FOUNDATIONS OF NATURAL LANGUAGE UNDERSTANDING
自然語言理解的基礎

張家瑋 博士
國立臺中科技大學資訊工程系專案助理教授
國立成功大學工程科學系兼任助理教授

FINTECH PROJECT
CREDIT RISK

徵審系統資料來源

客戶提供之資料
(財報、基本資料)

徵審
系統

經濟部
商業司網站

經濟部商業司為中華民國經濟部的業務單位之一，主管商業事務及公司登記等

財團法人
聯合徵信中心

成立於1975年，是國內唯一的跨金融機構間信用報告機構，同時蒐集個人與企業信用報告，建置全國信用資料庫，以提供信用紀錄及營運財務資訊予

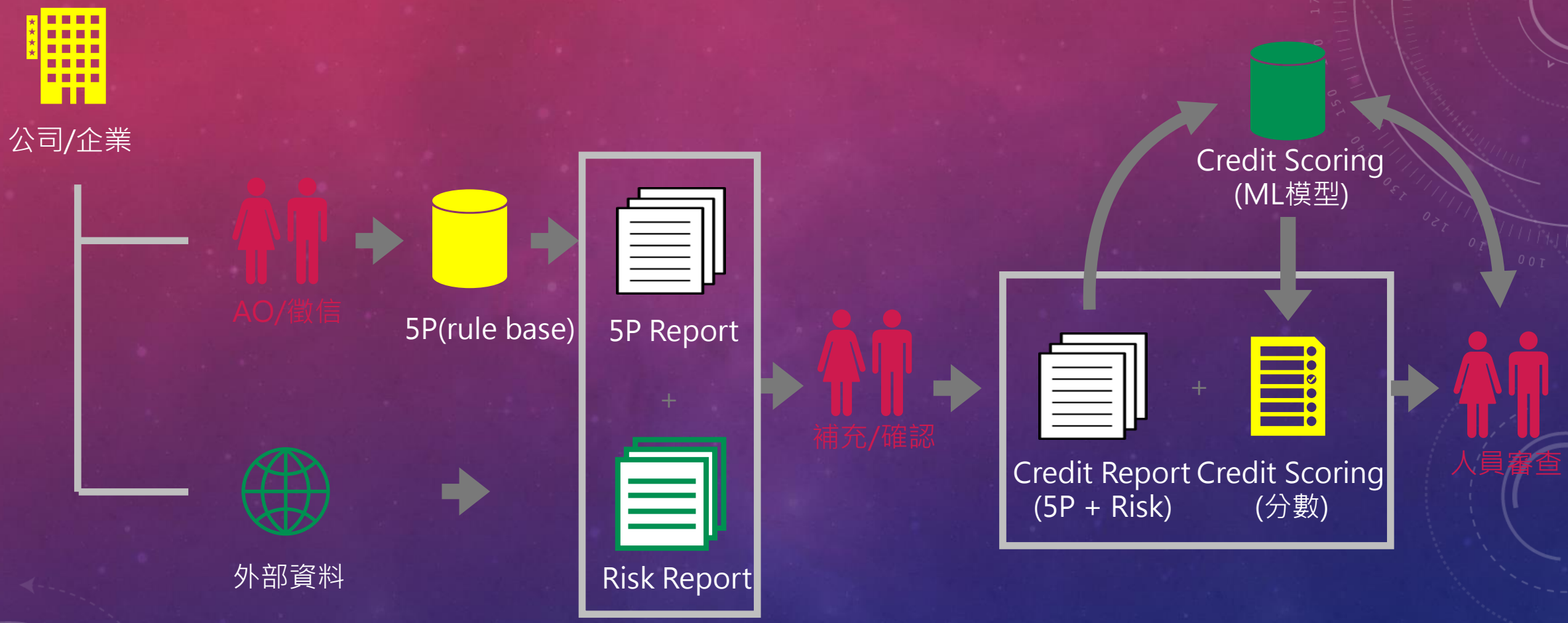
台灣經濟新報
資料庫(TEJ)

台灣經濟新報（TEJ）成立於1990年 4月，專門提供金融市場基本分析所需的資訊

台灣經濟研究院
資料庫

1976年設立，為台灣最早由民間設立之獨立學術研究機構。成立之宗旨在積極從事國內、外經濟及產業經濟之研究，並將研究成果提供政府、企業及學術界參考，以促進我國經濟發展

Credit Risk Framework



銀行人員作業　成大教授團隊　銀行自動化作業

永豐金用AI分析授信的風險管理 準確度已達9成

分享 留言 列印 存新聞 A- A+

2017-11-16 13:01 聯合報 記者林良齊／即時報導 讚 30 分享

成功大學台北辦公室過去成立30年、年久失修，今年5月與永豐金控簽定合作意象書後，永豐金捐贈經費協助修繕辦公室，今天舉辦啟用典禮，成大研發長謝孫源說，過去的產學、研發重鎮都在台南，希望透過台北辦公室啟用把觸角擴展至台北，深化產學合作。

成功大學校長蘇慧貞表示，成大陸續啟用包括在馬來西亞、越南、印尼等地中心，甚至在非洲馬達加斯加也有，預計明年也啟動歐洲的中心，但「台北反而是距離比較遠」，基地的開發宣誓要盡的社會責任不同。

蘇慧貞也說，成大將會秉持著為未來勾勒、讓民眾享受大數據年代的幸福，未來每個人到成大的每一個空間，都會有新數據的蒐集。

謝孫源指出，業界十分缺乏AI人才，因此成大也與永豐金合作，讓老師為他們上課，也期待未來在金融科技、區塊鏈技術後能夠與他們持續合作，包括客戶端的使用或金融端的風險管理等都是未來的使用範疇。

永豐金控營運長江威娜指出，目前透過初步的人工智慧分析授信的風險管理準確度已達9成，比起過去單純用統計模型的7成高了不少，未來能夠運用在中小企業的授信上，除了風險管理外，未來也會針對資產配置、壞帳警示等研究，希望能夠透過與成大合作加速上路。

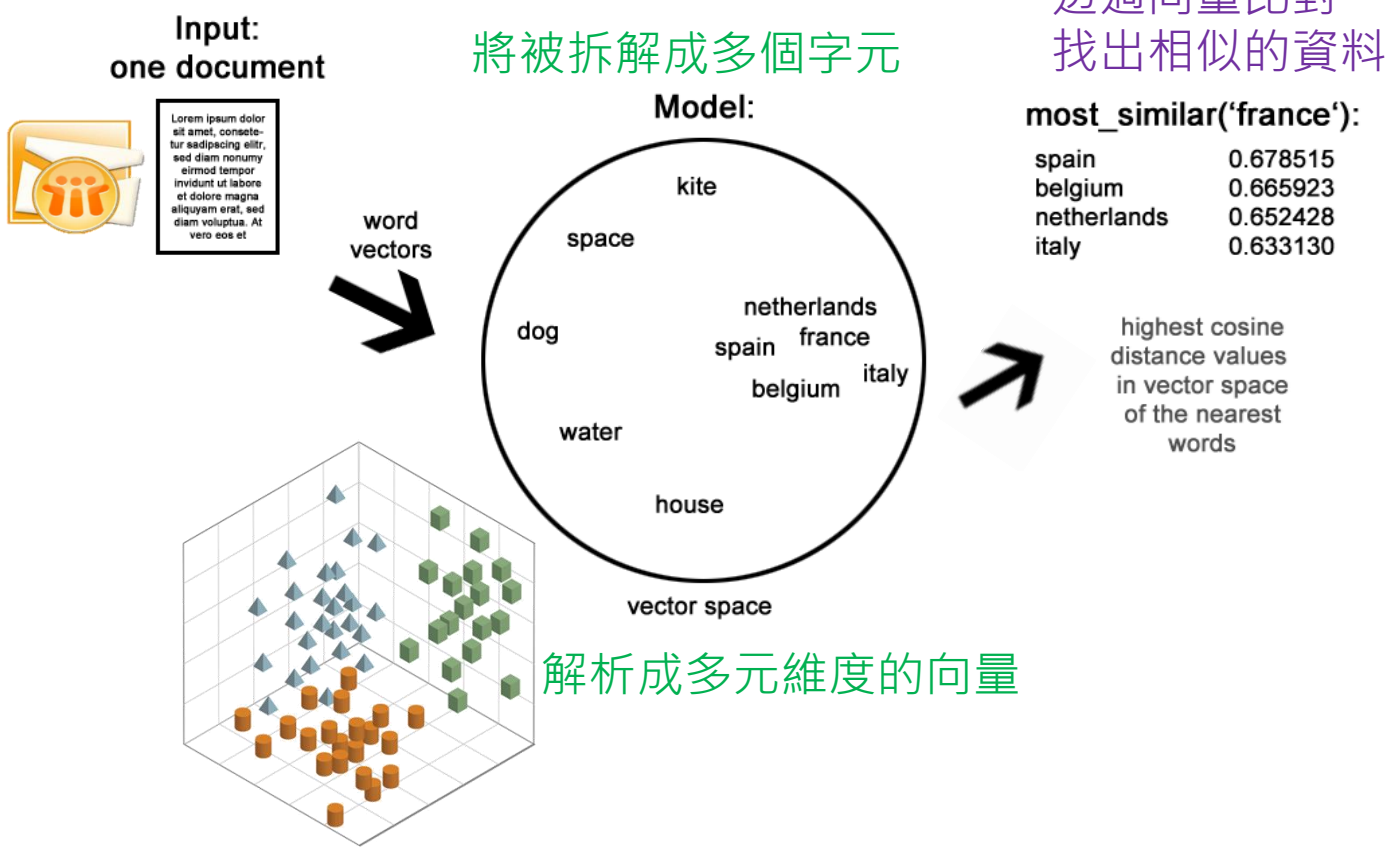
https://udn.com/news/story/7239/2822145

SEMANTIC SIMILARITY MEASURES

文字檔案

word2vec

透過向量比對
找出相似的資料

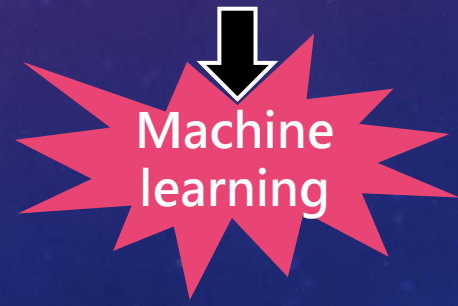


將被拆解成多個字元

解析成多元維度的向量

VECTOR REPRESENTATION

	w_1	w_2	w_3	..	..	..	w_{n-1}	w_n	label
D_1	0.11	0.23	0	..	..	..	0.57	0	0
D_2	0	0	0	..	..	..	0.29	0.7	1
D_3	0	0.81	0.44	..	..	..	0	0	0
D_4	0	0.37	0	..	..	..	0	0.16	1
..	..	..	..	..	..	..	..	..	..
D_k	..	..	..	..	..	..	..	..	1



TF-IDF

TF-IDF

- TF: term frequency: $\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
- IDF: inverse document frequency: $\mathrm{idf}_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$

where:

- $|D|$: total number of documents in the corpus
- $|\{j : t_i \in d_j\}|$: number of documents where term t_i appears

Then:

- $\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \mathrm{idf}_i$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

- The calculation of tf–idf for the term "this" is performed as follows:
  - $$\mathrm{tf}(\text{"this"}, d_1) = \frac{1}{5} = 0.2$$
    $$\mathrm{tf}(\text{"this"}, d_2) = \frac{1}{7} \approx 0.14$$

$$\mathrm{idf}(\text{"this"}, D) = \log\left(\frac{2}{2}\right) = 0$$

- So tf–idf is zero for the word "this", which implies that the word is not very informative as it appears in all documents.

$$\mathrm{tfidf}(\text{"this"}, d_1) = 0.2 \times 0 = 0$$
$$\mathrm{tfidf}(\text{"this"}, d_2) = 0.14 \times 0 = 0$$

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

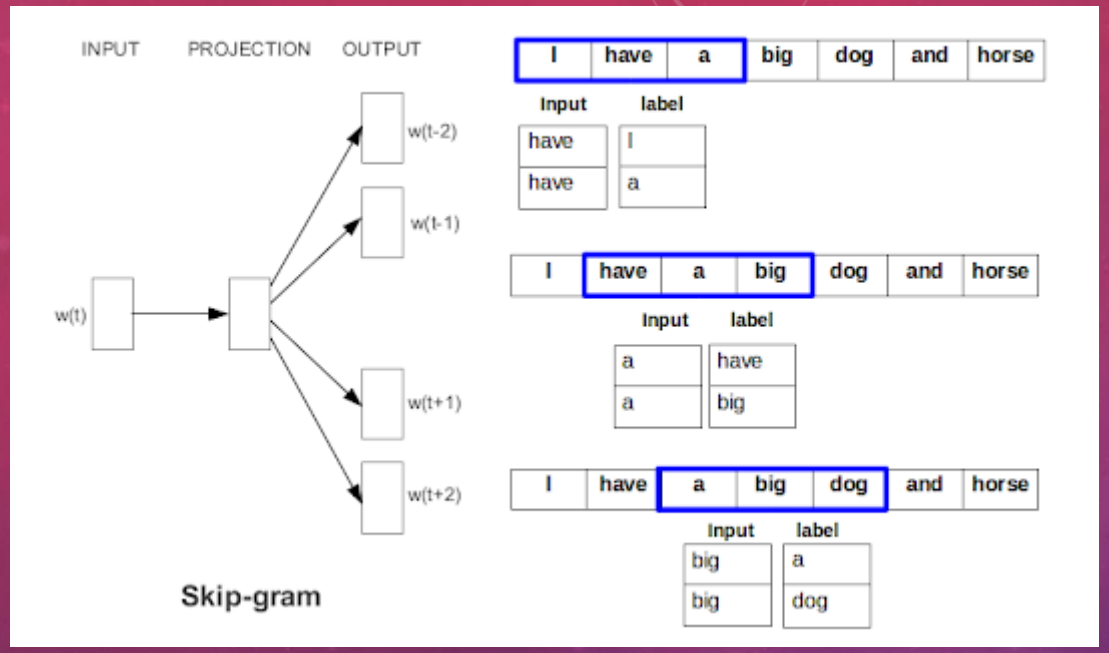
Term	Term Count
this	1
is	1
another	2
example	3

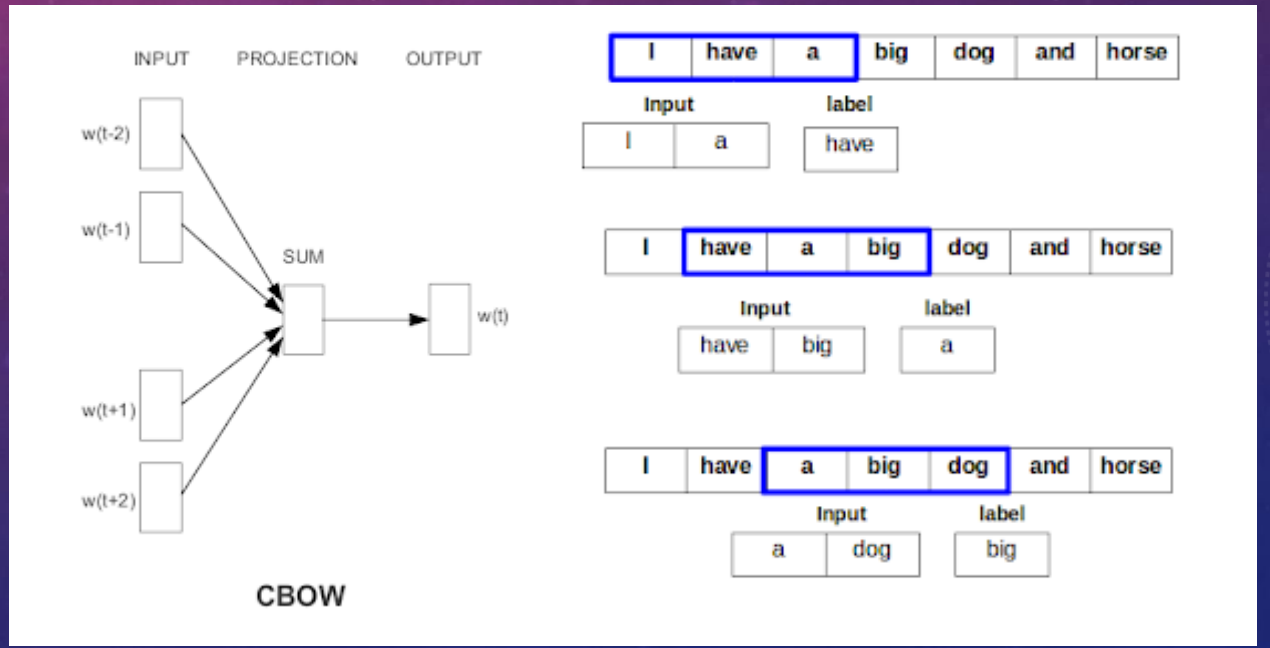
- A slightly more interesting example arises from the word "example", which occurs three times only in the second document:
  - $\mathrm{tf}(\text{"example"}, d_1) = \frac{0}{5} = 0$
  - $\mathrm{tf}(\text{"example"}, d_2) = \frac{3}{7} \approx 0.429$

$$\mathrm{idf}(\text{"example"}, D) = \log\left(\frac{2}{1}\right) = 0.301$$

$$\mathrm{tfidf}(\text{"example"}, d_1) = \mathrm{tf}(\text{"example"}, d_1) \times \mathrm{idf}(\text{"example"}, D) = 0 \times 0.301 = 0$$
$$\mathrm{tfidf}(\text{"example"}, d_2) = \mathrm{tf}(\text{"example"}, d_2) \times \mathrm{idf}(\text{"example"}, D) = 0.429 \times 0.301 \approx 0.13$$

WORD2VEC

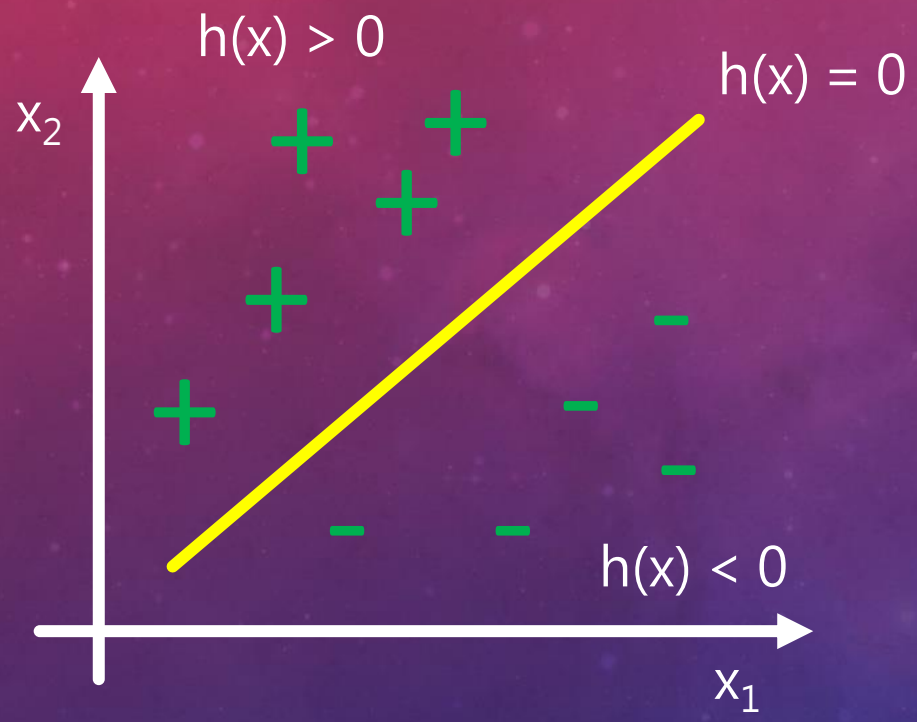

INPUT
PROJECTION
OUTPUT
w(t)
w(t-2)
w(t-1)
w(t+1)
w(t+2)
I have a big dog and horse
Input label
have I
have a
I have a big dog and horse
Input label
a have
a big
I have a big dog and horse
Input label
big a
big dog
Skip-gram


INPUT
PROJECTION
OUTPUT
w(t-2)
w(t-1)
w(t+1)
w(t+2)
SUM
w(t)
I have a big dog and horse
Input label
I a
have
I have a big dog and horse
Input label
have big
a
I have a big dog and horse
Input label
a dog
big
CBOW

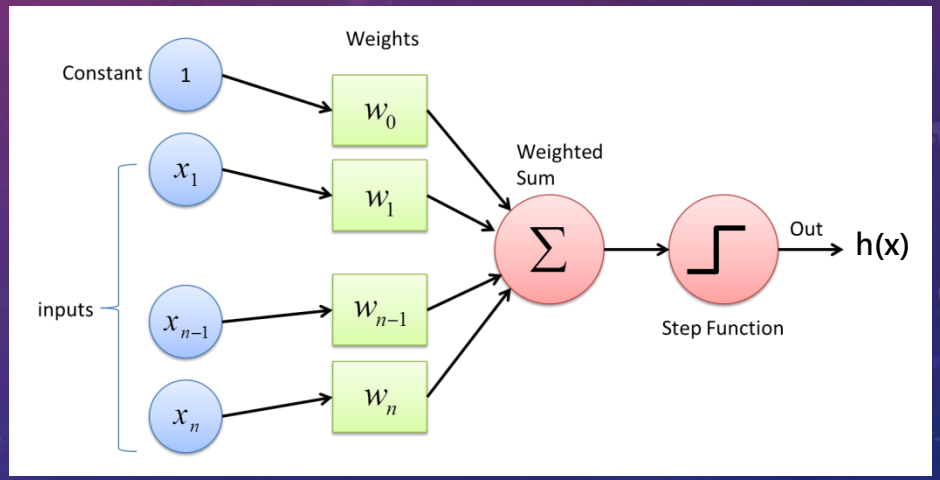
One Hot Encoding

```
The -> [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
cat -> [0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
jump -> [0. 0. 0. 1. 0. 0. 0. 0. 0. 0.]
over -> [0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
the -> [0. 0. 0. 0. 0. 1. 0. 0. 0. 0.]
dog -> [0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
The -> [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
dog -> [0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
ate -> [0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
my -> [0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
homework -> [0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
```

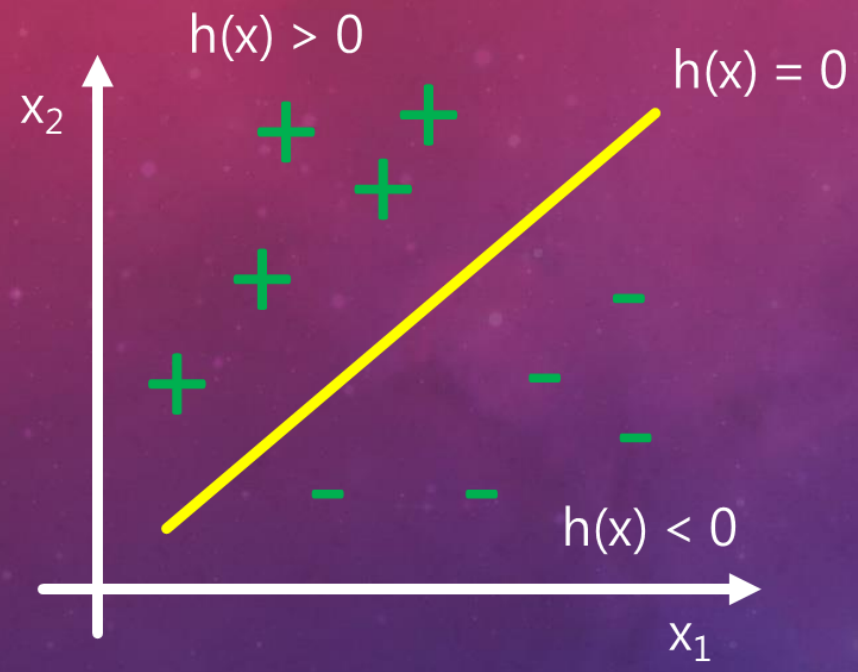
Perceptron Linear Algorithm



- Features: $x = (x_1, x_2)$
- Target: $y = +1$ or -1
- $h(x) = w_0 + w_1x_1 + w_2x_2$



Perceptron Linear Algorithm



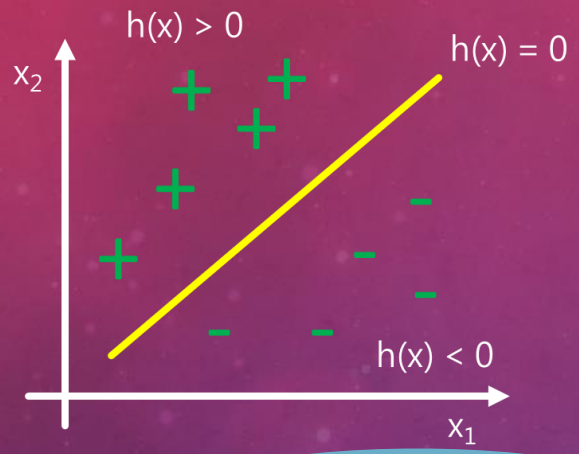
$h(x) = w_0 + w_1x_1 + w_2x_2$

$$scores = \sum_{i}^{N} w_i x_i + b$$

$$scores = \sum_{i}^{N+1} w_i x_i$$

- 若 $scores \geq 0$ ，则 $\hat{y} = 1$
- 若 $scores < 0$ ，则 $\hat{y} = -1$

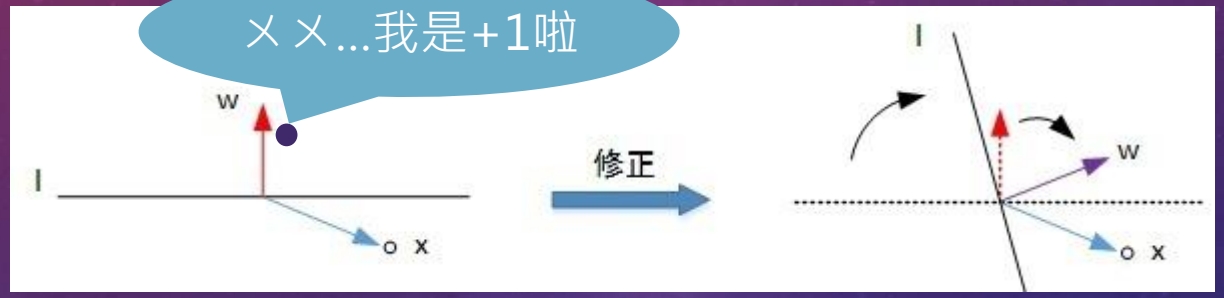
Perceptron Linear Algorithm



- 若 $scores \geq 0$ ，则 $\hat{y} = 1$
- 若 $scores < 0$ ，则 $\hat{y} = -1$

+ -

$w_{t+1} = w_t + y_t x_t$



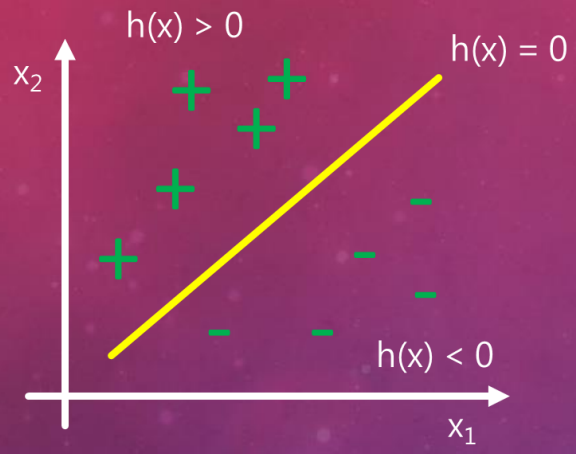
[Case 1]
y = 1 錯分成 y = -1

- +

$w_{t+1} = w_t + y_t x_t$

[Case 2]
y = -1 錯分成 y = 1

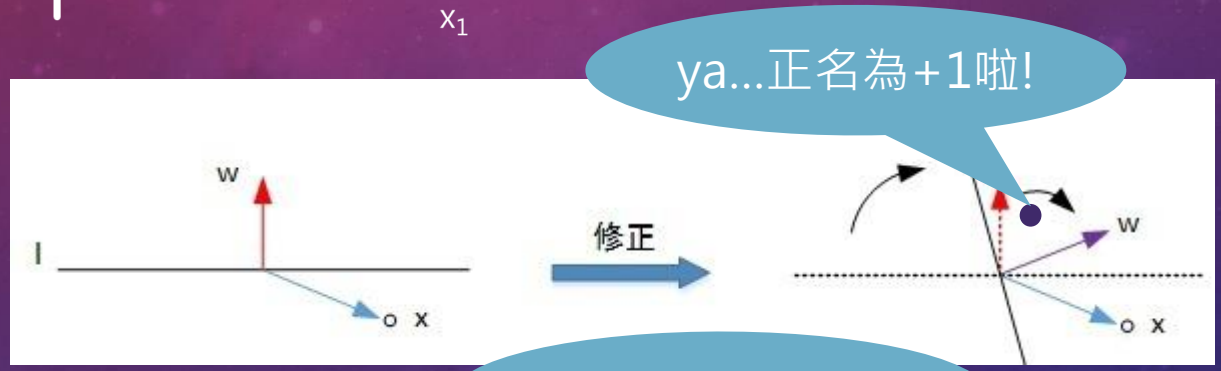
Perceptron Linear Algorithm



- 若 $scores \geq 0$ ，则 $\hat{y} = 1$
- 若 $scores < 0$ ，则 $\hat{y} = -1$

\+ - +

$w_{t+1} = w_t + y_t x_t$



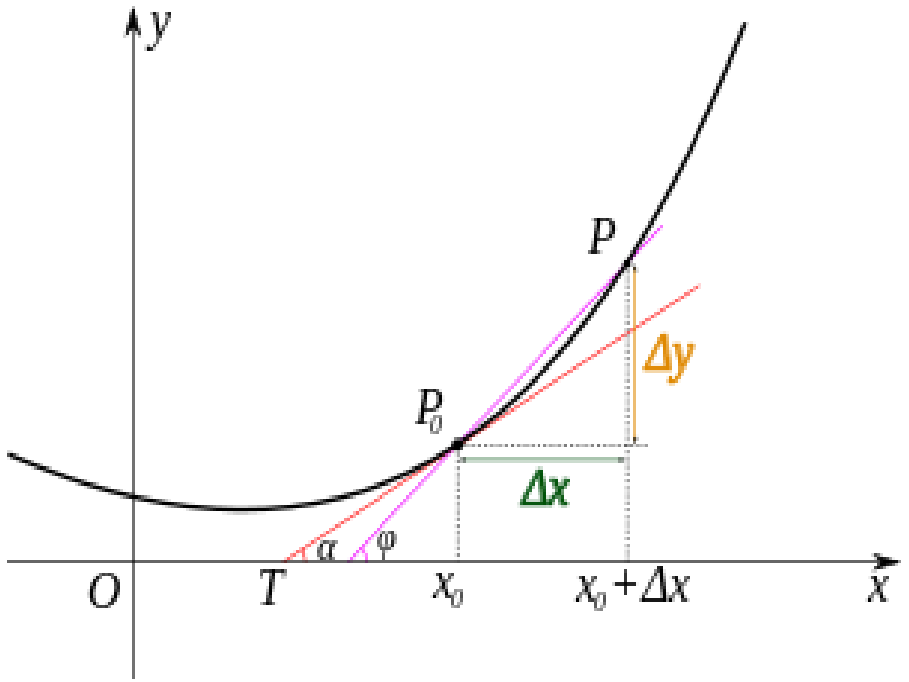
[Case 1]
y = 1 錯分成 y = -1

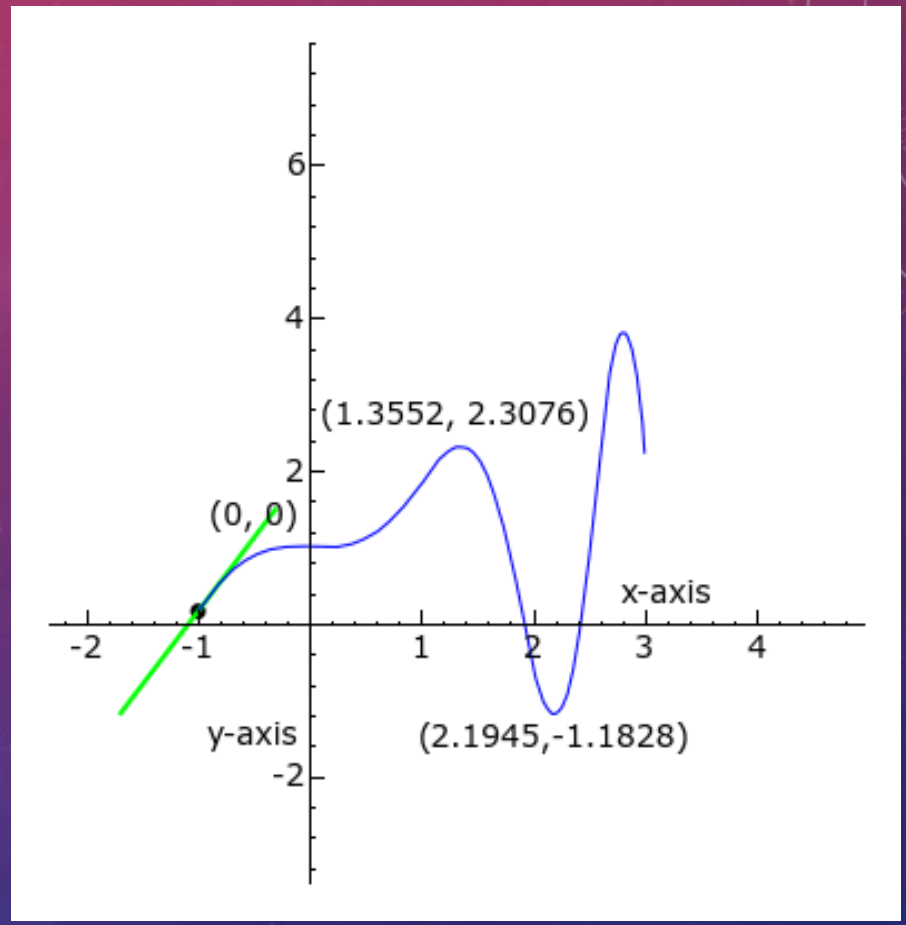
\- + -

$w_{t+1} = w_t + y_t x_t$

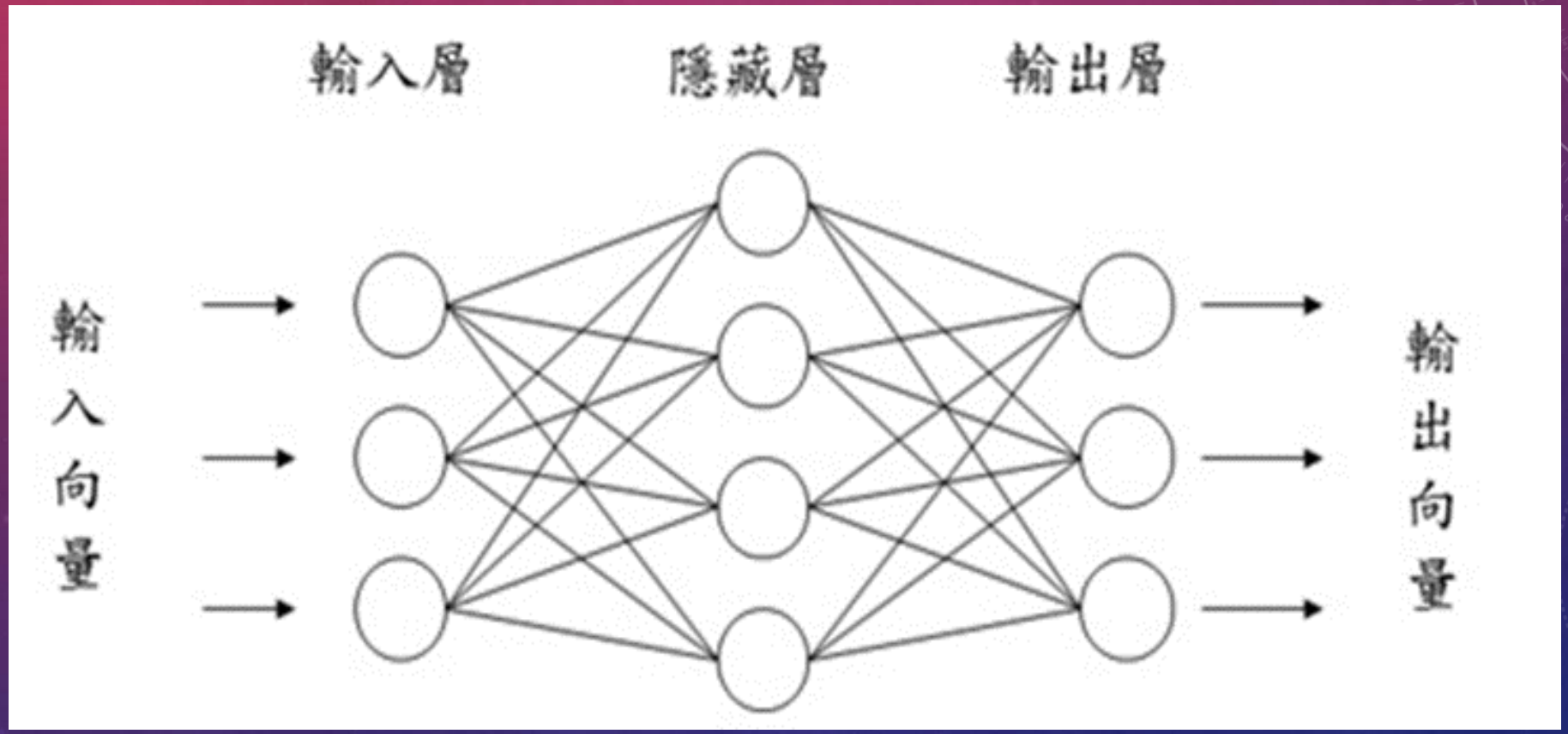
[Case 2]
y = -1 錯分成 y = 1

$$\tan\alpha = \lim_{\Delta x \to 0} \tan\varphi = \lim_{\Delta x \to 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}$$

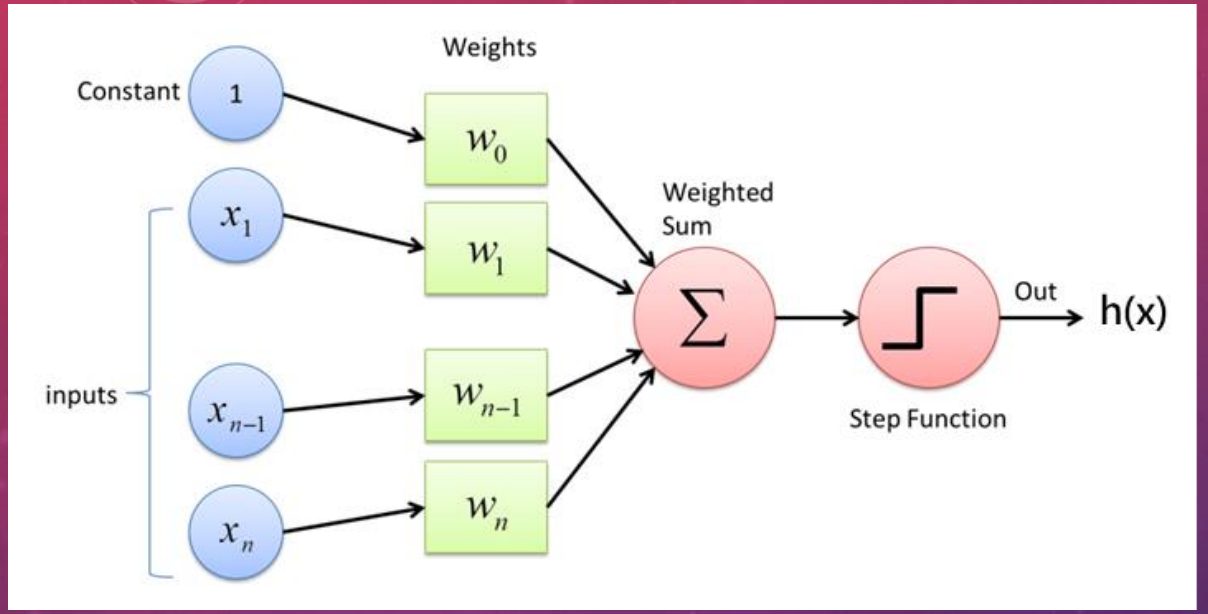




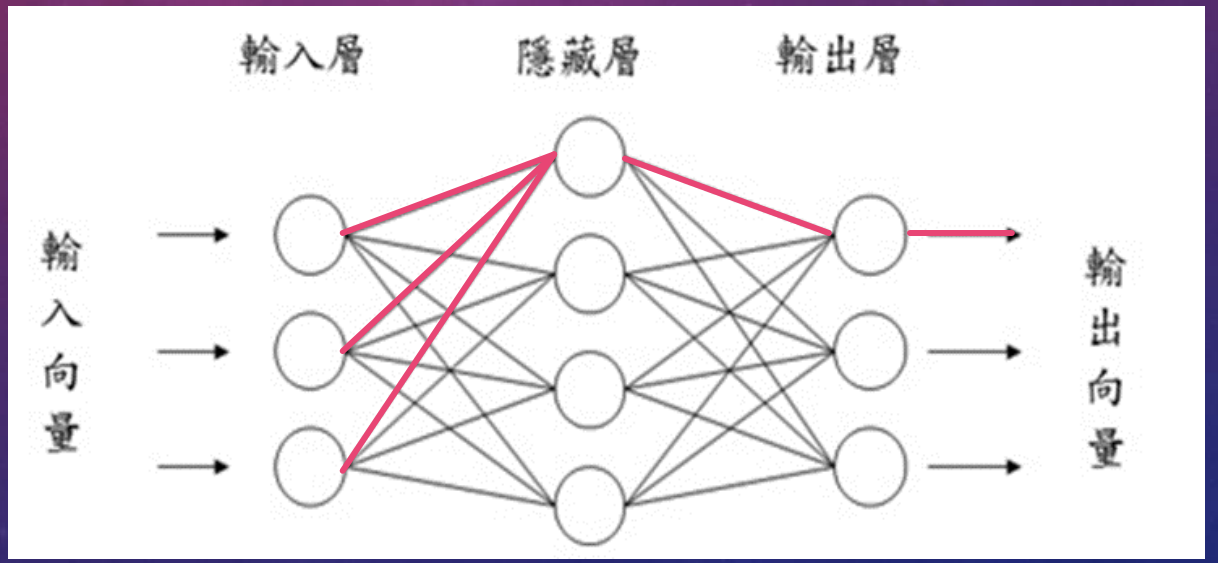
Multi-Layer Perceptron (MLP)



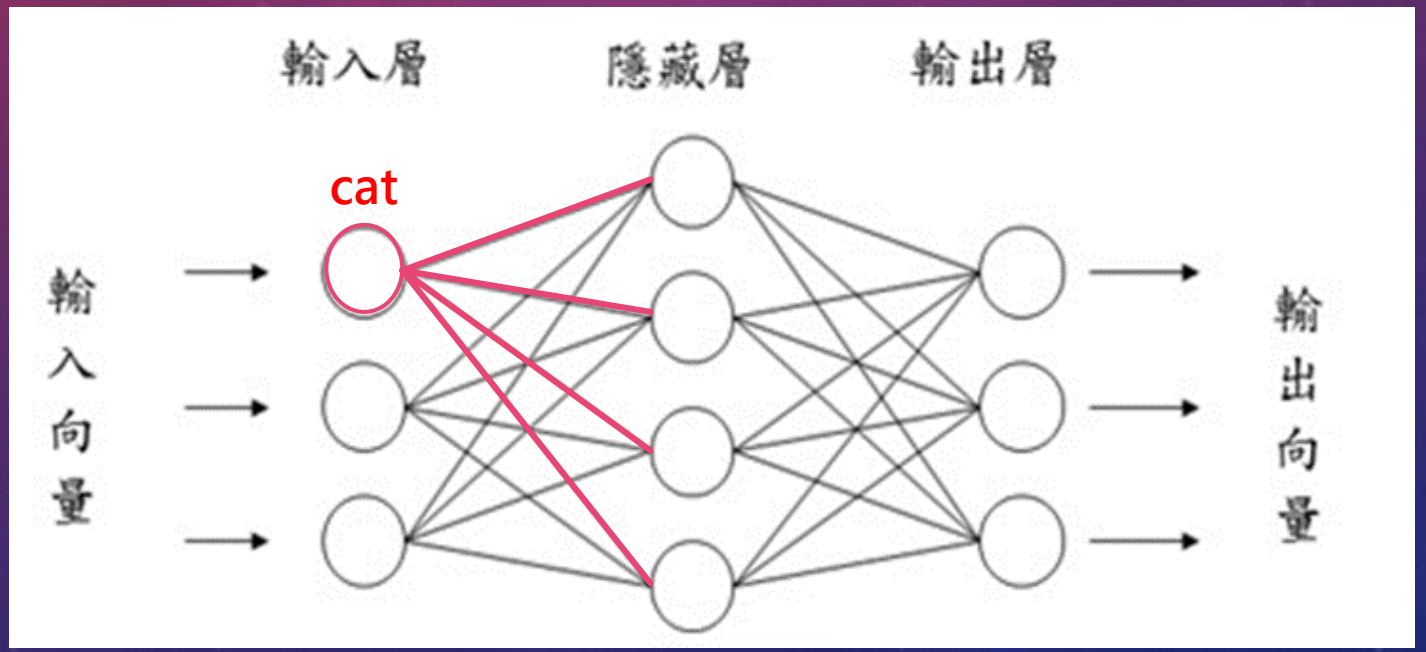
PLA

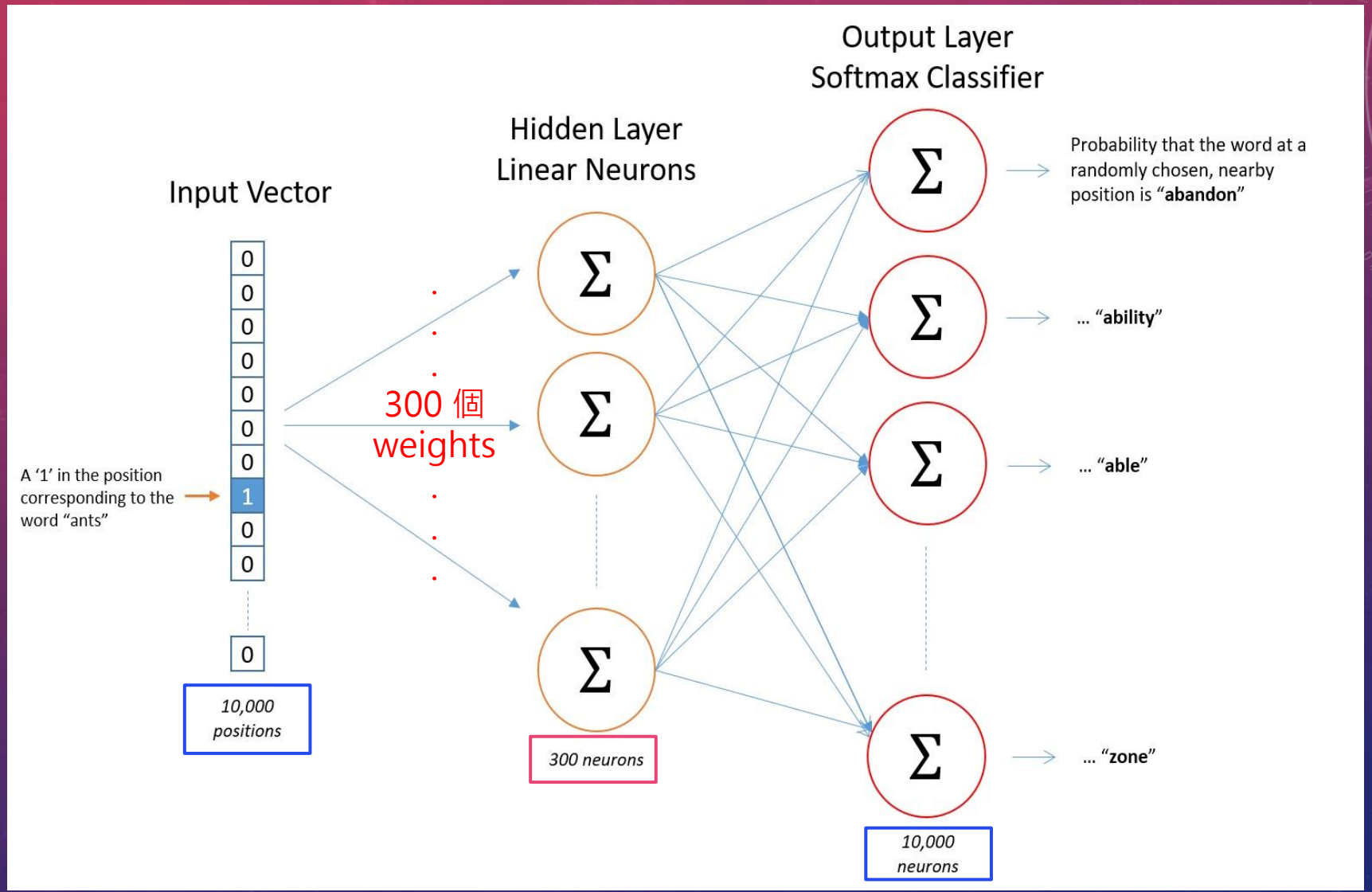

Weights
Constant
1
w_0
x_1
w_1
Weighted
Sum
Σ
Out
h(x)
inputs
x_{n-1}
w_{n-1}
x_n
w_n
Step Function

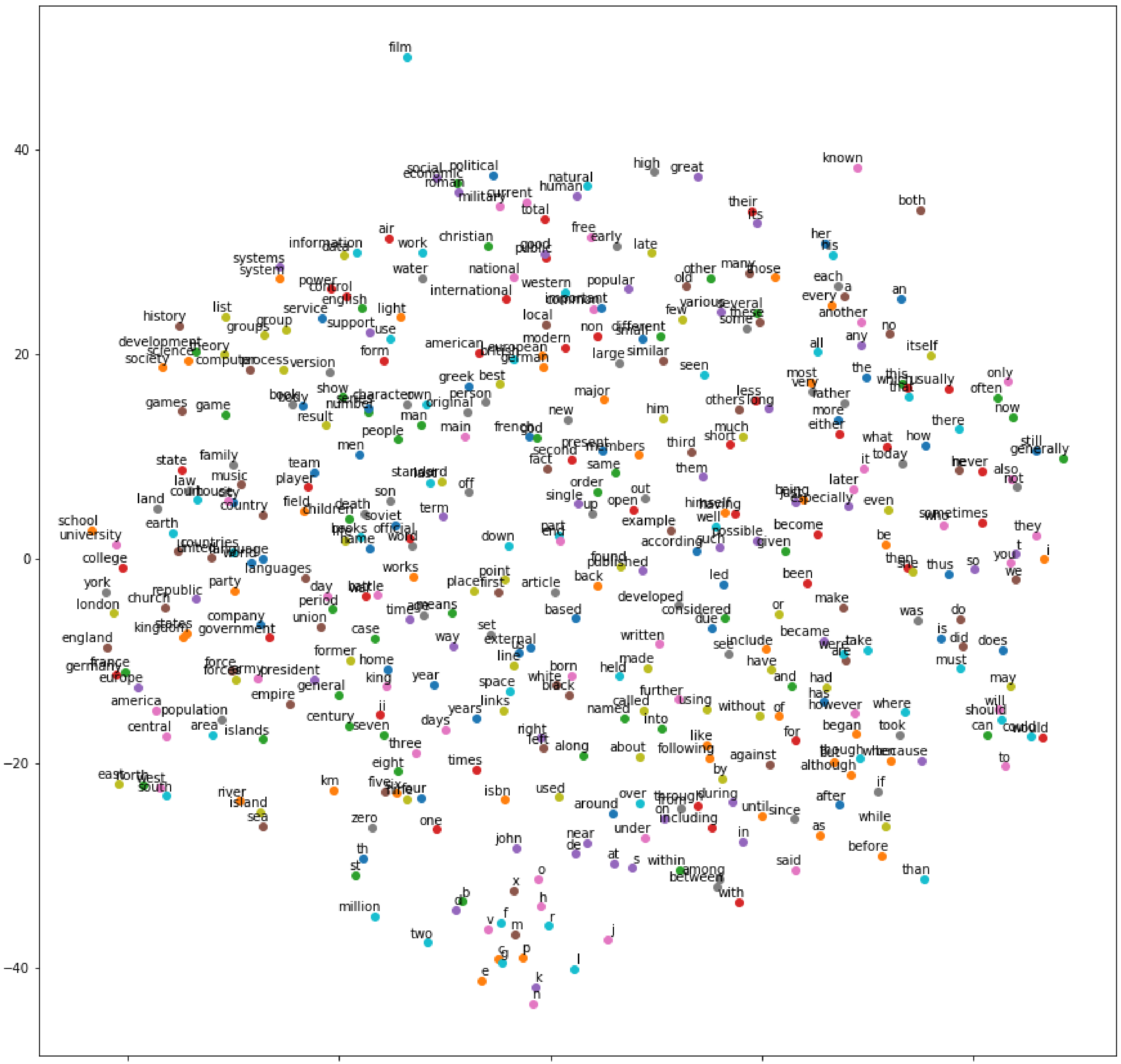
MLP


輸入層
隱藏層
輸出層
輸入向量
輸出向量

```
The -> [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
cat -> [0. 0. 1. 0. 0. 0. 0. 0. 0. 0.]
jump -> [0. 0. 0. 1. 0. 0. 0. 0. 0. 0.]
over -> [0. 0. 0. 0. 1. 0. 0. 0. 0. 0.]
the -> [0. 0. 0. 0. 0. 1. 0. 0. 0. 0.]
dog -> [0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
The -> [0. 1. 0. 0. 0. 0. 0. 0. 0. 0.]
dog -> [0. 0. 0. 0. 0. 0. 1. 0. 0. 0.]
ate -> [0. 0. 0. 0. 0. 0. 0. 1. 0. 0.]
my -> [0. 0. 0. 0. 0. 0. 0. 0. 1. 0.]
homework -> [0. 0. 0. 0. 0. 0. 0. 0. 0. 1.]
```




Output Layer
Softmax Classifier
Hidden Layer
Linear Neurons
Input Vector
Probability that the word at a randomly chosen, nearby position is "abandon"
... "ability"
... "able"
... "zone"
A '1' in the position corresponding to the word "ants"
0
0
0
0
0
0
0
1
0
0
0
10,000 positions
300 個 weights
300 neurons
10,000 neurons



film
40
20
0
−20
−40
known
high
great
social
political
economic
roman
natural
human
military
current
total
their
its
both
free
early
late
air
work
christian
information
data
her
his
systems
system
water
national
many
other
those
each
power
control
english
international
western
popular
old
every
a
an
history
list
service
light
local
few
various
several
another
group
groups
support
use
non
different
small
these
some
no
any
all
itself
development
theory
science
society
computer
process
version
form
american
european
british
modern
german
large
similar
seen
the
most
this
very
usually
only
often
greek
best
major
less
long
rather
book
show
person
own
original
new
him
others
more
now
games
game
result
number
man
main
french
much
either
there
what
how
still
generally
people
present
third
short
second
members
fact
same
today
men
it
never
also
not
state
family
team
them
later
music
player
standard
off
order
son
single
up
out
open
law
country
field
death
children
soviet
term
land
especially
even
sometimes
who
school
university
earth
books
official
part
end
example
become
be
they
well
possible
according
such
given
i
countries
united
language
name
world
down
college
languages
found
published
been
works
point
place
first
article
back
then
thus
so
you
we
york
day
battle
war
led
republic
party
church
period
developed
london
time
means
based
considered
make
was
do
union
company
due
kingdom
states
government
case
set
written
include
became
is
did
does
england
external
way
see
were
take
former
line
made
have
france
home
germany
born
held
europe
army
president
force
forces
king
year
space
white
black
must
and
had
general
empire
further
has
may
america
using
links
named
called
without
of
however
where
will
population
years
should
into
central
century
days
right
left
seven
along
like
about
began
took
can
would
could
area
islands
three
following
against
for
though
but
whether
because
times
by
although
to
eight
south
west
east
north
five
six
four
km
isbn
used
river
island
over
through
from
during
after
if
around
until
since
on
including
sea
zero
one
while
as
under
near
de
john
in
th
at
within
among
between
said
before
than
s
st
o
with
x
h
million
b
r
f
d
v
j
m
two
g
p
l
e
c
k
n

THANK YOU