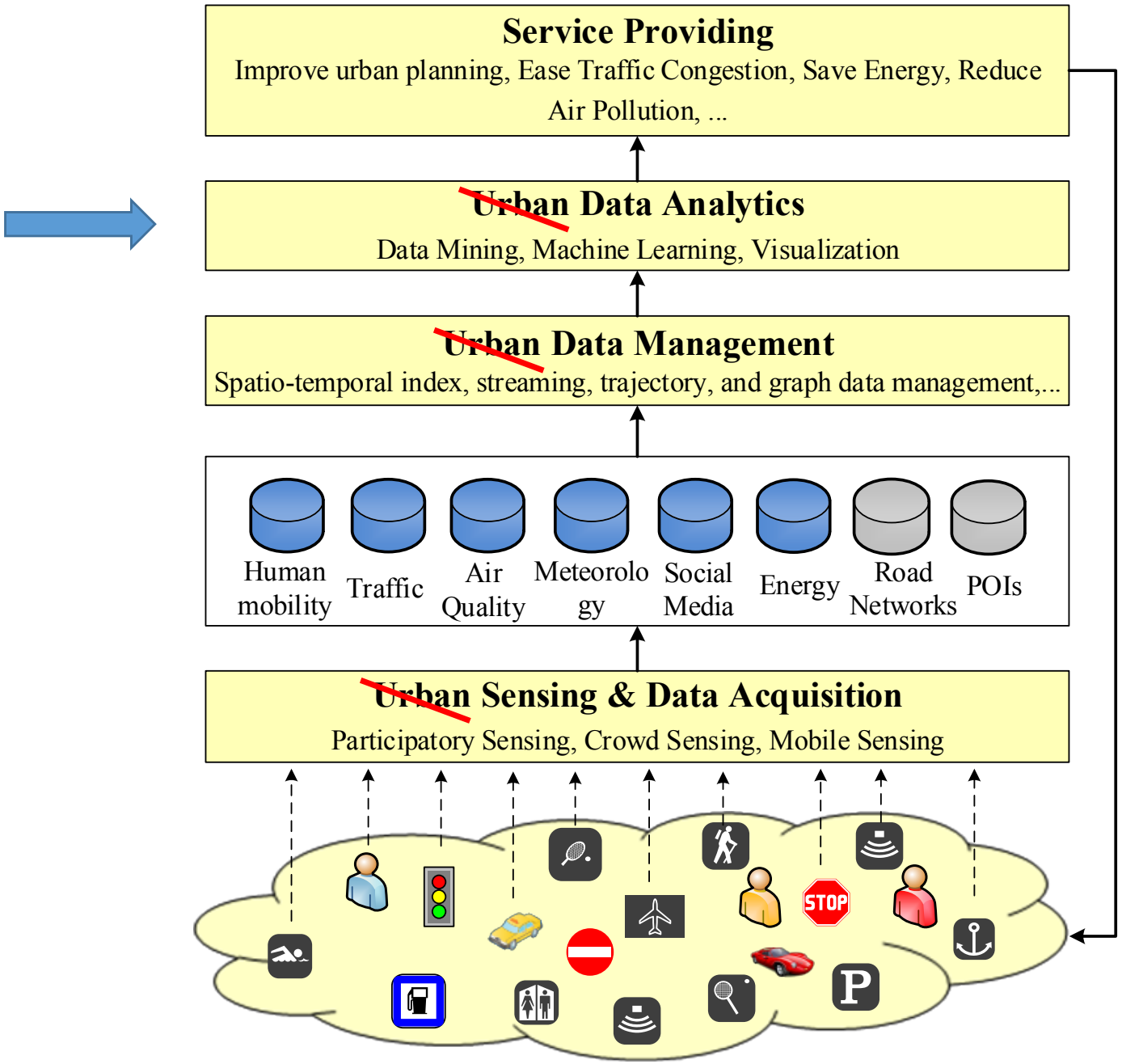
Welcome to

DS3010:
DS-III: Computational Data Intelligence
Classification

Prof. Yanhua Li

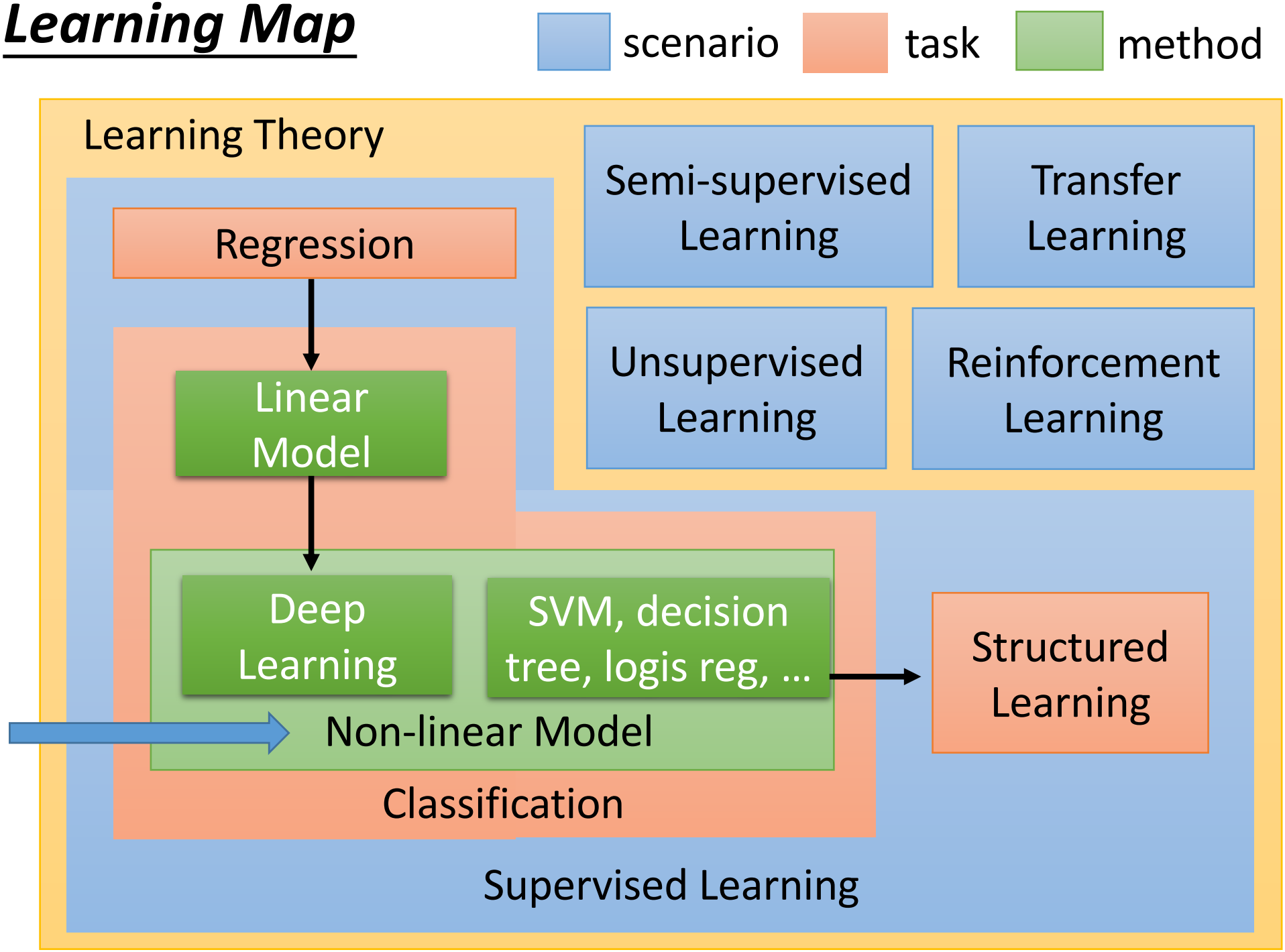
Time: 11:00am – 12:50pm M & R
Location: HL 114
D-term 2022

Data pipeline



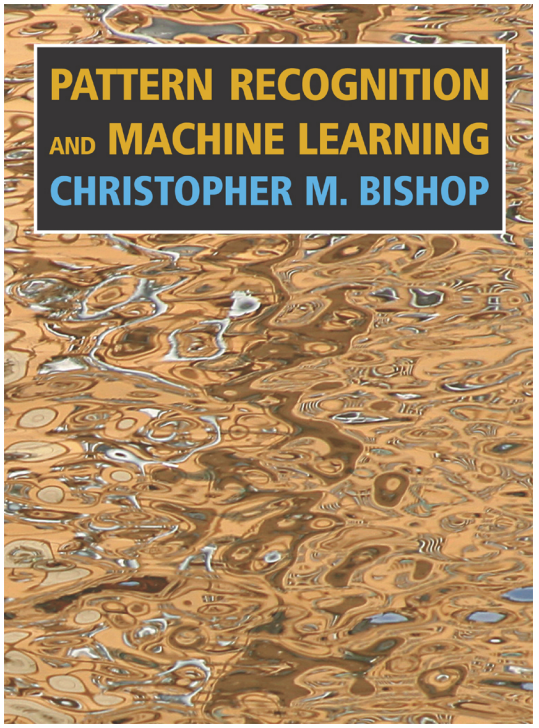
Urban Computing: concepts, methodologies, and applications.
Zheng, Y., et al. *ACM transactions on Intelligent Systems and Technology*.

Learning Map



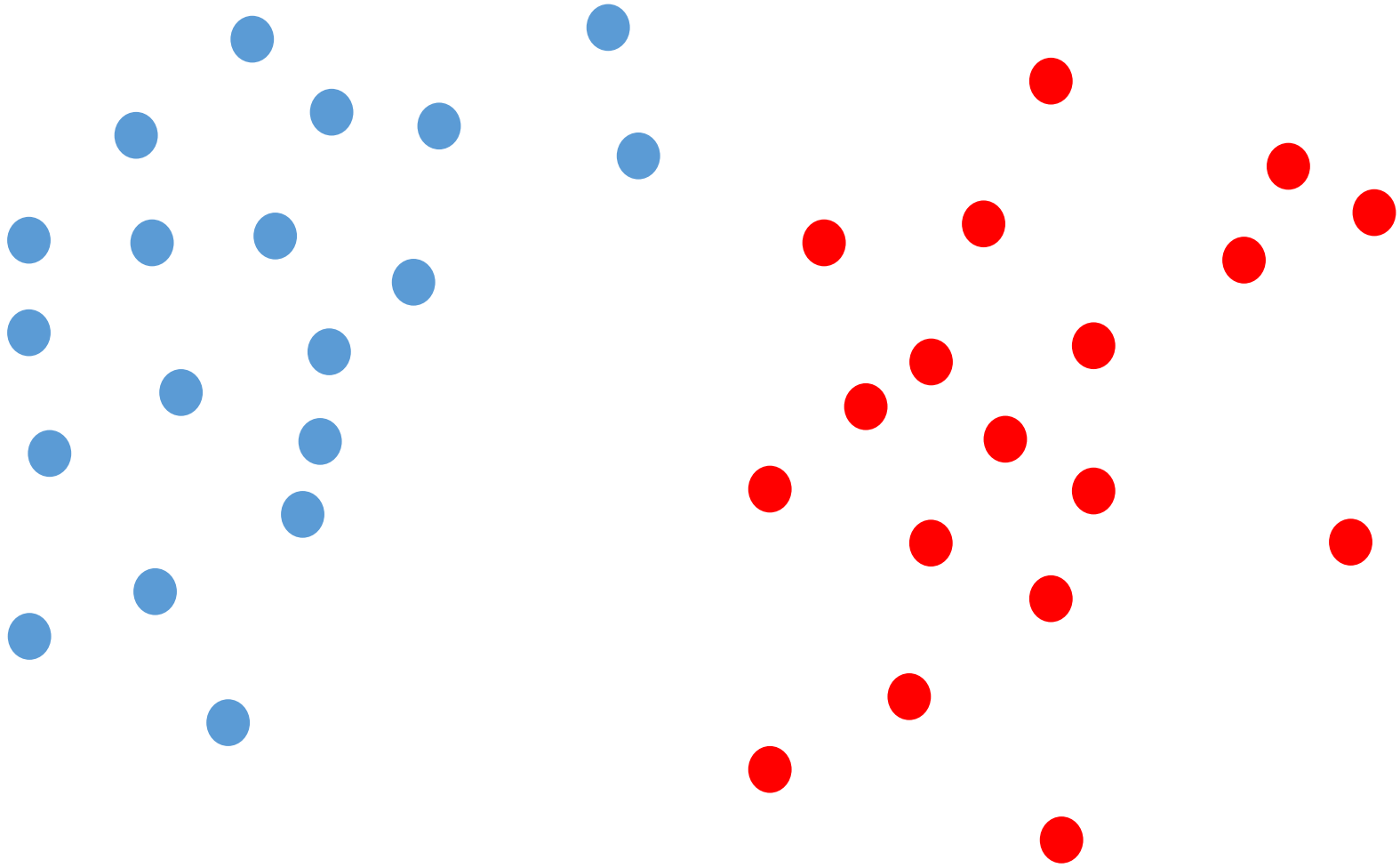
References

Classification
SVM (a much simplified discussion)

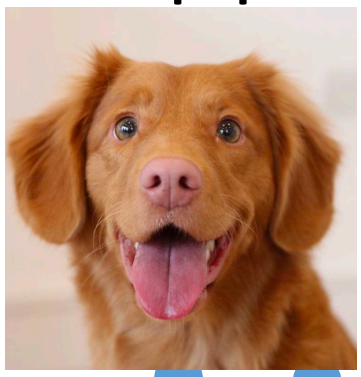


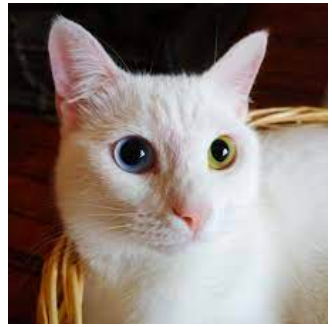
Bishop: Chapter 7.1

Support Vector Machine -- SVM

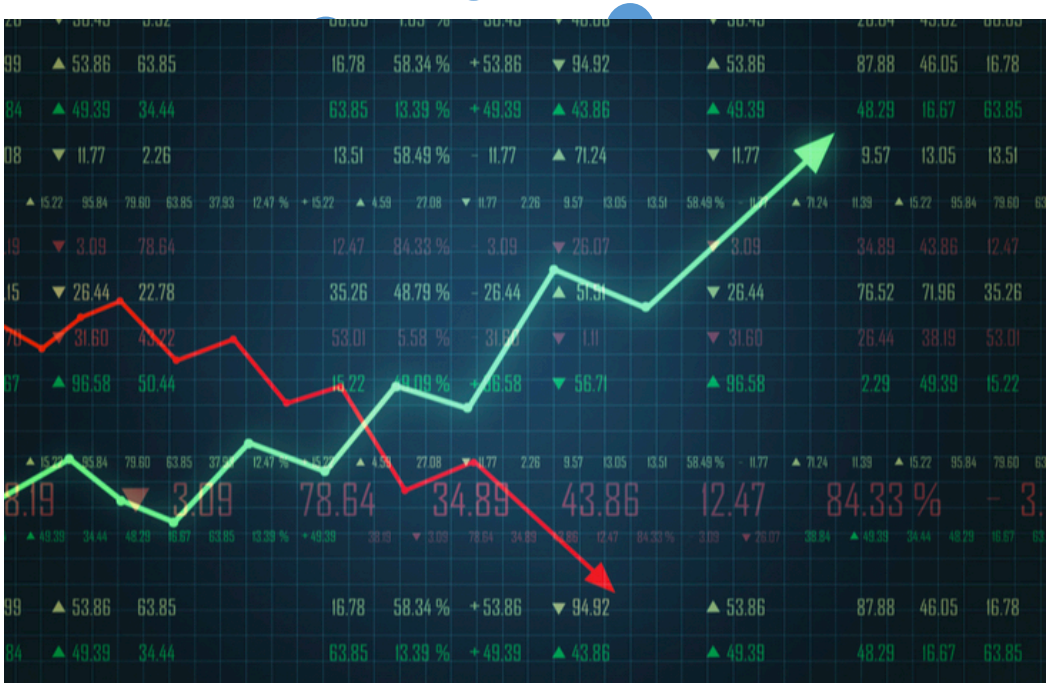


Support Vector Machine -- SVM

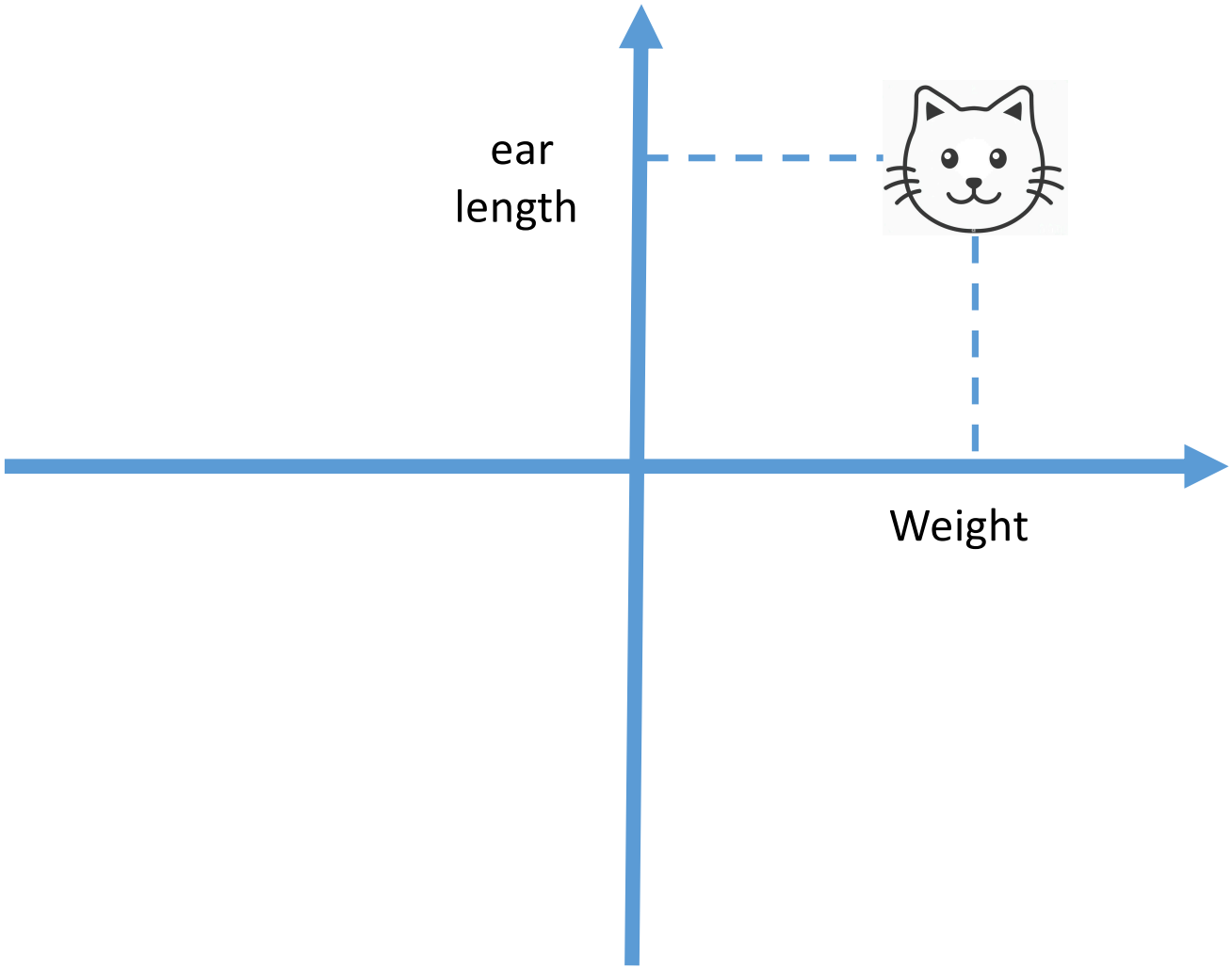




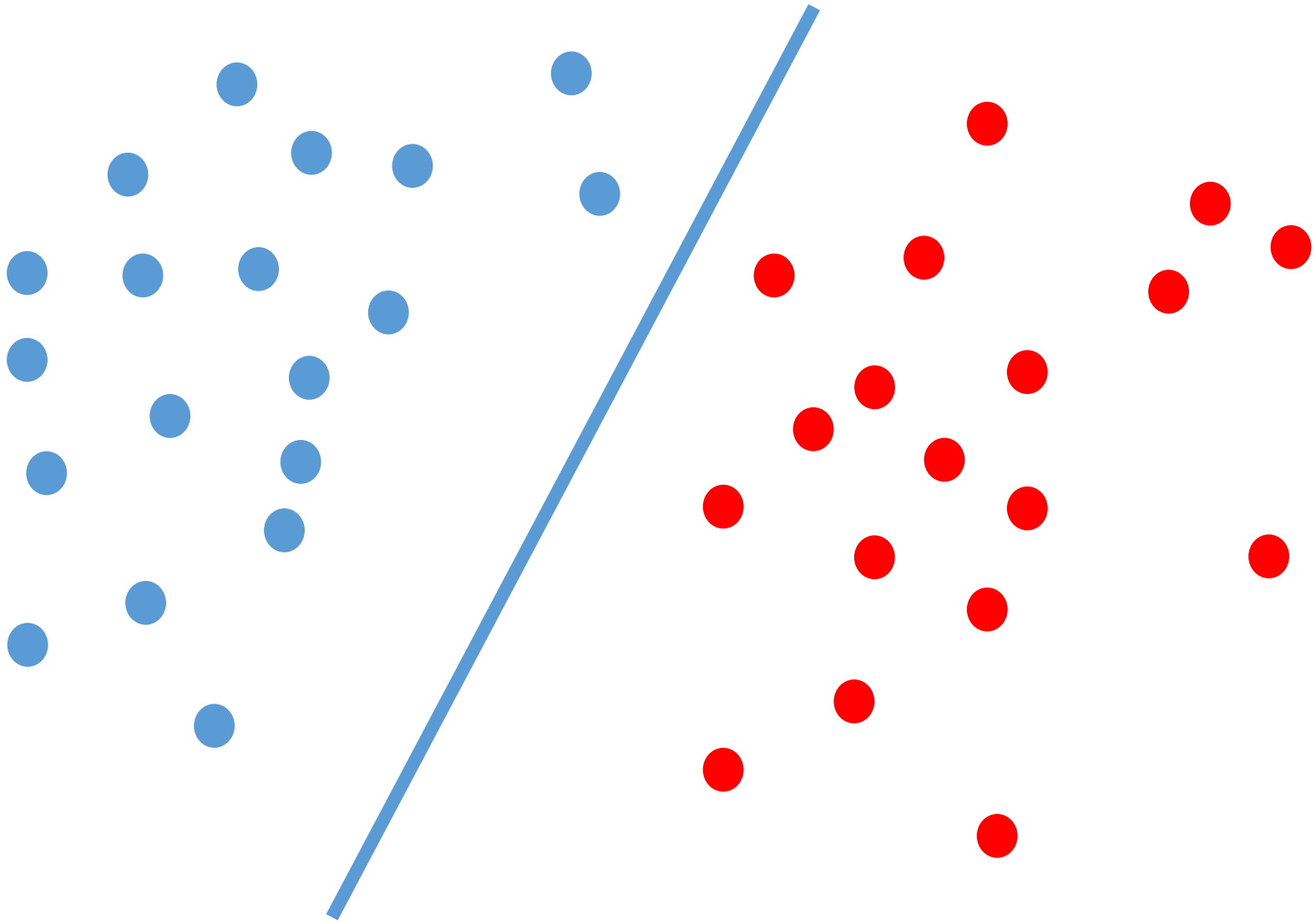
Support Vector Machine -- SVM



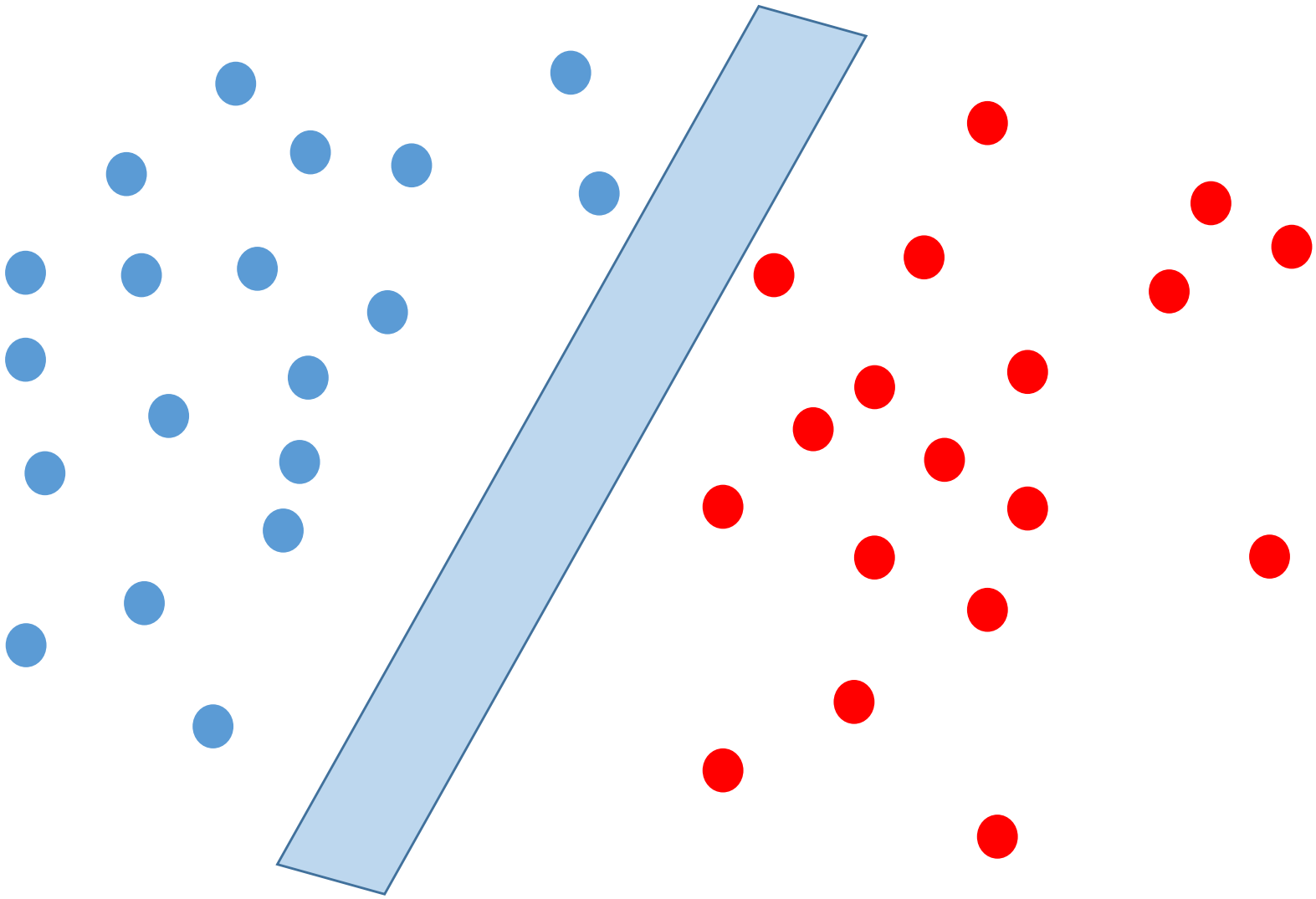
Support Vector Machine -- SVM



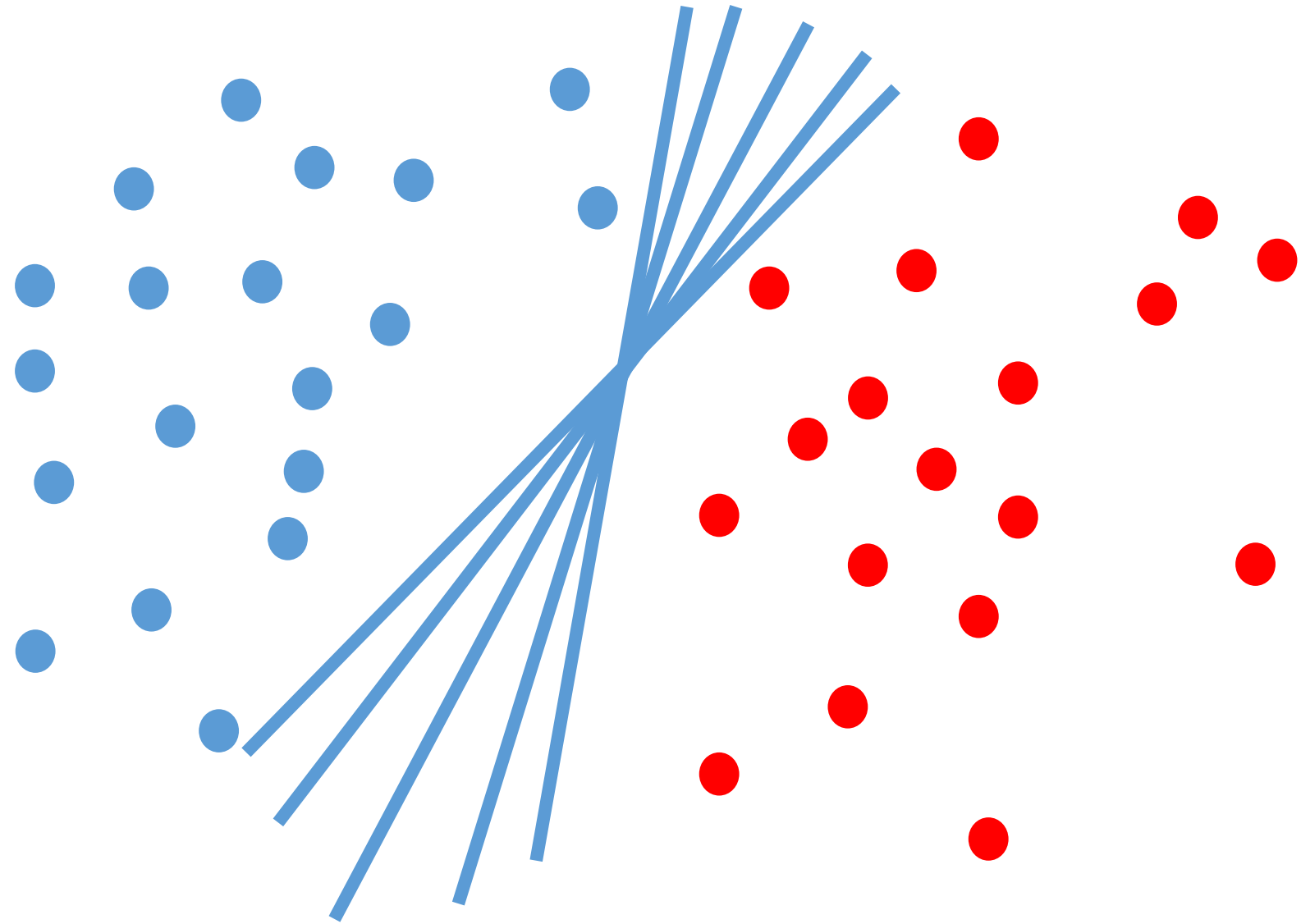
Support Vector Machine – SVM 1D



Support Vector Machine – SVM 2D

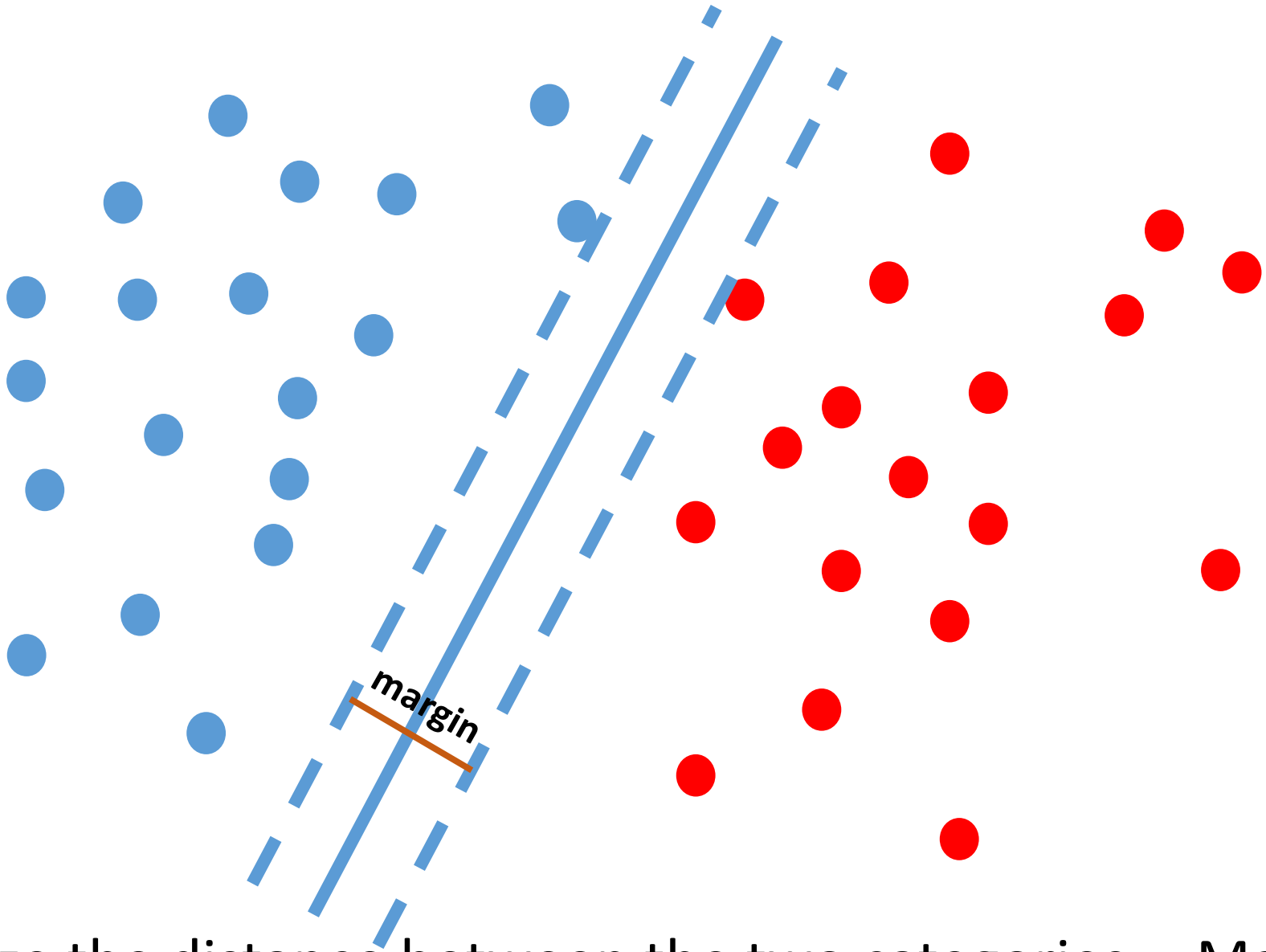


Support Vector Machine – SVM 1D



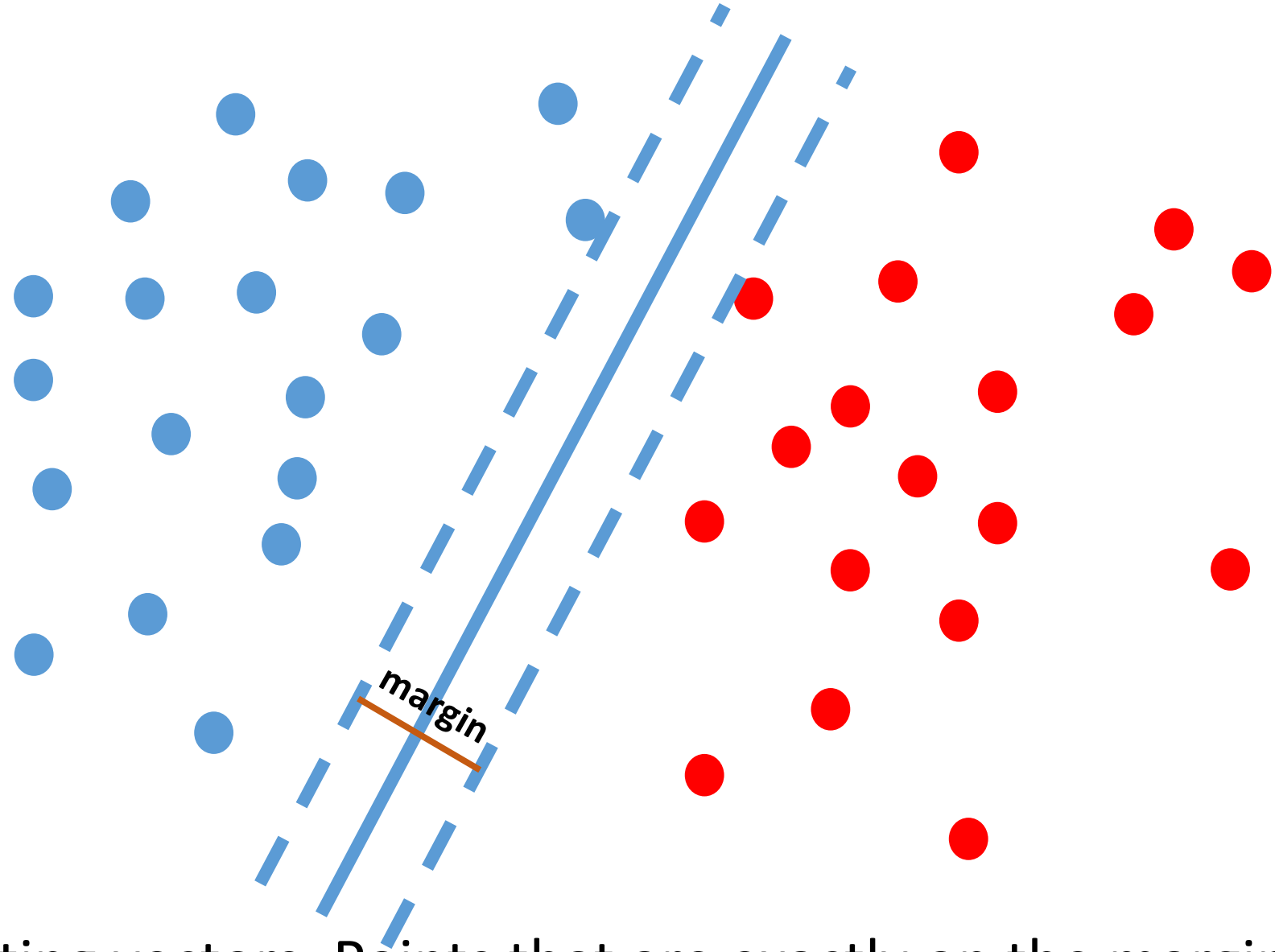
There could be multiple such hyper planes

Support Vector Machine – SVM 1D



Maximize the distance between the two categories -- Margin

Support Vector Machine – SVM 1D



Supporting vectors: Points that are exactly on the margin

Support Vector Machine – SVM

Points	Feature 1	Feature 2	…	Category
1	0.97	0.27	…	
2	0.77	0.19	…	
3	0.33	0.85	…	
4	0.41	0.93	…	
5	0.27	0.32	…	

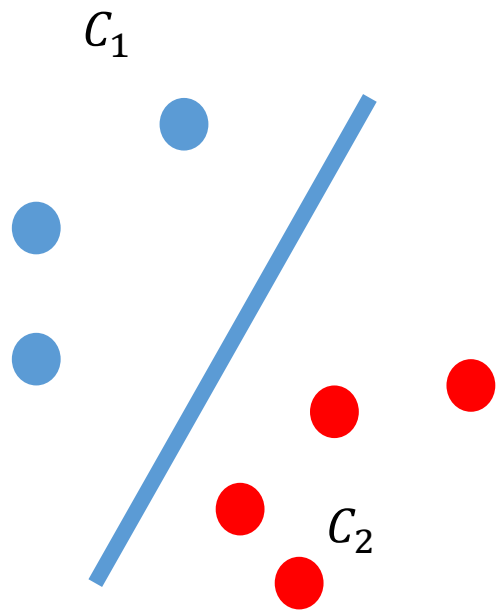
Supervised learning algorithm

Convex Optimization Problem

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

$$\text{s.t. } w^T x + b \geq 1, \forall x \in C_1$$

$$w^T x + b \leq 1, \forall x \in C_2$$

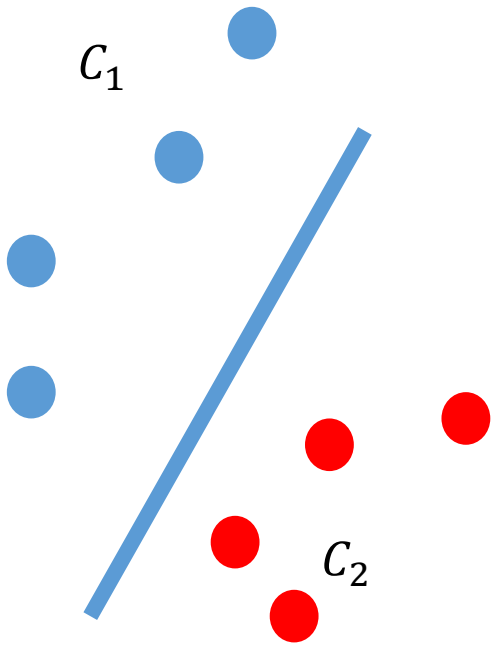


Supervised learning algorithm

from sklearn import svm

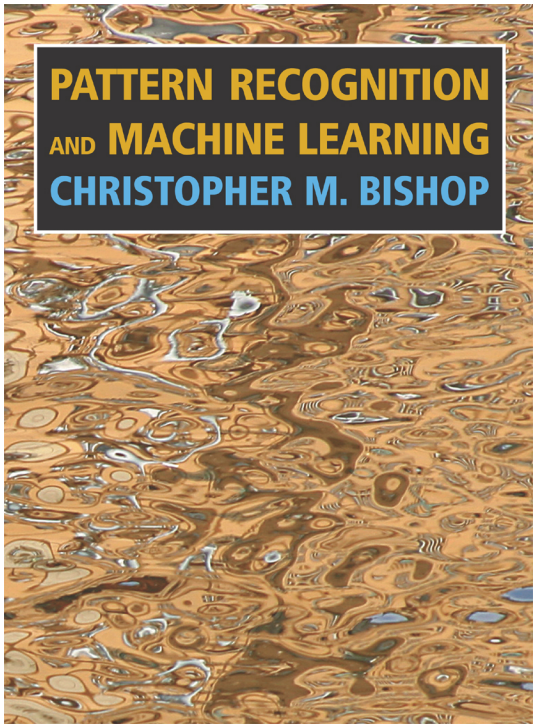
```
from sklearn import svm
# features
X = [
      [-3, -1]
      [0, -2]
      [-2.5, 2]
      [-1, -1]
      [3, .5]
      [.5, 3]
      [3, 3]
]

# labels
y = [0, 1, 0, 1, 1, 0, 1]
# fit
clf = svm.SVC(kernel='linear').fit(x,y)
# predict
clf.predict([[2, 4]])
```



References

Classification
(Logistic regression)



Bishop: Chapter 4.3

Step 1: Function Set

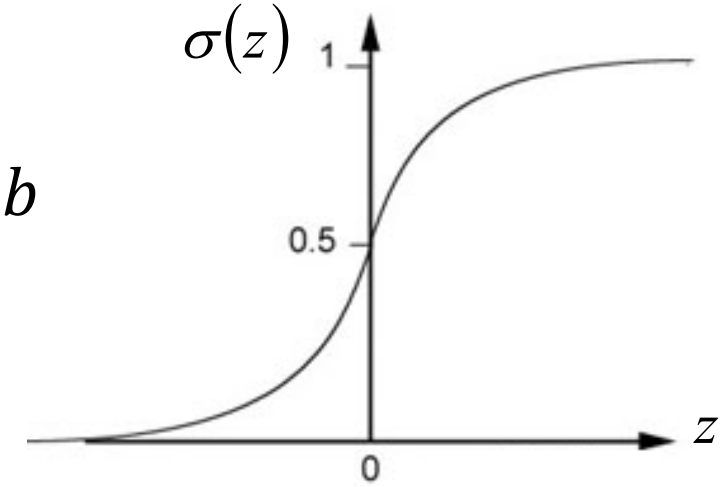
Function set: Including all different w and b

z	$\geq$	0	class 1
z	$<$	0	class 2

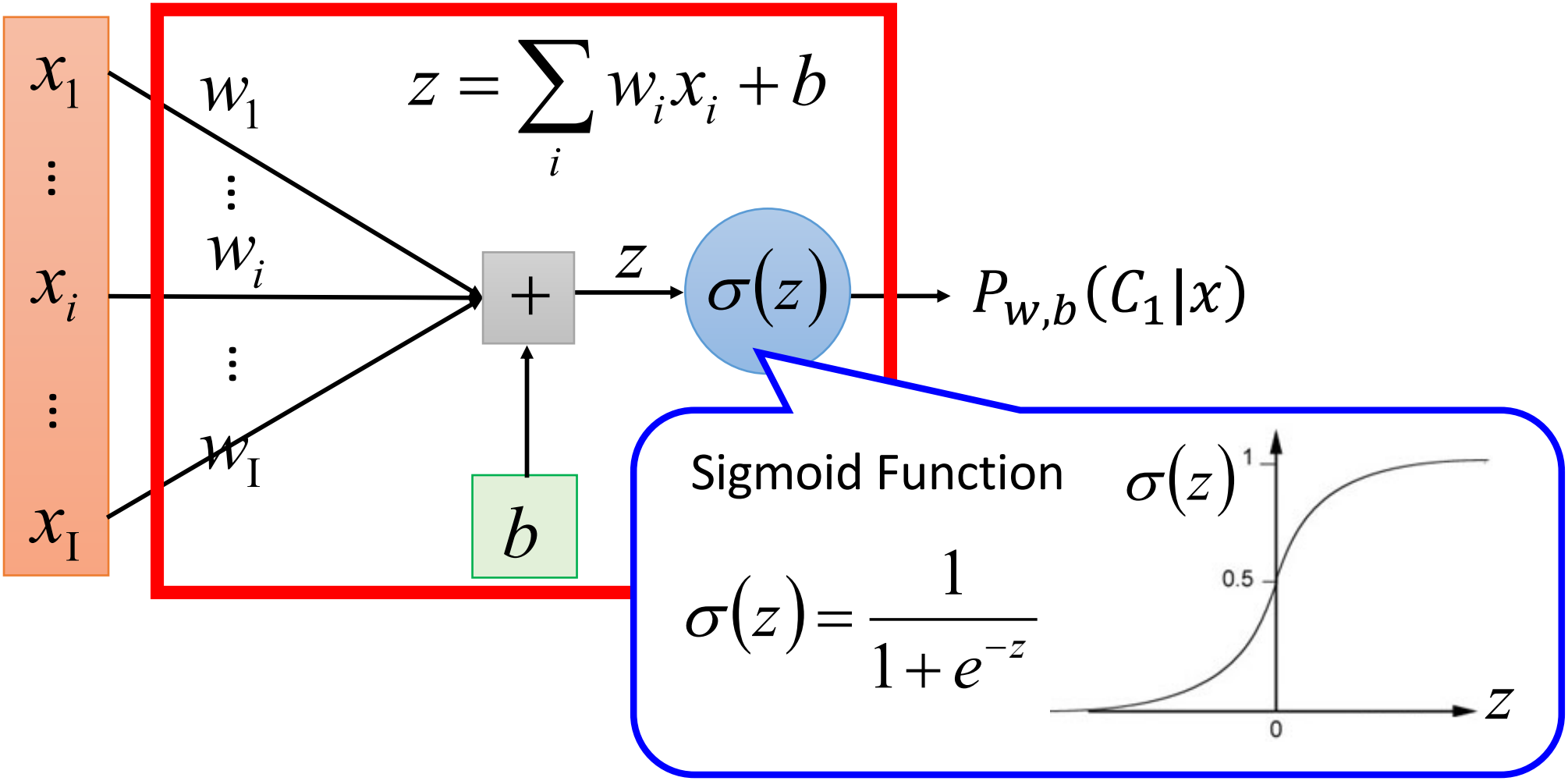
$$P_{w,b}(C_1|x) = \sigma(z)$$

$$z = w \cdot x + b = \sum_i w_i x_i + b$$

$$\sigma(z) = \frac{1}{1 + exp(-z)}$$



Step 1: Function Set



Step 2: Goodness of a Function

Training Data

x^1	x^2	x^3	……	x^N
C_1	C_1	C_2		C_1

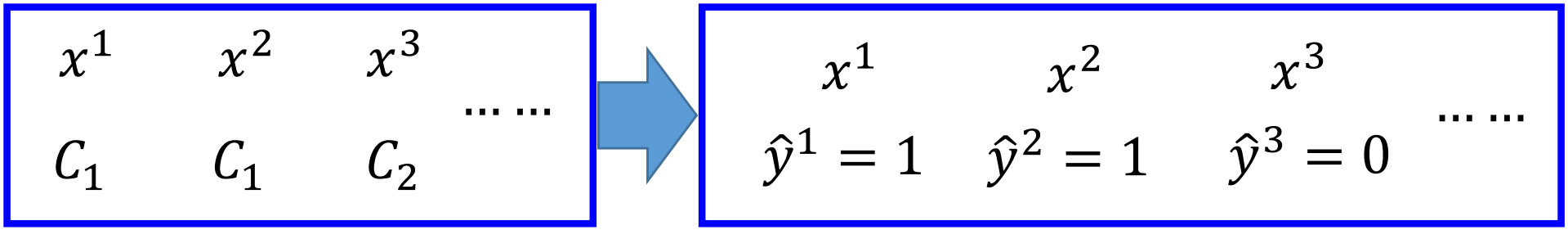
Assume the data is generated based on $f_{w,b}(x) = P_{w,b}(C_1|x)$

Given a set of w and b, what is its probability of generating the data?

$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}(x^N)$$

The most likely w^* and b^* is the one with the largest $L(w,b)$.

$$w^*, b^* = arg \max_{w,b} L(w,b)$$



$\hat{y}^n$: 1 for class 1, 0 for class 2

$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots$$

$$w^*, b^* = arg \max_{w,b} L(w,b) = w^*, b^* = arg \min_{w,b} -lnL(w,b)$$

$$-lnL(w,b)$$

$$= -lnf_{w,b}(x^1) \rightarrow -[1\ lnf(x^1) + 0\ ln(1 - f(x^1))]$$

$$-lnf_{w,b}(x^2) \rightarrow -[1\ lnf(x^2) + 0\ ln(1 - f(x^2))]$$

$$-ln\left(1 - f_{w,b}(x^3)\right) \rightarrow -[0\ lnf(x^3) + 1\ ln(1 - f(x^3))]$$

$$\vdots$$

Step 2: Goodness of a Function

$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}(x^N)$$

$$-lnL(w,b) = lnf_{w,b}(x^1) + lnf_{w,b}(x^2) + ln\left(1 - f_{w,b}(x^3)\right) \cdots$$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$= \sum_n -\left[\hat{y}^n lnf_{w,b}(x^n) + (1-\hat{y}^n) ln\left(1 - f_{w,b}(x^n)\right)\right]$$

Cross entropy between two Bernoulli distribution

Distribution p:

$\mathrm{p}(x=1) = \hat{y}^n$

$\mathrm{p}(x=0) = 1 - \hat{y}^n$

cross entropy

Distribution q:

$\mathrm{q}(x=1) = f(x^n)$

$\mathrm{q}(x=0) = 1 - f(x^n)$

$$H(p,q) = -\sum_x [p(x) ln\big(q(x)\big) + \big(1-p(x)\big) ln\big(1-q(x)\big)]$$

Step 2: Goodness of a Function

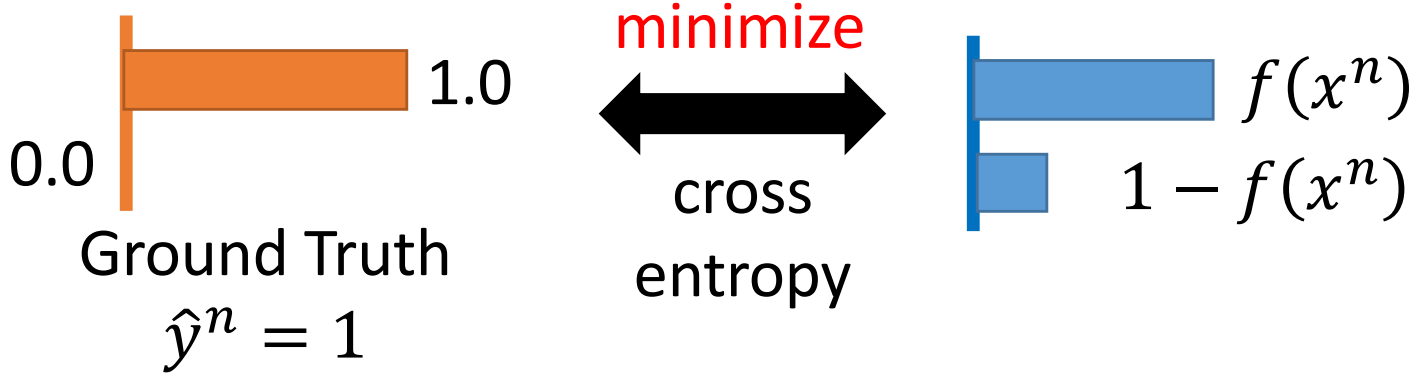
$$L(w,b) = f_{w,b}(x^1) f_{w,b}(x^2) \left(1 - f_{w,b}(x^3)\right) \cdots f_{w,b}(x^N)$$

$$-lnL(w,b) = lnf_{w,b}(x^1) + lnf_{w,b}(x^2) + ln\left(1 - f_{w,b}(x^3)\right) \cdots$$

$\hat{y}^n$: 1 for class 1, 0 for class 2

$$= \sum_n -\left[\hat{y}^n lnf_{w,b}(x^n) + (1-\hat{y}^n) ln\left(1 - f_{w,b}(x^n)\right)\right]$$

Cross entropy between two Bernoulli distribution



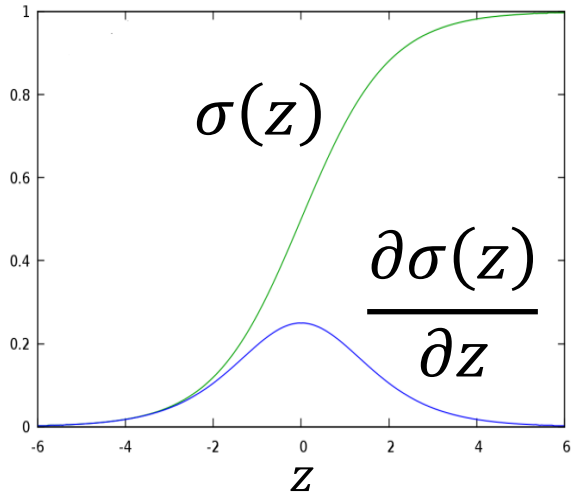
Step 3: Find the best function

$$-\frac{lnL(w,b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \frac{ln f_{w,b}(x^n)}{\partial w_i} + (1-\hat{y}^n)\frac{ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i}\right]$$

$$\left(1-f_{w,b}(x^n)\right)x_i^n$$

$$\frac{\partial ln f_{w,b}(x)}{\partial w_i} = \frac{\partial ln f_{w,b}(x)}{\partial z}\frac{\partial z}{\partial w_i} \qquad \frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial ln\sigma(z)}{\partial z} = \frac{1}{\sigma(z)}\frac{\partial \sigma(z)}{\partial z} = \frac{1}{\sigma(z)}\sigma(z)\left(1-\sigma(z)\right)$$



$$f_{w,b}(x) = \sigma(z)$$
$$= 1/(1+exp(-z))$$

$$z = w\cdot x + b = \sum_i w_i x_i + b$$

Step 3: Find the best function

$$-\frac{lnL(w,b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \frac{ln f_{w,b}(x^n)}{\partial w_i} + (1-\hat{y}^n) \frac{ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i}\right]$$

$$\left(1-f_{w,b}(x^n)\right)x_i^n \qquad -f_{w,b}(x^n)x_i^n$$

$$\frac{\partial ln\left(1-f_{w,b}(x)\right)}{\partial w_i} = \frac{\partial ln\left(1-f_{w,b}(x)\right)}{\partial z}\frac{\partial z}{\partial w_i} \qquad \frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial ln\left(1-\sigma(z)\right)}{\partial z} = -\frac{1}{1-\sigma(z)}\frac{\partial \sigma(z)}{\partial z} = -\frac{1}{1-\sigma(z)}\sigma(z)\left(1-\sigma(z)\right)$$

$$f_{w,b}(x) = \sigma(z)$$
$$= 1/1+exp(-z)$$

$$z = w \cdot x + b = \sum_i w_i x_i + b$$

Step 3: Find the best function

$$-\frac{lnL(w,b)}{\partial w_i} = \sum_n -\left[\hat{y}^n \frac{ln f_{w,b}(x^n)}{\partial w_i} + (1-\hat{y}^n) \frac{ln\left(1-f_{w,b}(x^n)\right)}{\partial w_i}\right]$$

$\left(1-f_{w,b}(x^n)\right)x_i^n$ $\quad -f_{w,b}(x^n)x_i^n$

$$= \sum_n -\left[\hat{y}^n \left(1-f_{w,b}(x^n)\right) x_i^n - (1-\hat{y}^n) f_{w,b}(x^n) x_i^n\right]$$

$$= \sum_n -\left[\hat{y}^n - \hat{y}^n f_{w,b}(x^n) - f_{w,b}(x^n) + \hat{y}^n f_{w,b}(x^n)\right] x_i^n$$

$$= \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$$

Larger difference, larger update

$$w_i \leftarrow w_i - \eta \sum_n -\left(\hat{y}^n - f_{w,b}(x^n)\right) x_i^n$$

Multi-class Classification (3 classes as example)

C_1: w^1, b_1 $\quad z_1 = w^1 \cdot x + b_1$

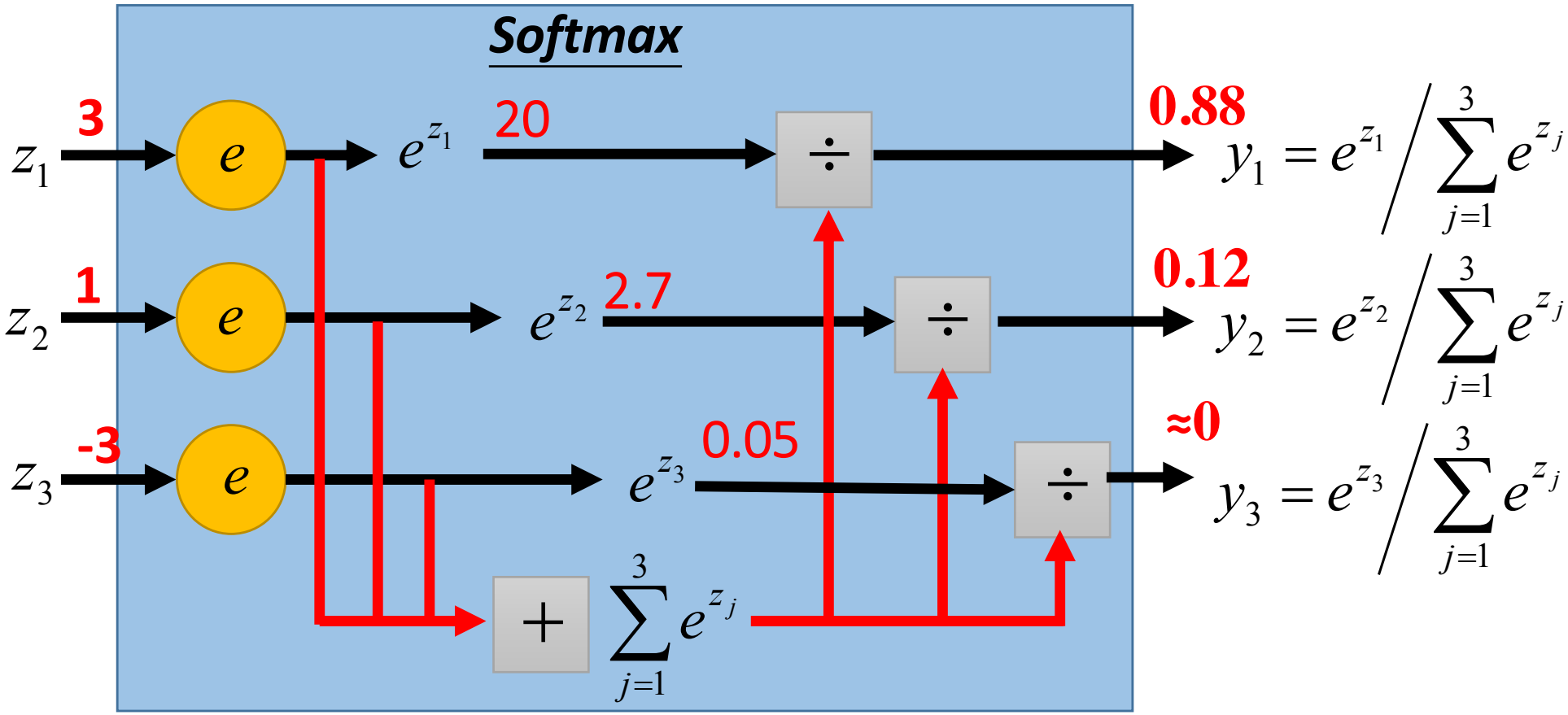
C_2: w^2, b_2 $\quad z_2 = w^2 \cdot x + b_2$

C_3: w^3, b_3 $\quad z_3 = w^3 \cdot x + b_3$

Probability:

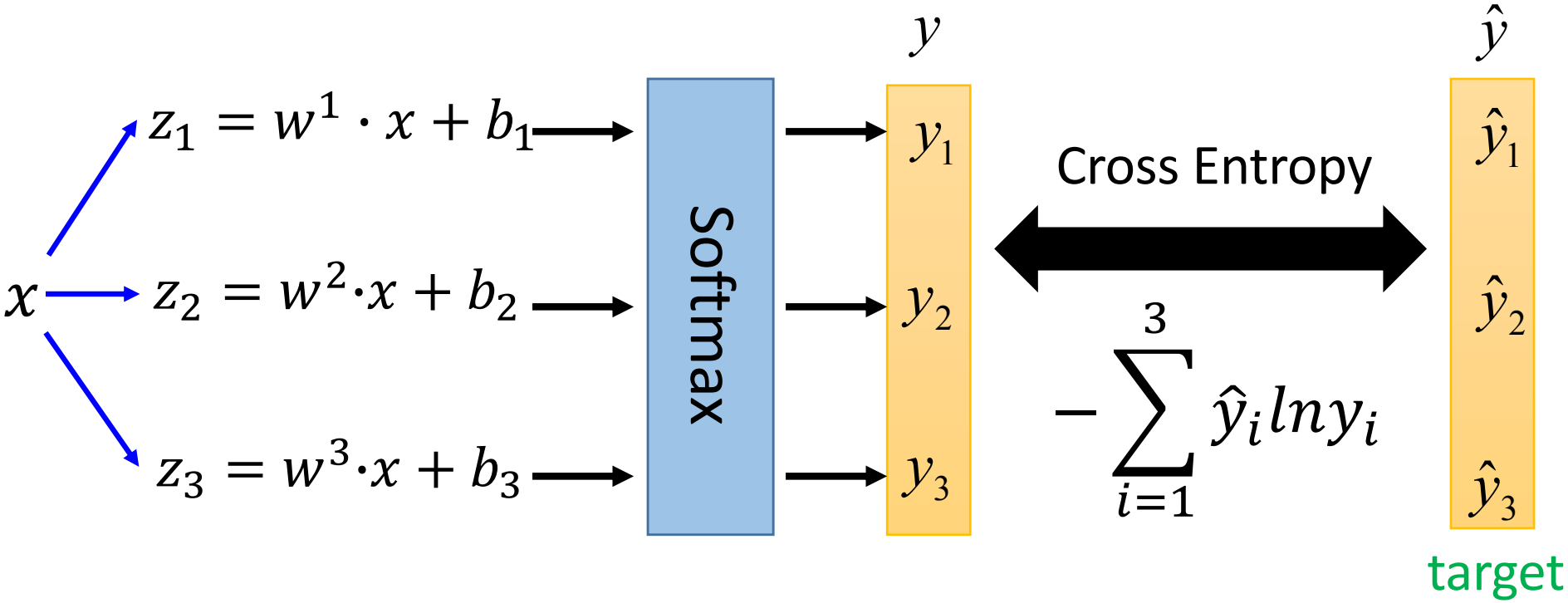
- $1 > y_i > 0$
- $\sum_i y_i = 1$

$$y_i = P(C_i \mid x)$$



Multi-class Classification (3 classes as example)

[Bishop, P209-210]



If x ∈ class 1

$$\hat{y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$-ln y_1$$

If x ∈ class 2

$$\hat{y} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

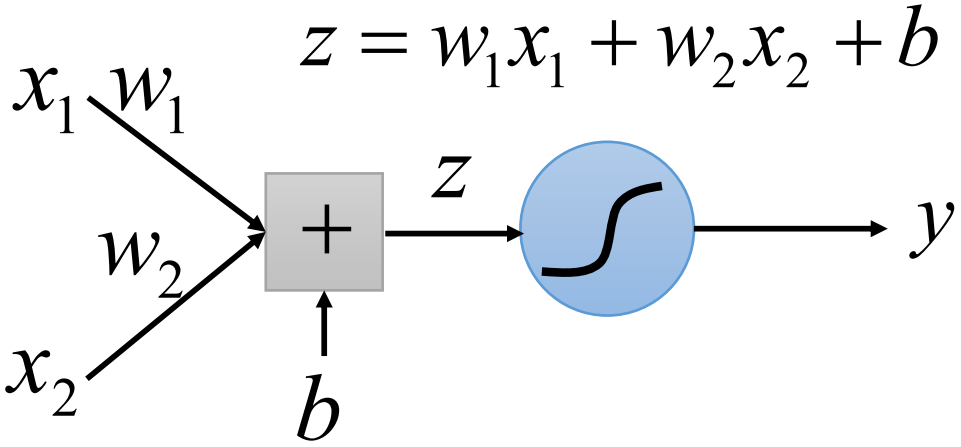
$$-ln y_2$$

If x ∈ class 3

$$\hat{y} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

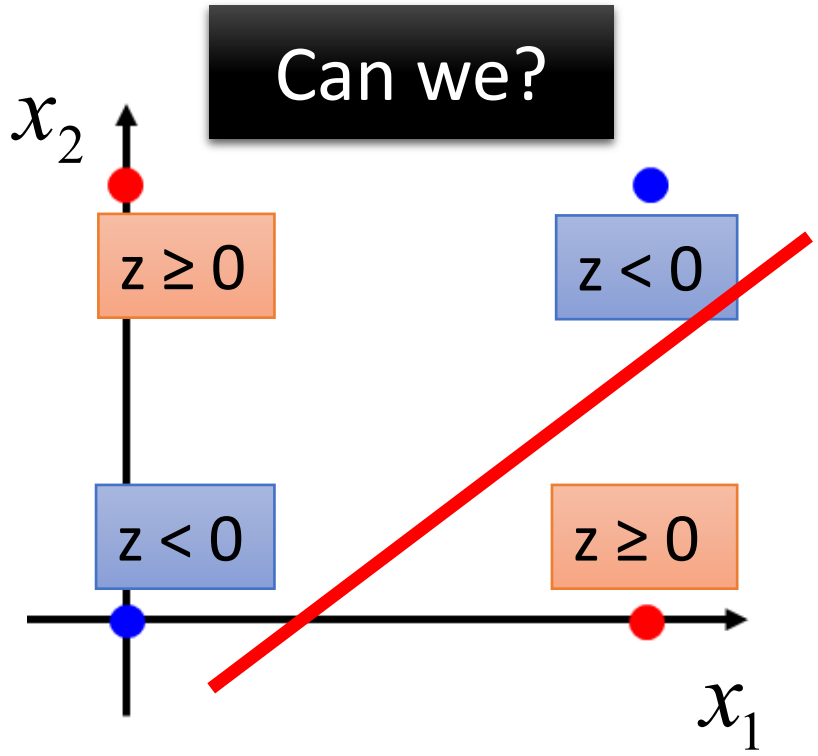
$$-ln y_3$$

Limitation of Logistic Regression



$$\begin{cases} Class1 & y \geq 0.5 \quad (z \geq 0) \\ Class2 & y < 0.5 \quad (z < 0) \end{cases}$$

Input Feature		Label
x_1	x_2	
0	0	Class 2
0	1	Class 1
1	0	Class 1
1	1	Class 2

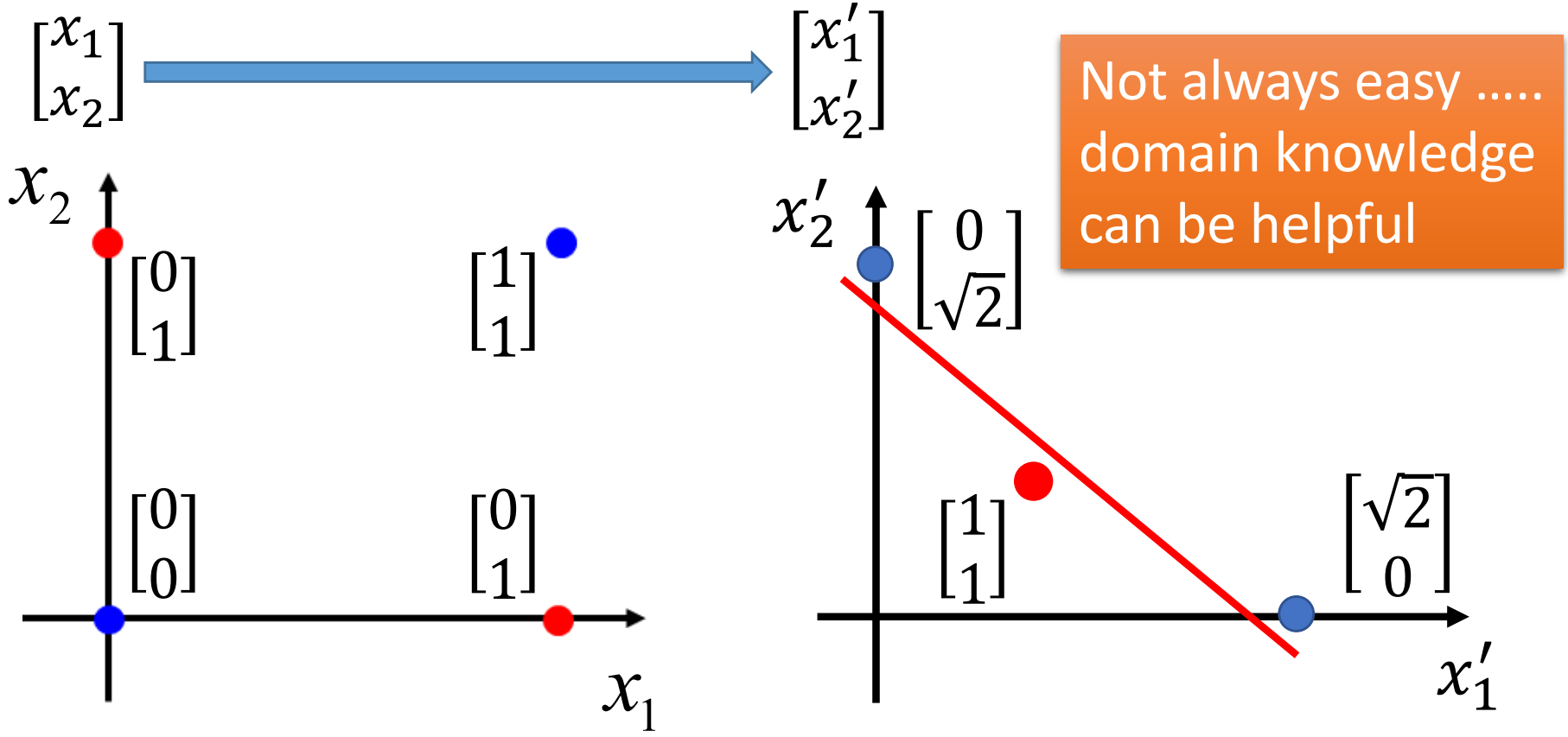


Limitation of Logistic Regression

- ***Feature transformation***

x_1': distance to $\begin{bmatrix}0\\0\end{bmatrix}$

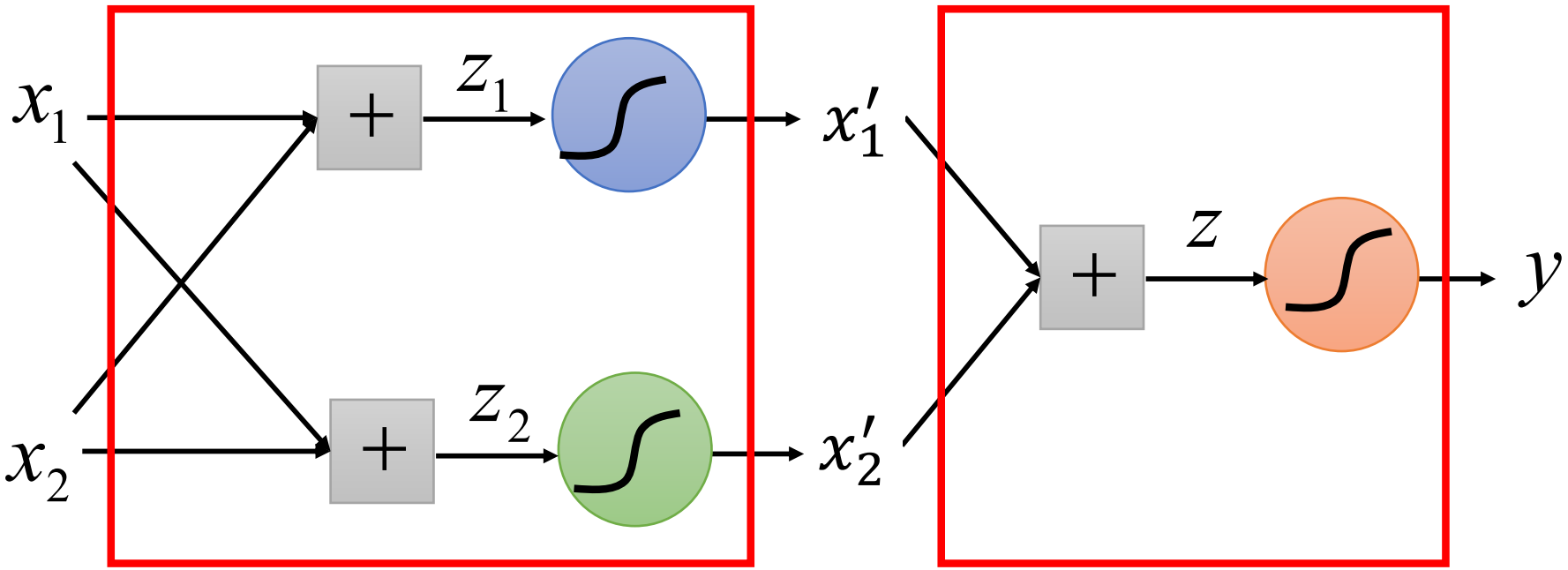
x_2': distance to $\begin{bmatrix}1\\1\end{bmatrix}$



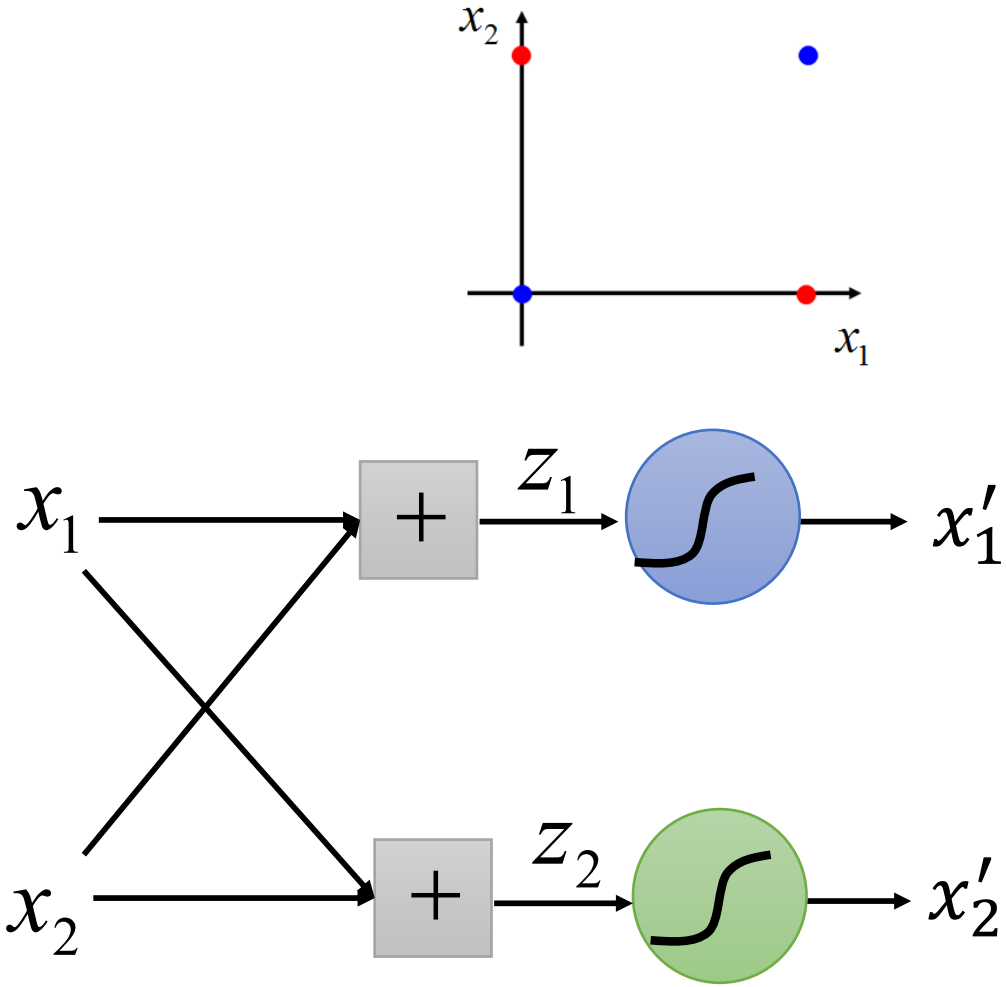
Not always easy ..... domain knowledge can be helpful

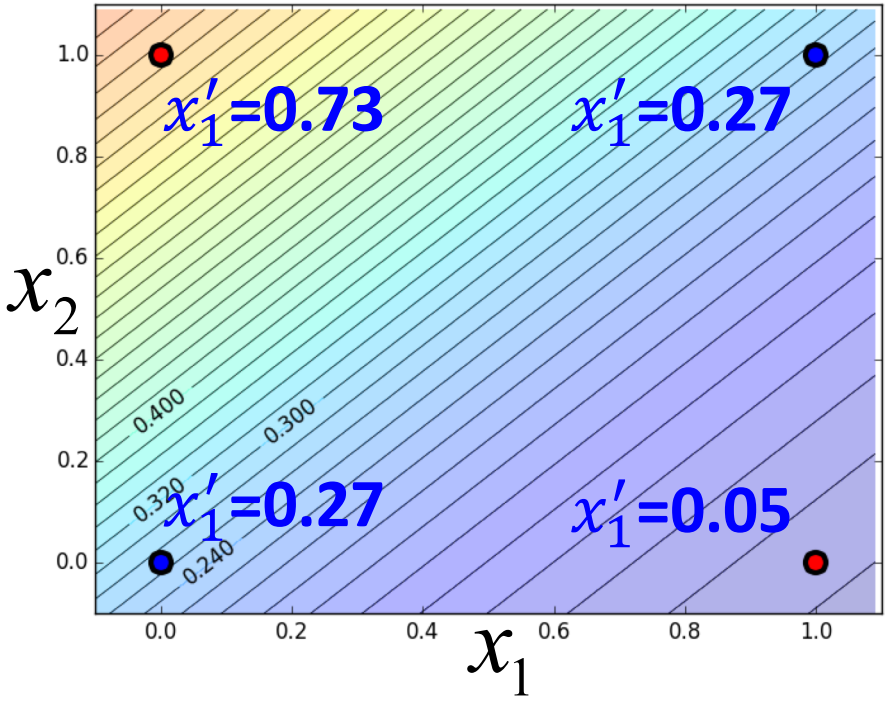
Limitation of Logistic Regression

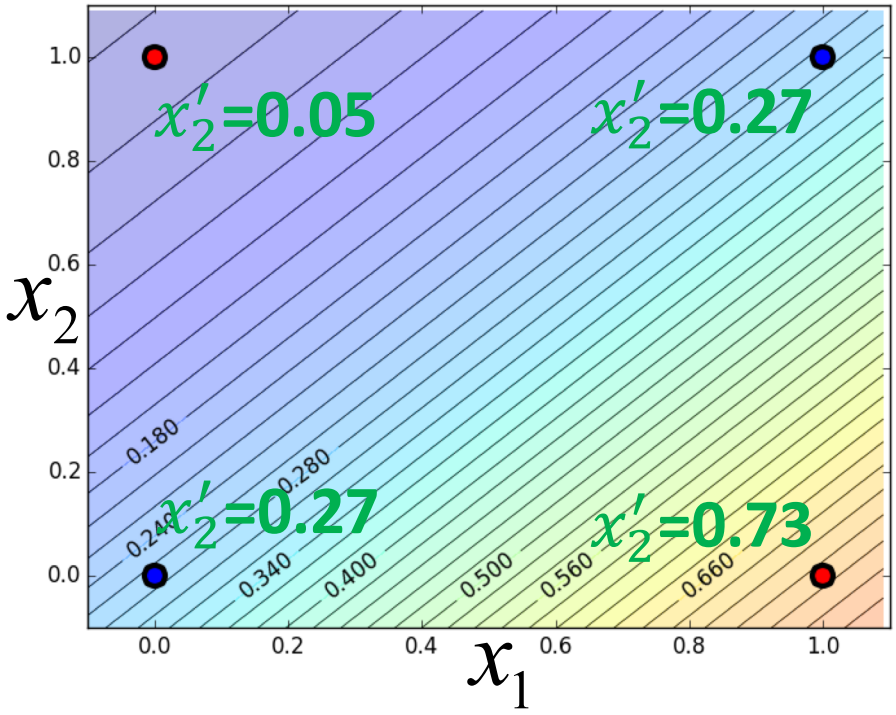
- Cascading logistic regression models

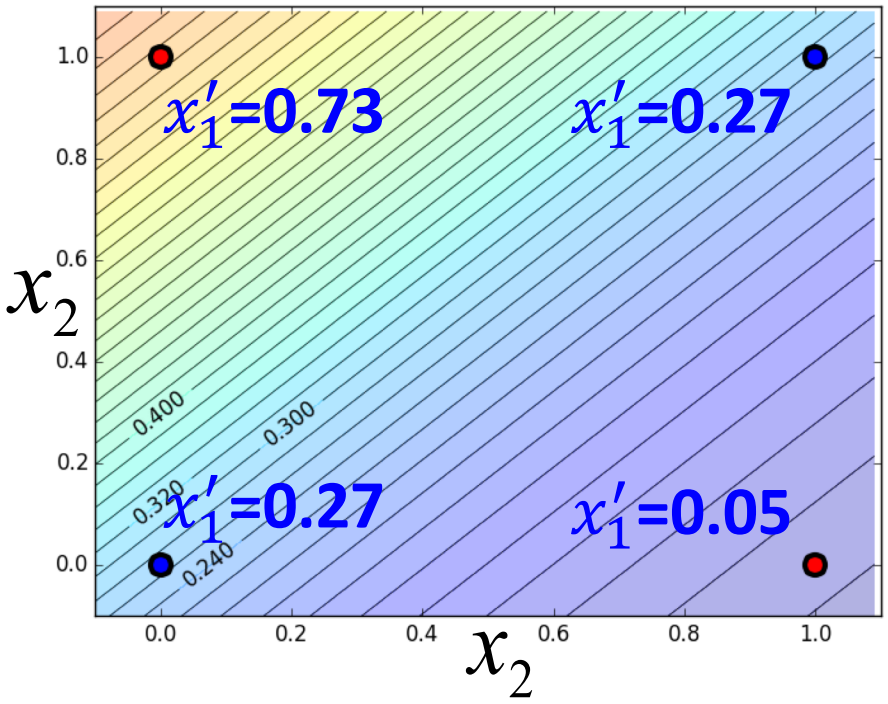


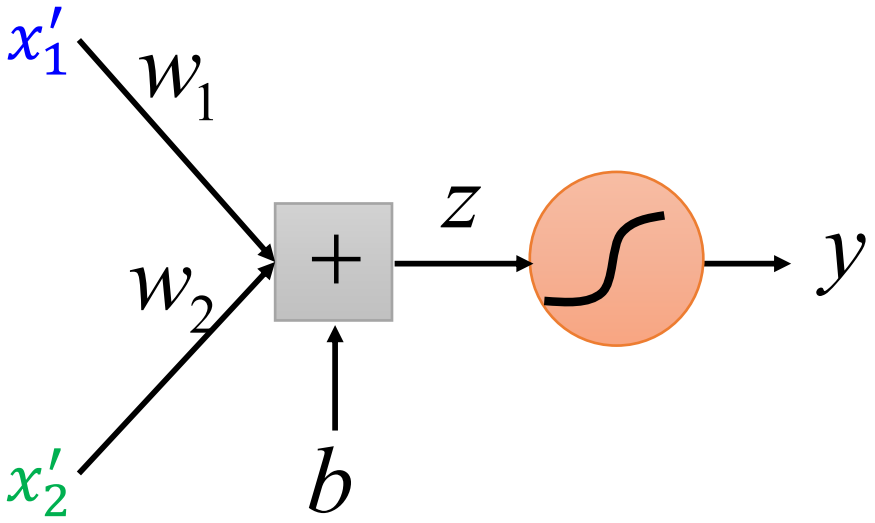
(ignore bias in this figure)

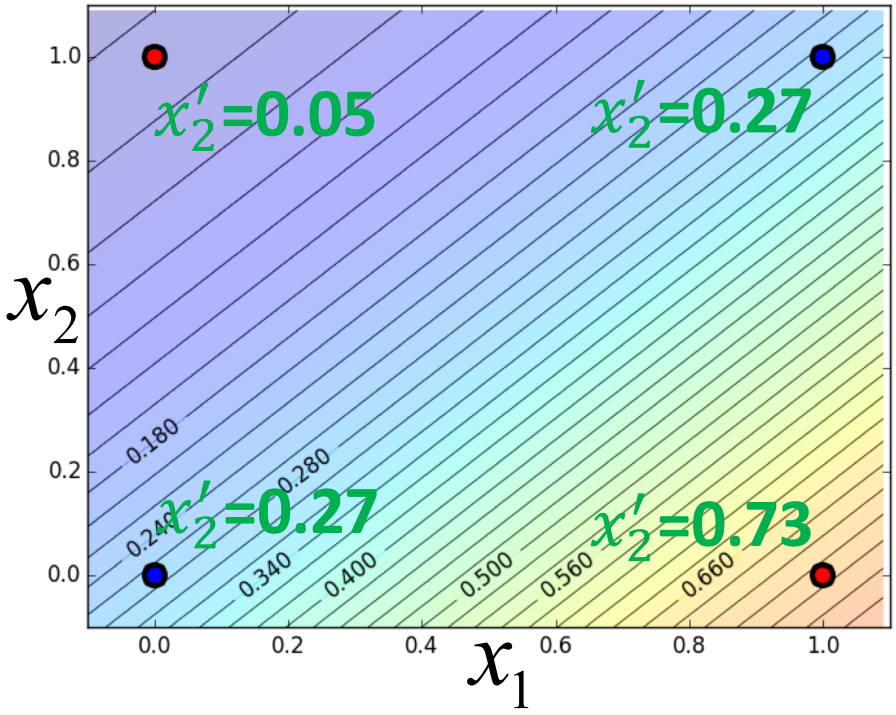

x_2
x_1
x_1
x_2
+
z_1
x_1'
+
z_2
x_2'

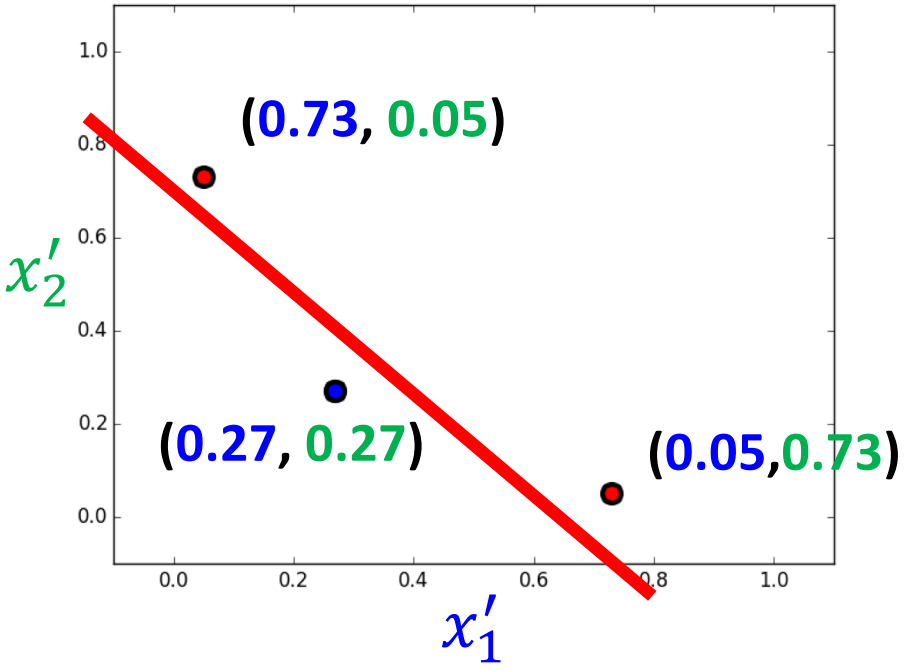

x_1'=0.73
x_1'=0.27
x_1'=0.27
x_1'=0.05
0.400
0.300
0.320
0.240
x_2
x_1
0.0
0.2
0.4
0.6
0.8
1.0


x_2'=0.05
x_2'=0.27
x_2'=0.27
x_2'=0.73
0.180
0.280
0.240
0.340
0.400
0.500
0.560
0.660
x_2
x_1
0.0
0.2
0.4
0.6
0.8
1.0


x'_1=0.73
x'_1=0.27
x'_1=0.27
x'_1=0.05
0.400
0.300
0.320
0.240
x_2
x_1

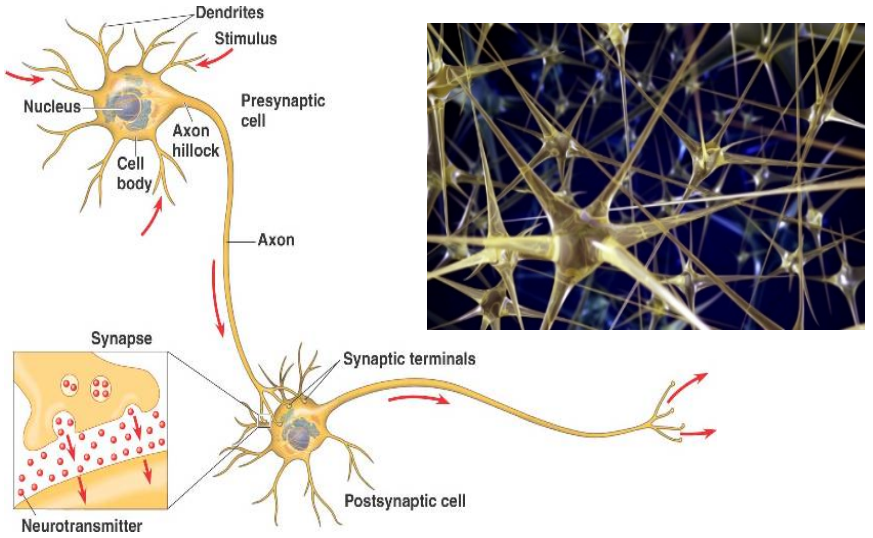

x'_1
w_1
w_2
x'_2
b
z
y

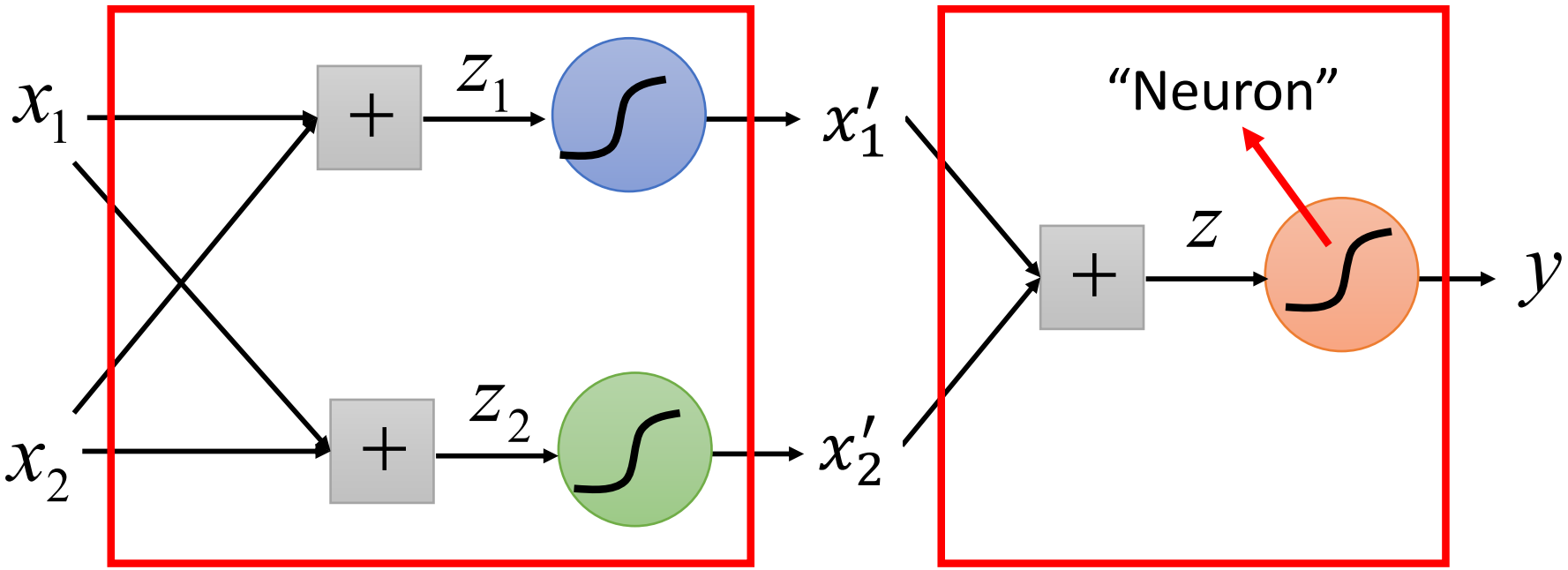

x'_2=0.05
x'_2=0.27
x'_2=0.27
x'_2=0.73
0.180
0.280
0.240
0.340
0.400
0.500
0.560
0.660
x_2
x_1


(0.73, 0.05)
(0.27, 0.27)
(0.05,0.73)
x'_2
x'_1

Deep Learning!

All the parameters of the logistic regressions are jointly learned.





Neural Network

Questions