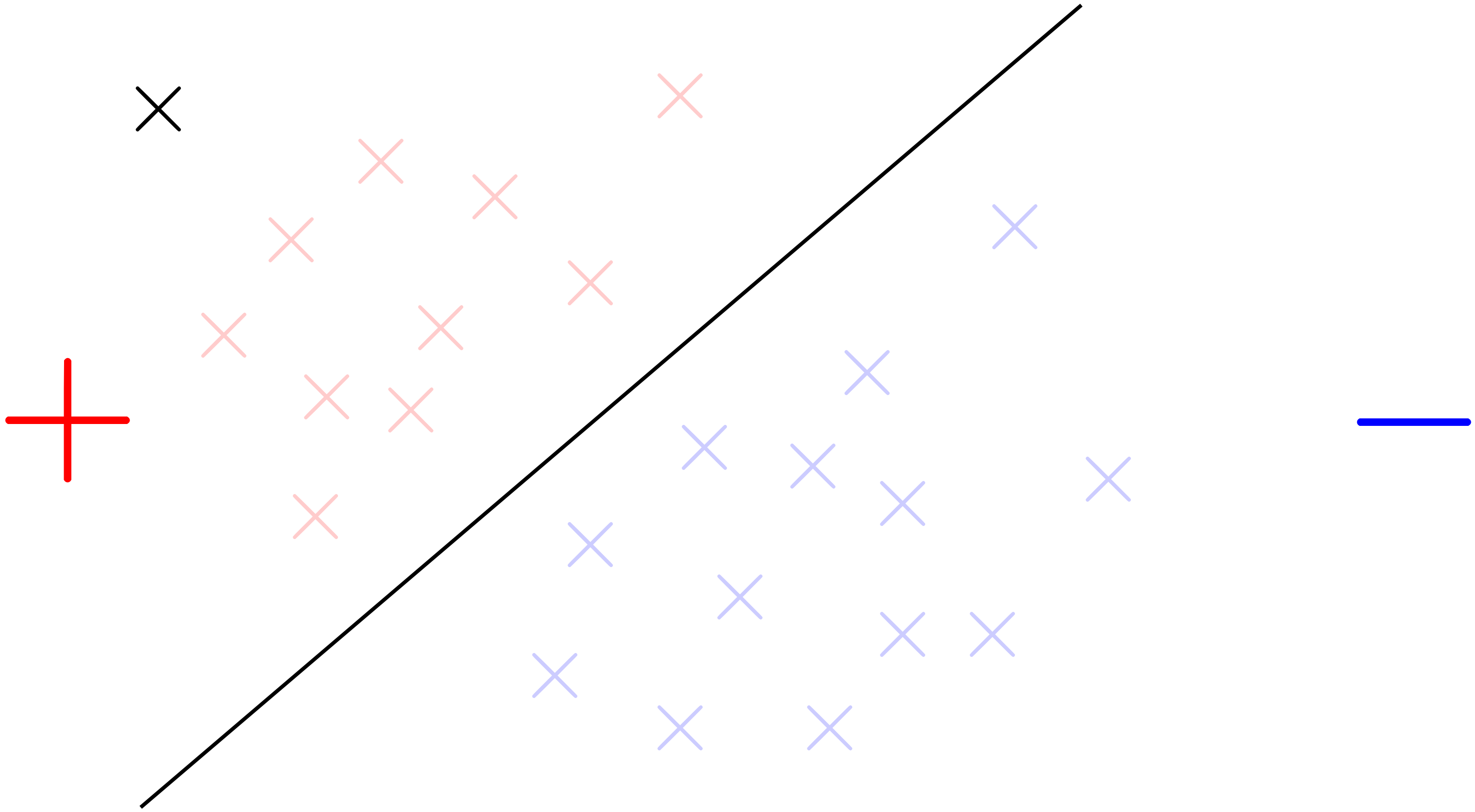
Classification with Probability



50% positive, 50% negative

Linear Classification

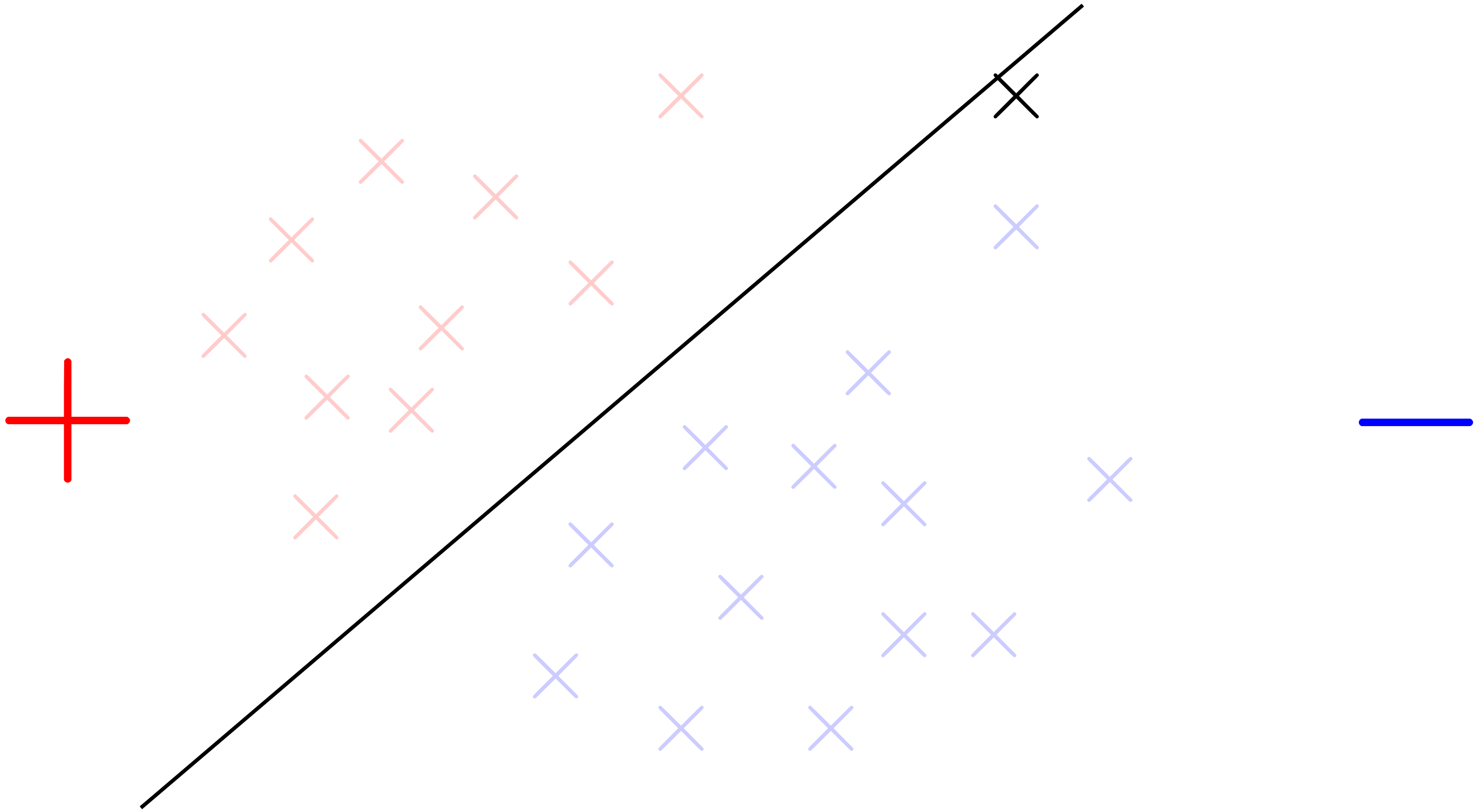
Classification with Probability



80% positive, 20% negative

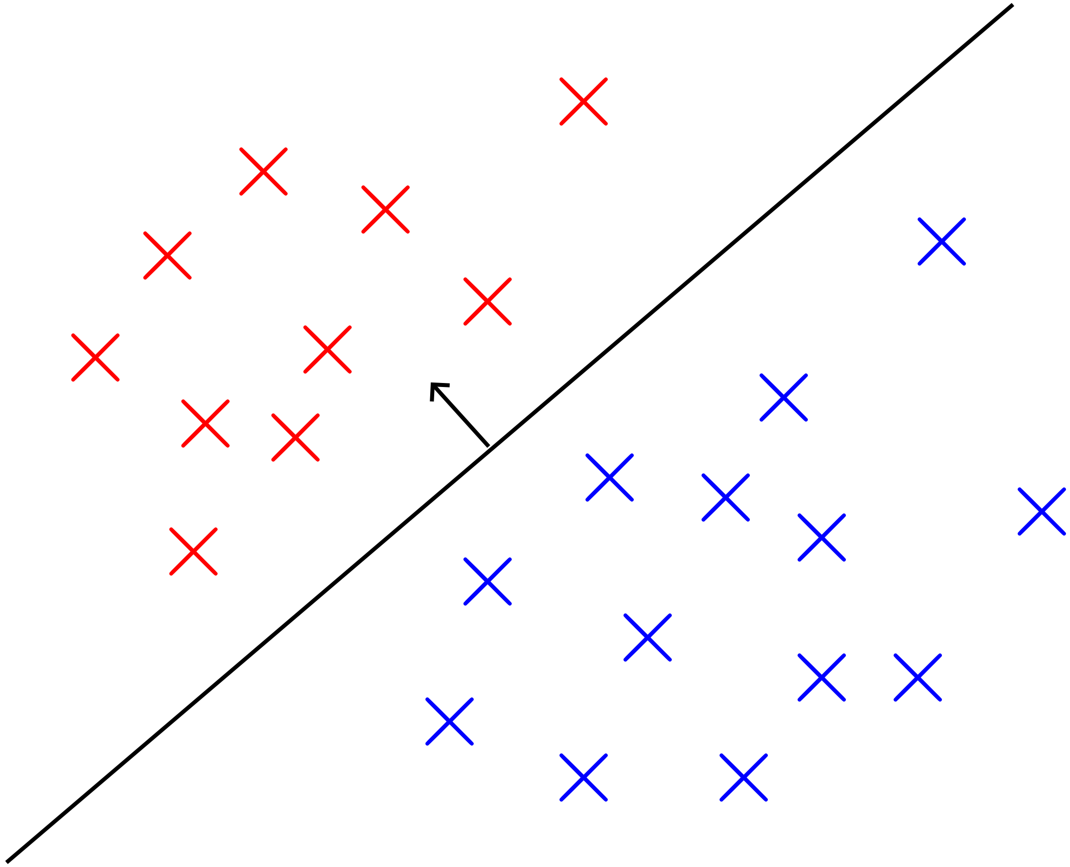
Linear Classification

Classification with Probability



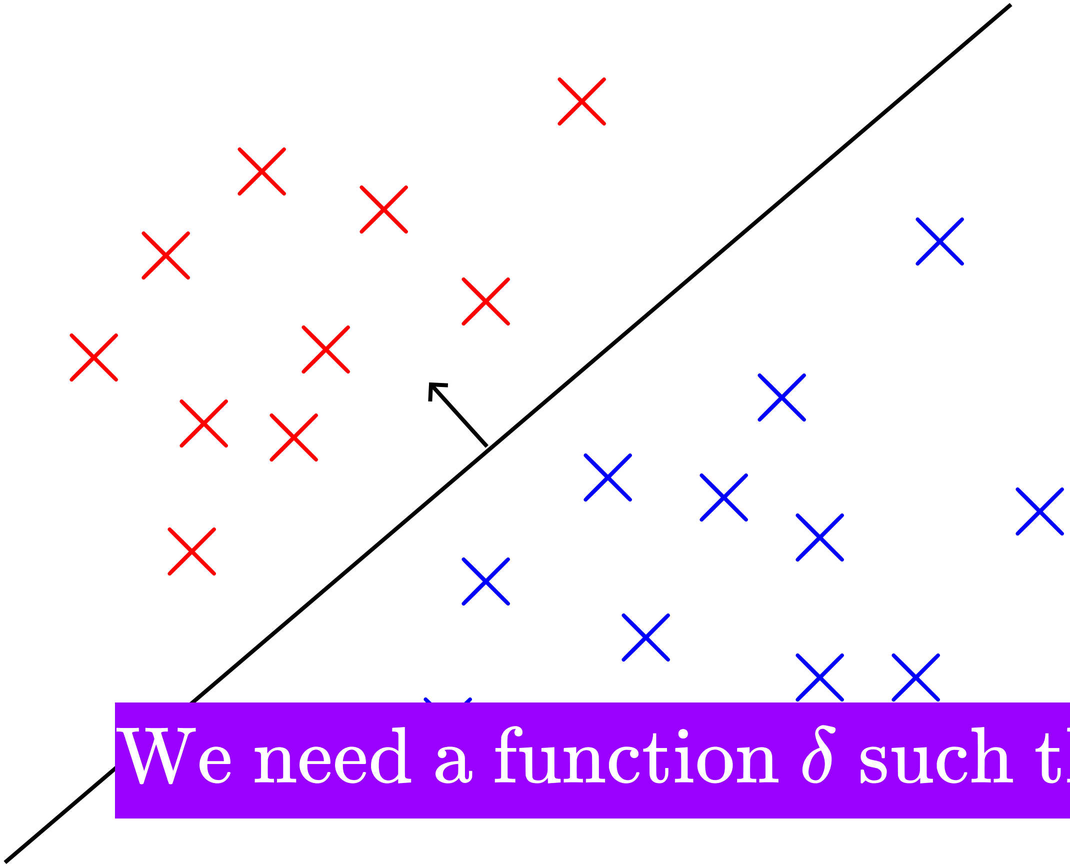
45% positive, 55% negative

Linear Classification



$$\theta \cdot x + \theta_0 = 0$$

Linear Classification



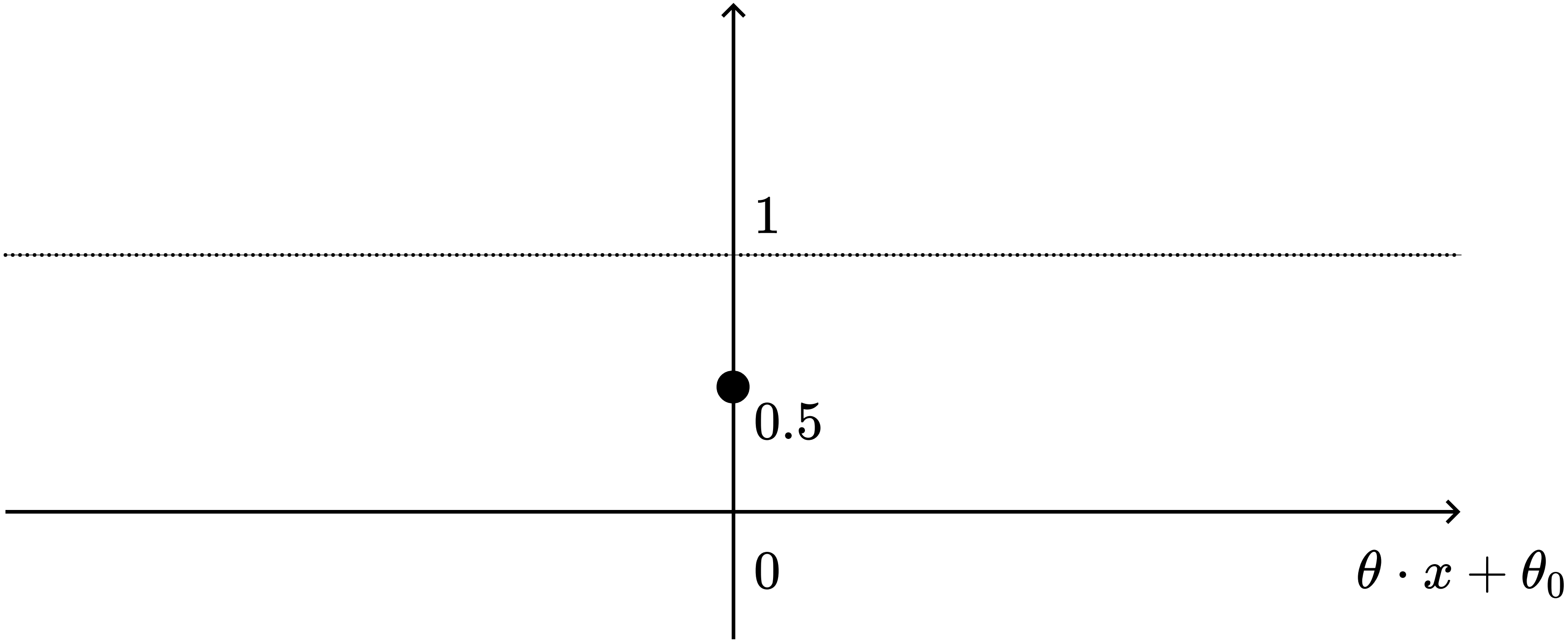
We need a function δ such that

$$\theta \cdot x + \theta_0 = 0 \qquad \delta(\theta \cdot x + \theta_0) = 0.5$$

$$\theta \cdot x + \theta_0 \to +\infty \qquad \delta(\theta \cdot x + \theta_0) \to 1.0$$

$$\theta \cdot x + \theta_0 \to -\infty \qquad \delta(\theta \cdot x + \theta_0) \to 0.0$$

Linear Classification

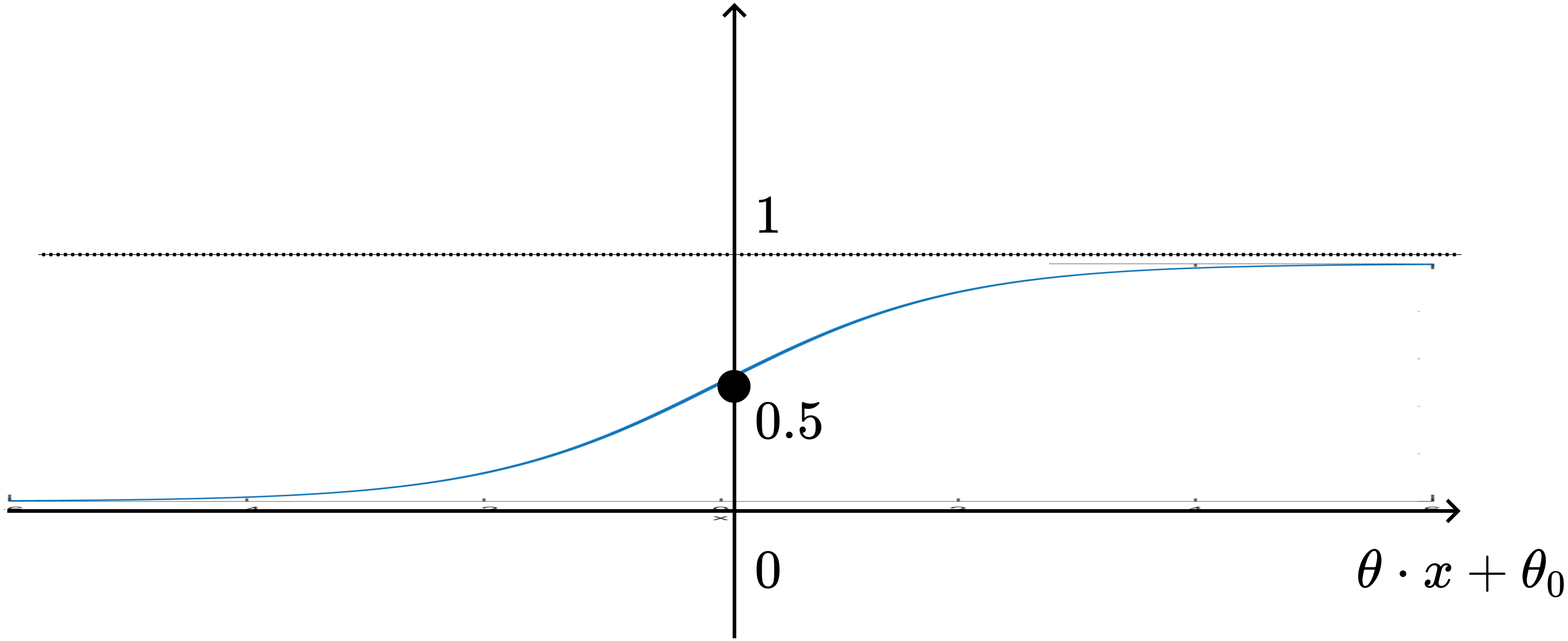


$$\theta \cdot x + \theta_0 = 0 \qquad \delta(\theta \cdot x + \theta_0) = 0.5$$

$$\theta \cdot x + \theta_0 \rightarrow +\infty \qquad \delta(\theta \cdot x + \theta_0) \rightarrow 1.0$$

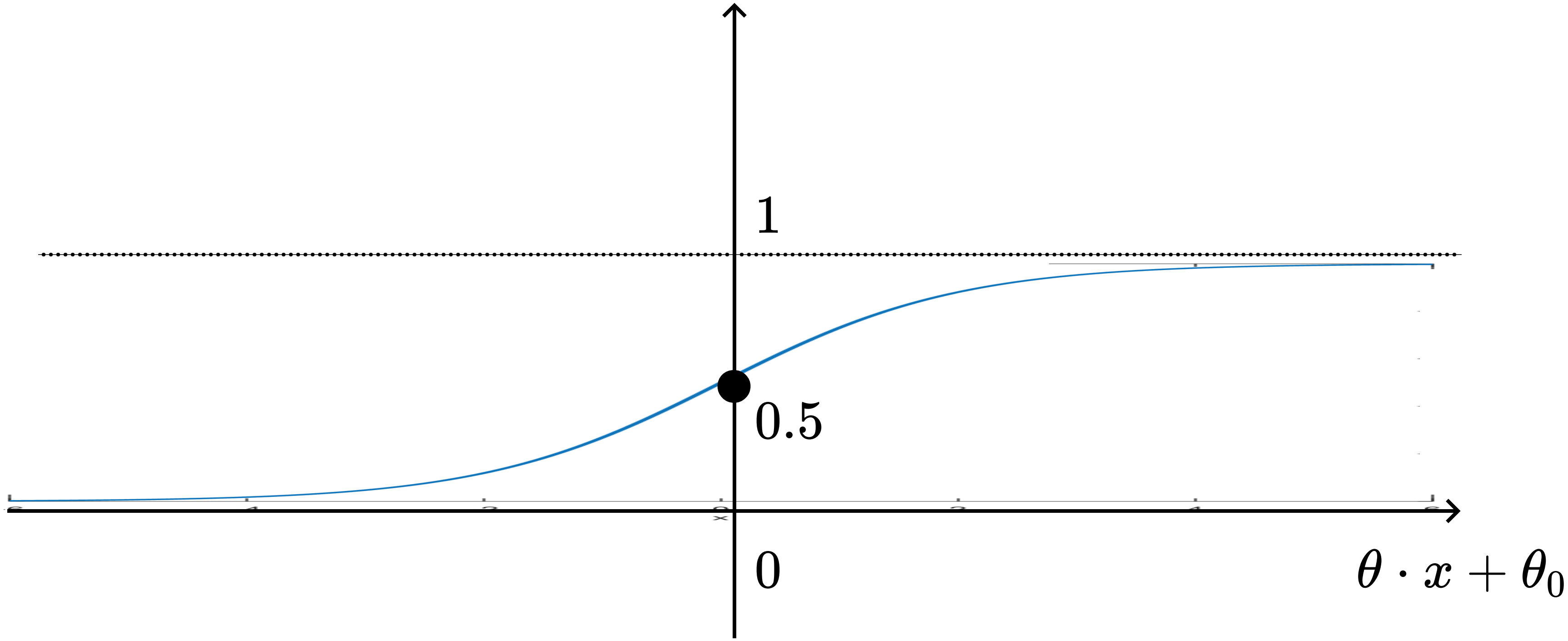
$$\theta \cdot x + \theta_0 \rightarrow -\infty \qquad \delta(\theta \cdot x + \theta_0) \rightarrow 0.0$$

Linear Classification



$$\delta(\theta \cdot x + \theta_0) = \frac{\exp(\theta \cdot x + \theta_0)}{1 + \exp(\theta \cdot x + \theta_0)}$$

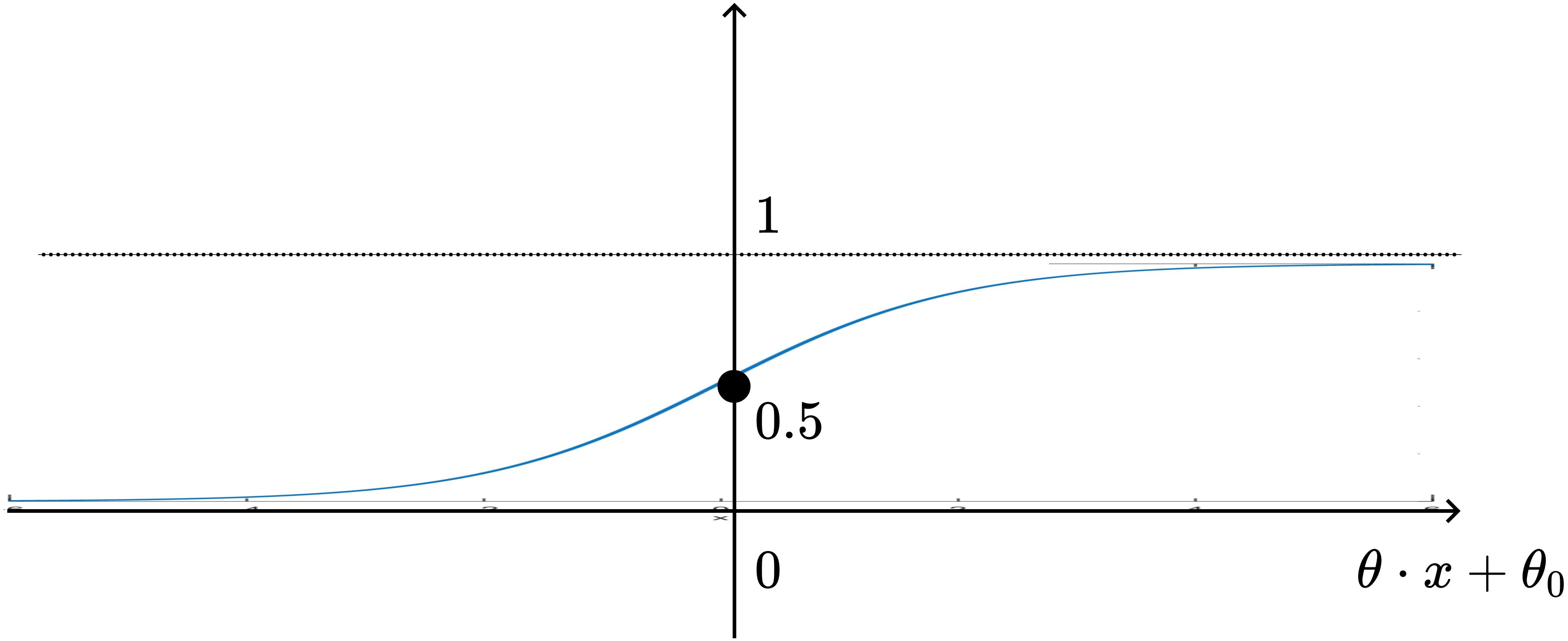
Linear Classification



$$h(x) = \frac{\exp(\theta \cdot x + \theta_0)}{1 + \exp(\theta \cdot x + \theta_0)}$$

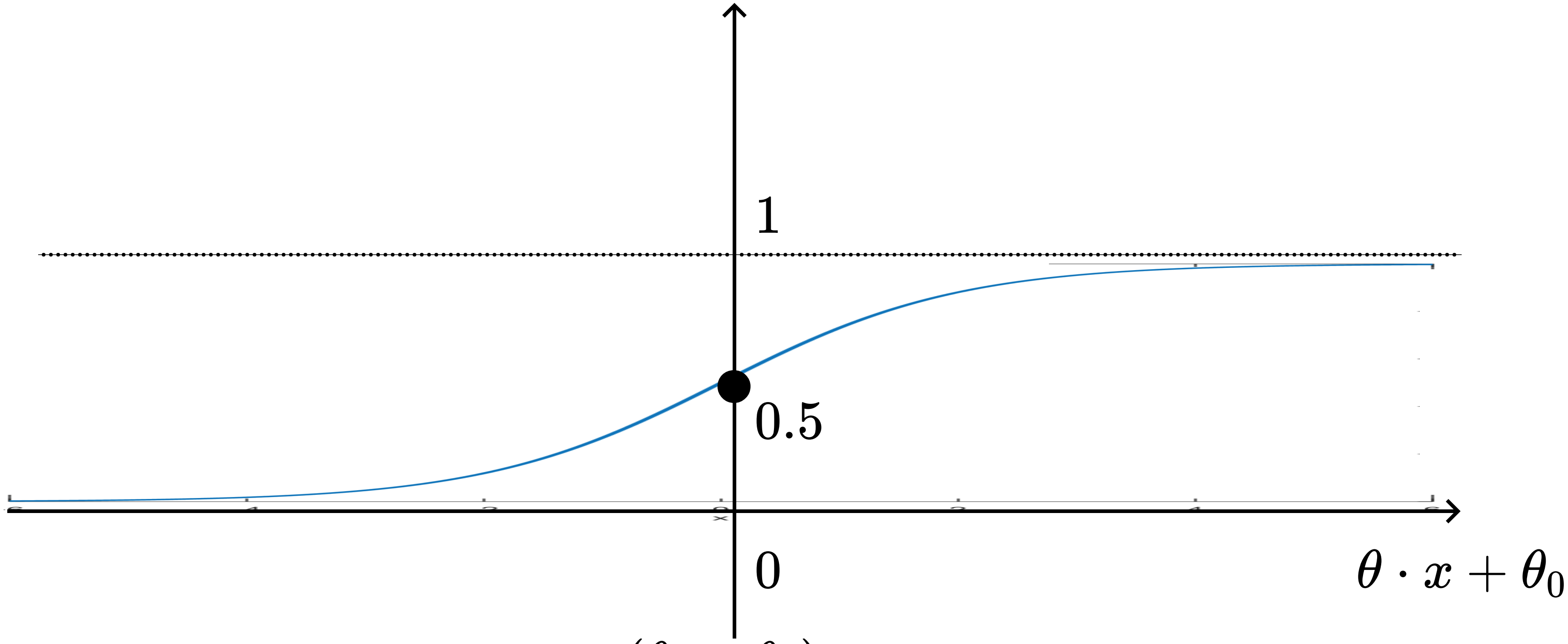
h is the probability of predicting positive $(y = +1)$

Linear Classification



$$p(y|x) = \begin{cases} h(x) & \text{for } y = +1 \\ 1 - h(x) & \text{for } y = -1 \end{cases}$$

Linear Classification

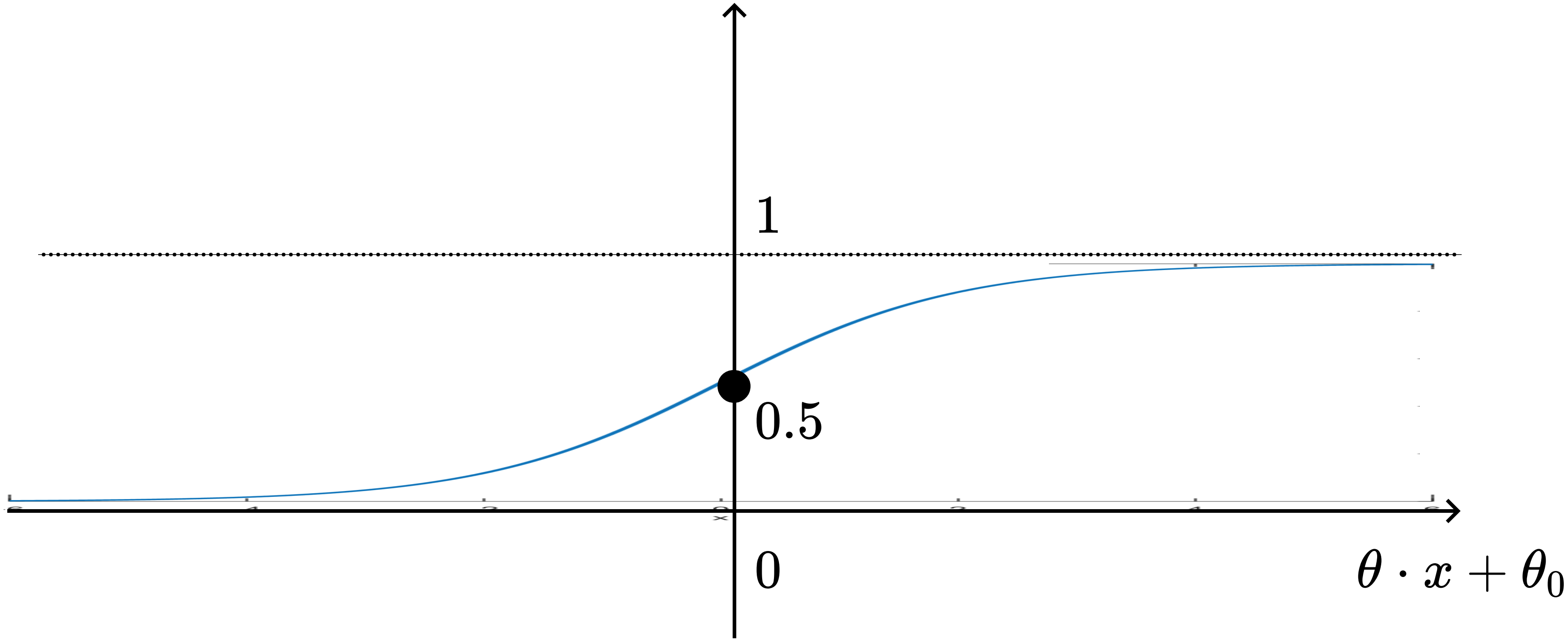


$$p(y = +1|x) = \frac{\exp(\theta \cdot x + \theta_0)}{1 + \exp(\theta \cdot x + \theta_0)} = \delta(\theta \cdot x + \theta_0)$$

$$p(y = -1|x) = \frac{1}{1 + \exp(\theta \cdot x + \theta_0)} = \frac{\exp(-(\theta \cdot x + \theta_0))}{1 + \exp(-(\theta \cdot x + \theta_0))}$$

$$= \delta(-(\theta \cdot x + \theta_0))$$

Linear Classification



$$p(y|x) = \delta(y(\theta \cdot x + \theta_0))$$

Linear Classification

Objective Function

Training set examples: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

What shall we optimize?

Linear Classification

Objective Function

Training set examples: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

$$\max_{\theta,\theta_0} \qquad \prod_{i=1}^{n} p(y^{(i)}|x^{(i)})$$

Linear Classification

Objective Function

Training set examples: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

$$\max_{\theta,\theta_0} \quad \prod_{i=1}^{n} p(y^{(i)}|x^{(i)})$$

$$\max_{\theta,\theta_0} \quad \log \prod_{i=1}^{n} p(y^{(i)}|x^{(i)})$$

$$\max_{\theta,\theta_0} \quad \sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)})$$

Linear Classification

Loss Function

Training set examples: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

$$\max_{\theta,\theta_0} \quad \prod_{i=1}^{n} p(y^{(i)}|x^{(i)})$$

$$\max_{\theta,\theta_0} \quad \log \prod_{i=1}^{n} p(y^{(i)}|x^{(i)})$$

$$\max_{\theta,\theta_0} \quad \sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)})$$

$$\min_{\theta,\theta_0} \quad \sum_{i=1}^{n} \log 1/p(y^{(i)}|x^{(i)})$$

Linear Classification

Loss Function

Training set examples: $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$

$$\textstyle\sum_{i=1}^{n} \log 1/p(y^{(i)}|x^{(i)})$$

$$\textstyle\sum_{i=1}^{n} \log \left(1 + \exp(-y^{(i)}(\theta \cdot x^{(i)} + \theta_0)\right)$$

What is the benefit of using logarithm? Why is this expression computationally more "convenient"?

Logistic Regression

Loss Function

$$\log\left(1+\exp(-y^{(t)}(\theta\cdot x^{(t)}+\theta_0))\right)$$

Hinge Loss

$$\max(0, 1-y^{(t)}(\theta\cdot x^{(t)}+\theta_0))$$

See the whiteboard to know the connections between the two loss functions.

Logistic Regression

Learning

Let us drop θ_0 for now:

$$e^{(t)}(\theta) = \log\left(1 + \exp(-y^{(t)}(\theta \cdot x^{(t)}))\right)$$

Logistic Regression

Learning

Let us drop θ_0 for now:

$$e^{(t)}(\theta) = \log\left(1 + \exp(-y^{(t)}(\theta \cdot x^{(t)})\right)$$

$$\nabla e^{(t)}(\theta) = \frac{-y^{(t)}x^{(t)}}{1+\exp(y^{(t)}(\theta \cdot x^{(t)}))}$$

Logistic Regression

Learning

Let us drop θ_0 for now:

$$e^{(t)}(\theta) = \log\left(1 + \exp(-y^{(t)}(\theta \cdot x^{(t)})\right)$$

$$\nabla e^{(t)}(\theta) = \frac{-y^{(t)} x^{(t)}}{1+\exp(y^{(t)}(\theta \cdot x^{(t)}))}$$

$$\theta \leftarrow \theta - \eta \nabla e^{(t)}(\theta)$$

Logistic Regression

Learning

Let us drop θ_0 for now:

$$e^{(t)}(\theta) = \log\left(1 + \exp(-y^{(t)}(\theta \cdot x^{(t)})\right)$$

$$\nabla e^{(t)}(\theta) = \frac{-y^{(t)}x^{(t)}}{1+\exp(y^{(t)}(\theta \cdot x^{(t)}))}$$

$$\theta \leftarrow \theta - \eta \nabla e^{(t)}(\theta)$$

What if we include θ_0? Can you write down the complete stochastic gradient descent procedure?

Logistic Regression

Prediction

We now have a new input x

How shall we predict the output label?

Logistic Regression

Prediction

We now have a new input x

$$p(y = +1|x)$$

$$p(y = -1|x)$$

Logistic Regression

Prediction

We now have a new input x

$$p(y=+1|x)$$

$$\vee \ ?$$

$$p(y=-1|x)$$

If yes, positive, otherwise negative!

Logistic Regression

Prediction

$\text{We now have a new input } x$

$$\frac{p(y=+1|x)}{p(y=-1|x)} \quad > \quad 1\,?$$

If yes, positive, otherwise negative!

Logistic Regression

Prediction

We now have a new input x

$$\log \frac{p(y=+1|x)}{p(y=-1|x)} > 0\ ?$$

If yes, positive, otherwise negative!

Logistic Regression

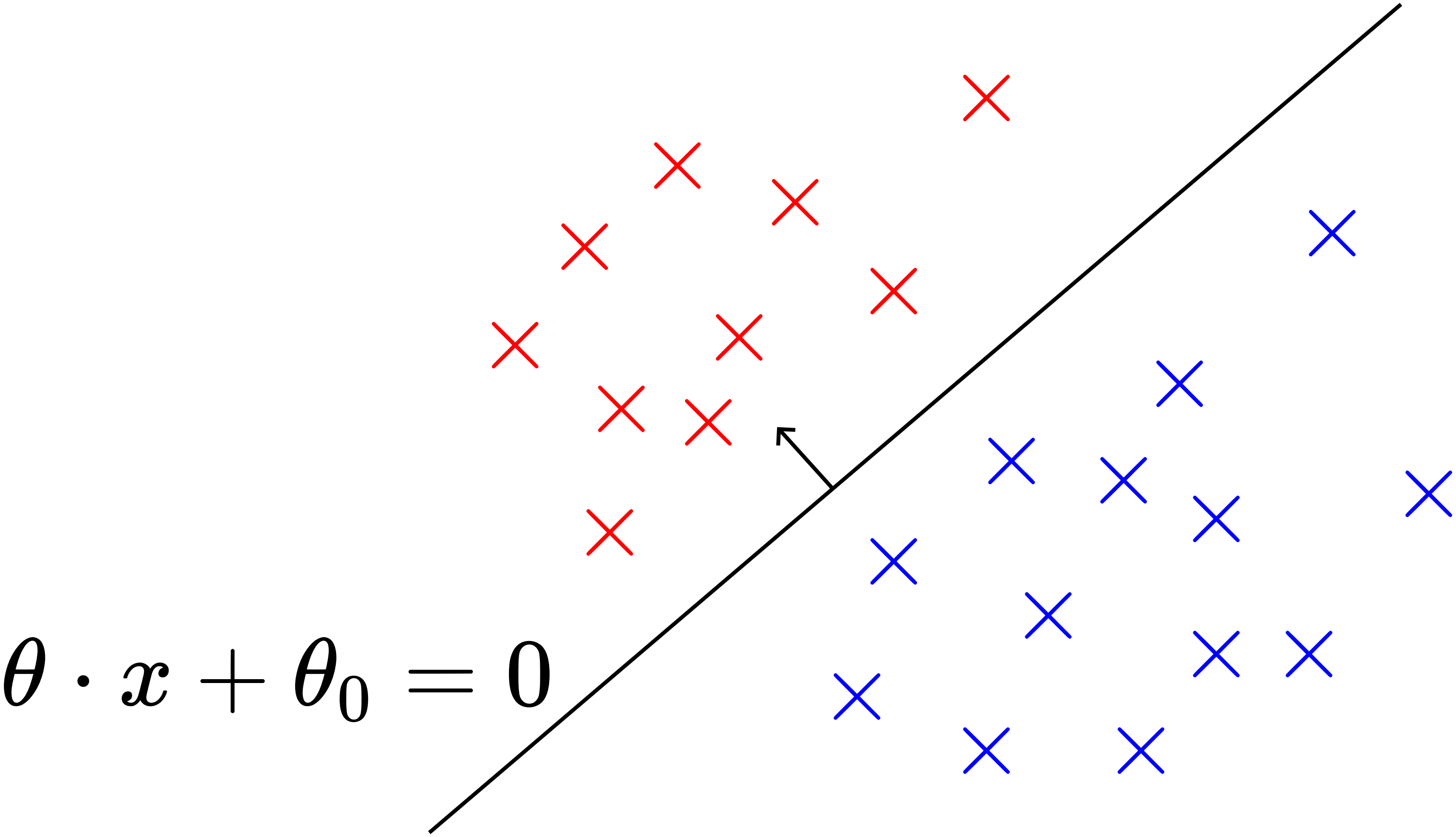
Prediction

We now have a new input x

$$\log \frac{P(y=+1|x)}{P(y=-1|x)} = \log \exp(\theta \cdot x + \theta_0) = \theta \cdot x + \theta_0$$

Note that this is a linear function. We shall check if this value is larger than 0. In other words, we still arrived at a linear decision boundary (but using a different approach)

Logistic Regression



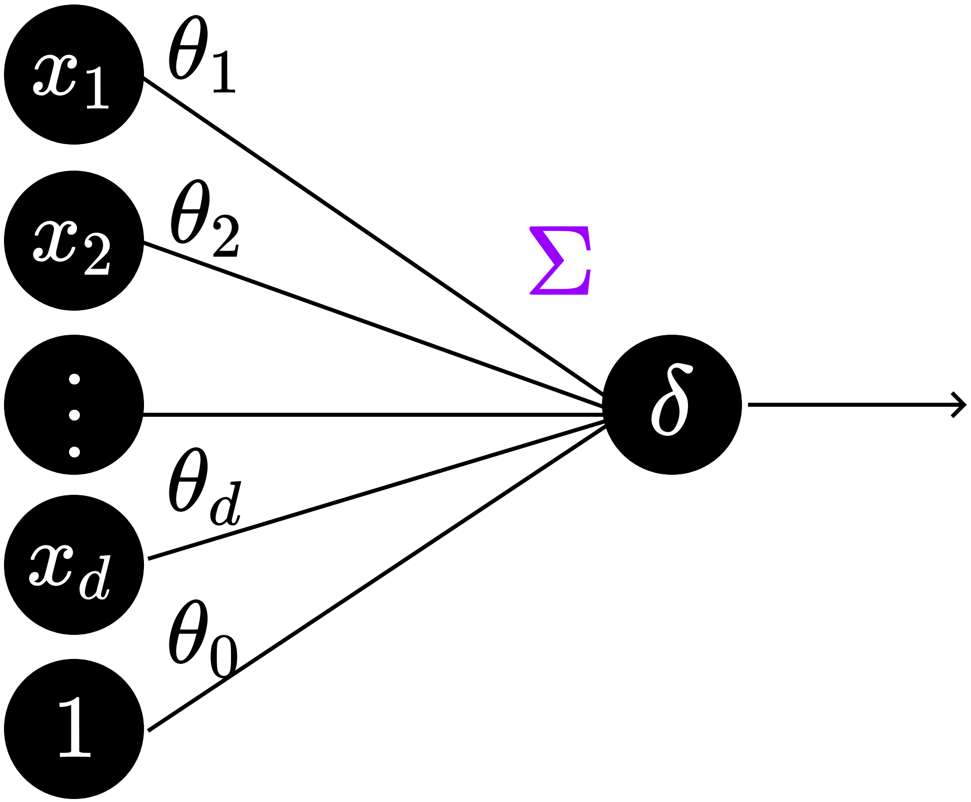
Note that this is a linear function. We shall check if this value is larger than 0. In other words, we still arrived at a linear decision boundary (but using a different approach)

Logistic Regression

$$h(x) = \frac{\exp(\theta \cdot x + \theta_0)}{1 + \exp(\theta \cdot x + \theta_0)}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

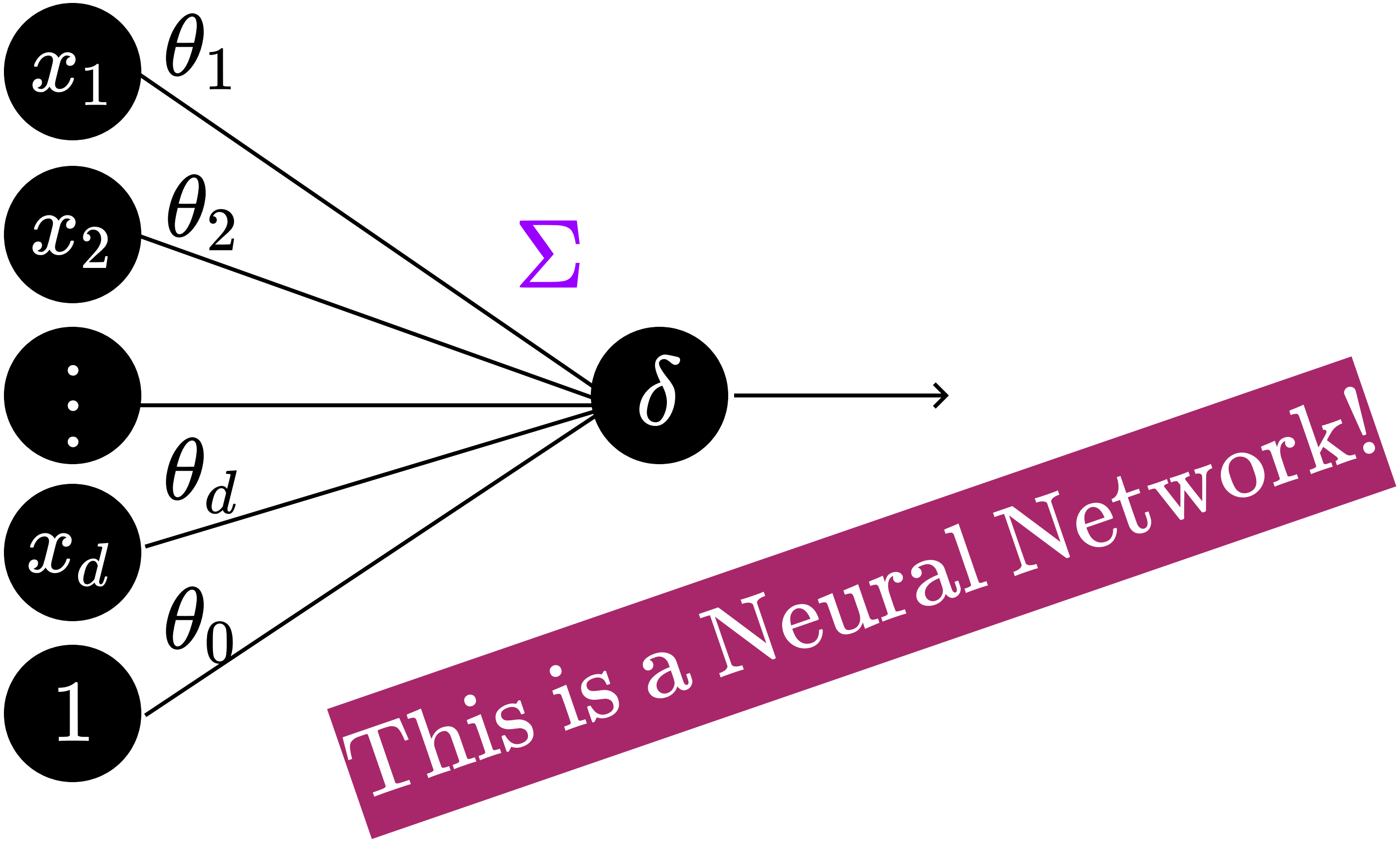
$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_d \end{bmatrix}$$



There is another way to interpret the above function

Logistic Regression

$$h(x) = \frac{\exp(\theta\cdot x + \theta_0)}{1+\exp(\theta\cdot x+\theta_0)}$$



There is another way to interpret the above function