機器學習、文本分類模型與融合式人工智慧

台師大通識教育課程

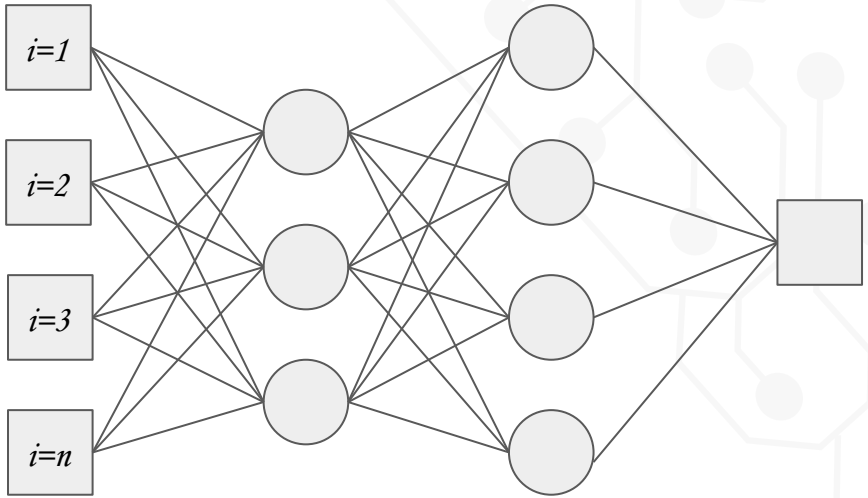
文本分析與程式設計

授課：卓騰語言科技 _ PeterWolf

機器學習究竟學習了什麼？

- 類神經網路就是智慧嗎？
- 智慧就是類神經網路嗎？

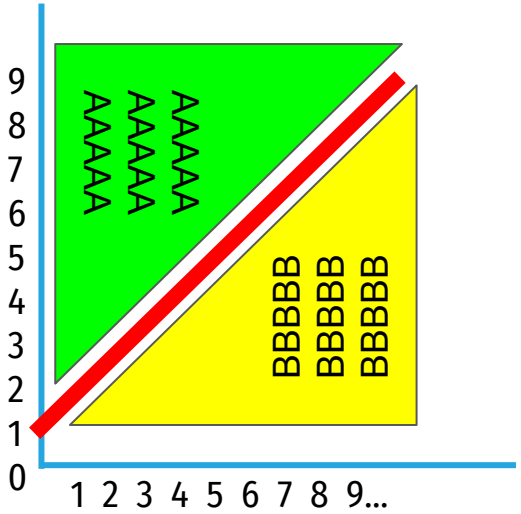
$$f(\sum_{i=1}^{n} w_i x_i + b)$$



如何描述一條直線？


AAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAA
BBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBB
BBBBBBBBBBBBBBBBBBBBBBBBB
f(x) = 6
9
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9...

如何描述一條斜線？



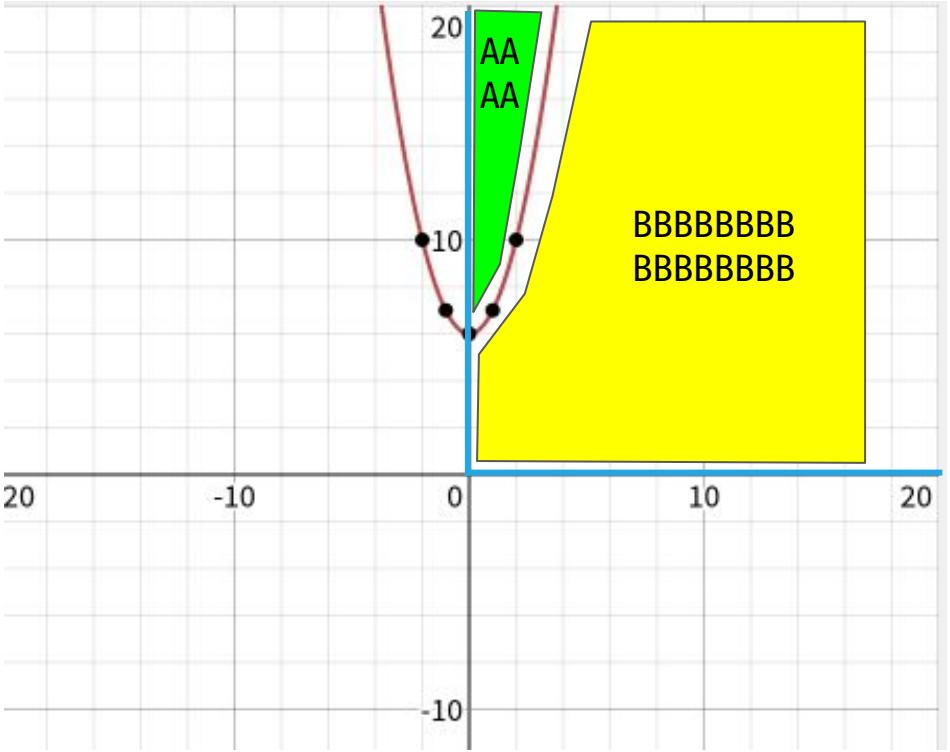
f(0) = 1
f(1) = 2
f(2) = 3

...

f(x) = x+1

如何描述一條曲線？

https://www.mathway.com/popular-problems/Calculus/504676



$f(0) = 6$

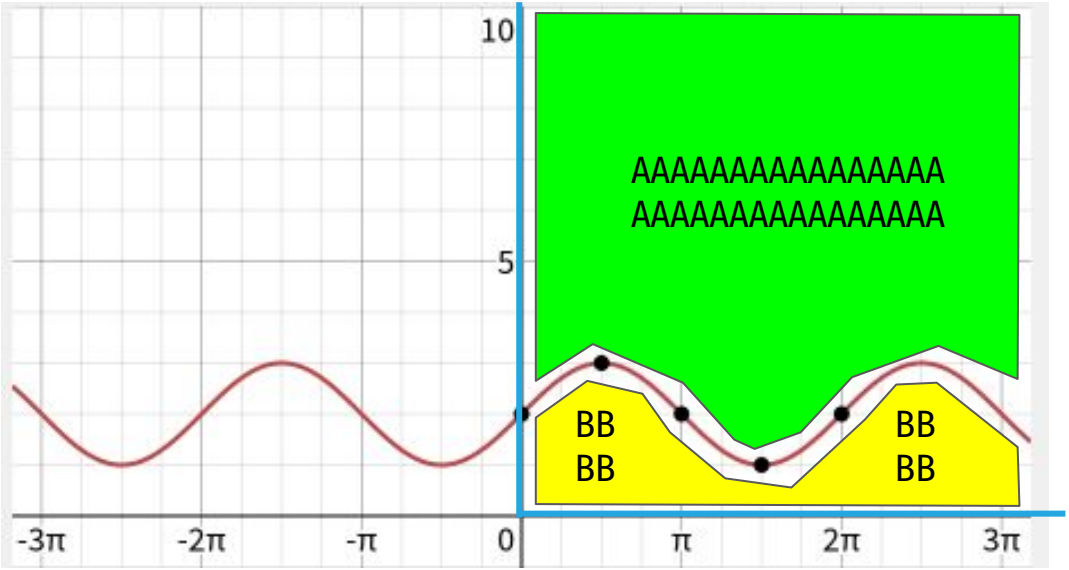
$f(1) = 7$

$f(2) = 10$

...

$f(x) = x^2+6$

如何描述一條規律地扭來扭去的曲線？



f(x) = sine(x)+2

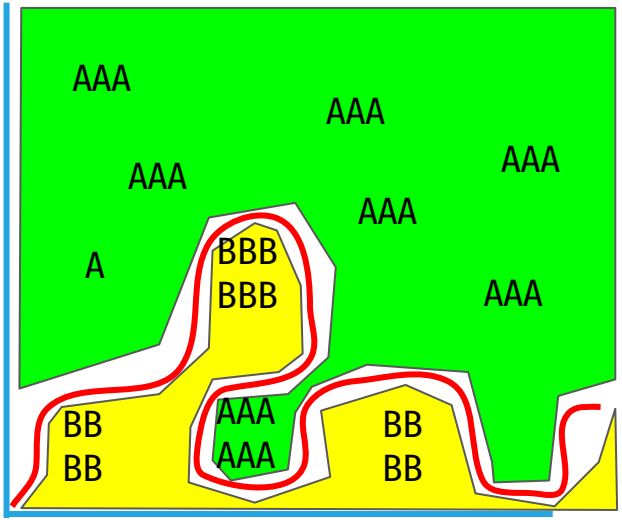
如何描述一條好像不規律地扭來扭去的曲線？

$$f(x) = \sin(x)+2 + \sin(2x)+3$$

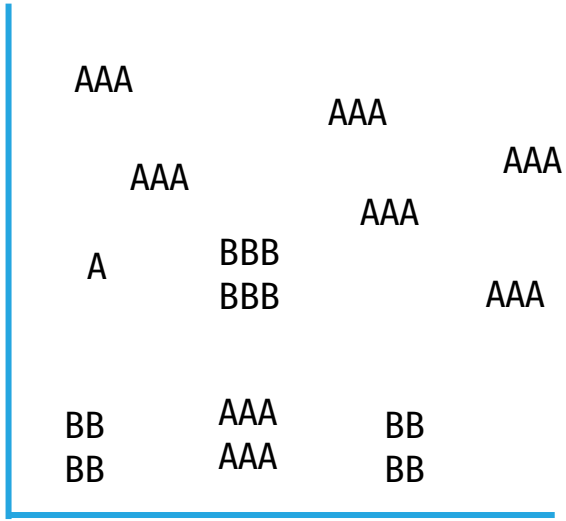
https://zh.wikipedia.org/zh-tw/傅立葉級數

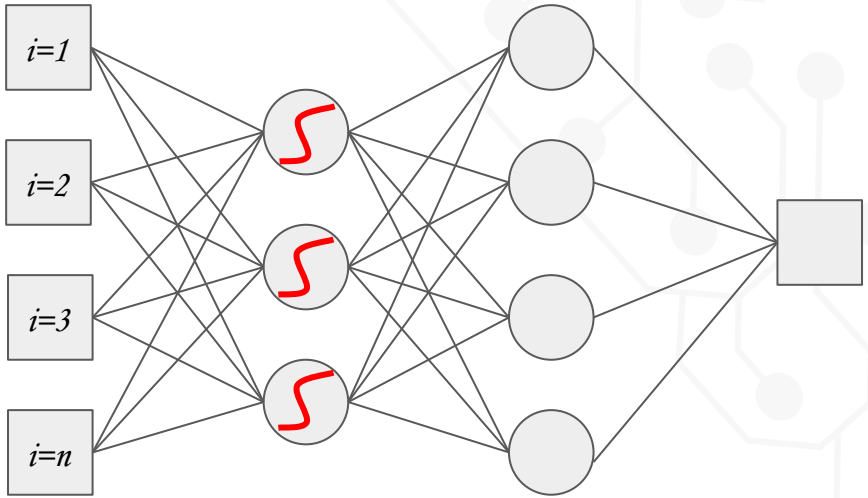
[警告]前方高能！
非修課人員請儘速撤離！

如何描述一條不曉得在扭什麼鬼的曲線？



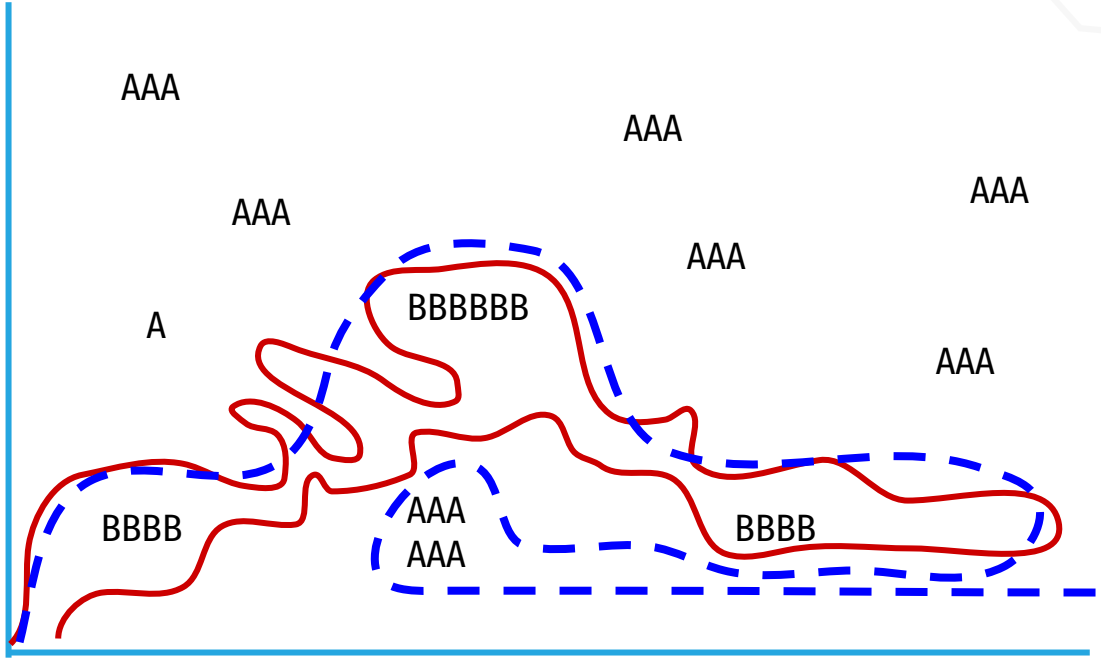
人不描述它，讓機器用類神經網路去「逼近」這條線





https://youtu.be/CqOfi41LfDw?t=475

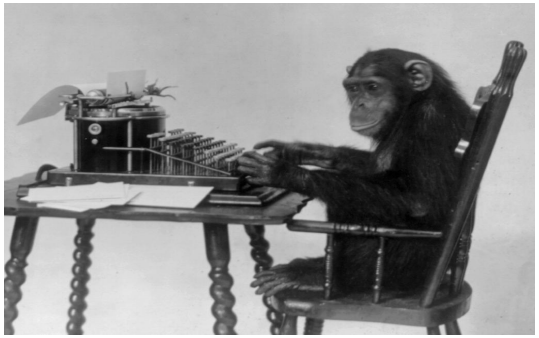
決定哪一條線最好？
是因為「邏輯」因為「分佈」或其它因素？



https://mropengate.blogspot.com/2015/06/ch15-4-neural-network.html

無限猴子定理

https://zh.wikipedia.org/wiki/無限猴子

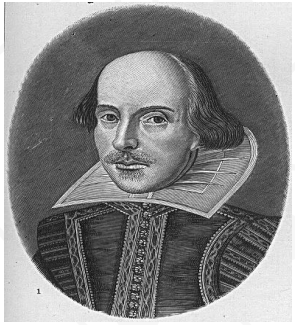


×

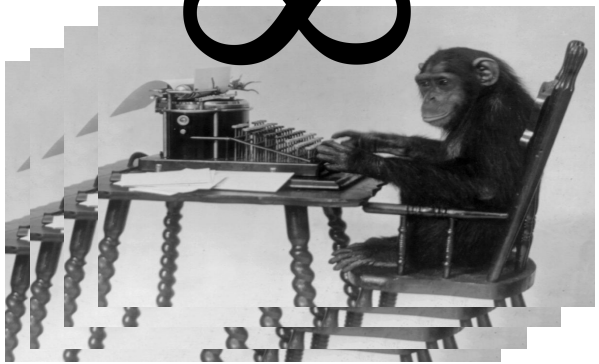
∞

time

=



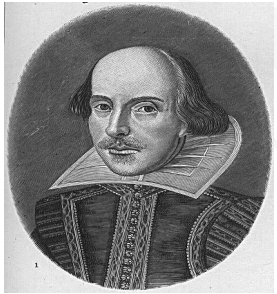
∞



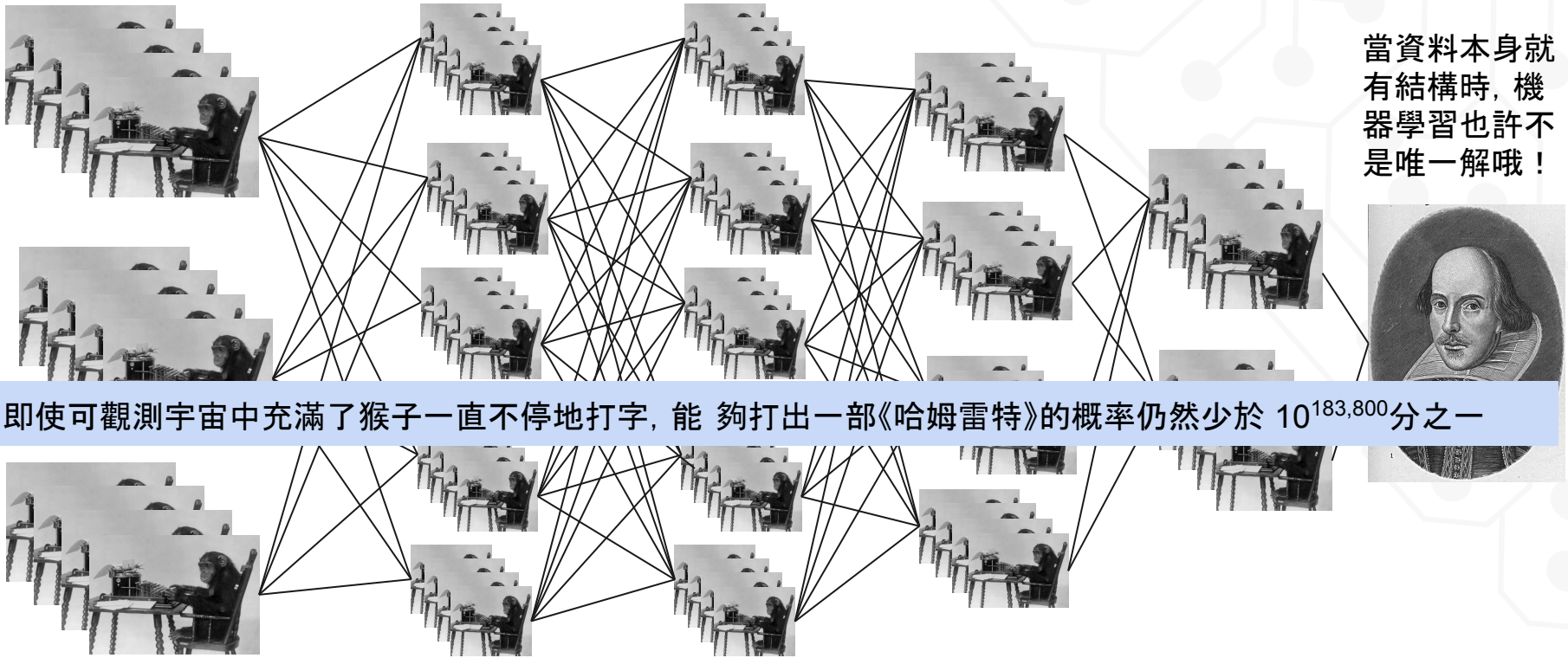
×

1 sec.
time

=



深度無限猴子 (ML : Monkey Learning)



當資料本身就有結構時，機器學習也許不是唯一解哦！

即使可觀測宇宙中充滿了猴子一直不停地打字，能 夠打出一部《哈姆雷特》的概率仍然少於 $10^{183,800}$分之一

NLP 的機器模型

先決條件：

大量的資料：最好什麼資料都有

平衡的分佈：最好資料分佈的模樣和語言系統一致

優點：

執行分類任務時，速度很快。不論是模型的資料型態或是硬體的電路、晶片設計都是為了讓機器學習的模型可以快速被載入執行設計。(i.e., 頻寬很大、容量很大、平行處理能力強...等)

缺點：

執行資料學習、產生模型的速度很慢。

粗略的分類(分類就是智慧嗎？)

缺乏邏輯、因果和知識的理解能力

缺乏一致性和呈現語言系統內部結構的能力

非機器學習 NLP

先決條件:

對應用場景的領域知識或是應用目標的理解

優點:

對文本細節的理解。
不需訓練模型，開箱即用。
容易實現邏輯、因果、知識的高階應用。
輸出一致性高，系統硬體需求低，容易佈署。

缺點:

開發團隊需具備領域知識
開發團隊需瞭解應用目標

何不摻在一起用呢？

1. 利用「機器學習模型」粗分類的結果，再給「非模型」的解法處理。將大幅提高最終結果的正確率，也降低了全部丟給「非模型」的解法時，因為硬體並不是為了這種方法優化時損失的優勢。

2. 資料「少」時，先用「非模型」的解法實作應用，待資料「夠多」，足以產生效果優良的模型時，再採用前述 1 的架構調整應用。如此一來，不論資料「夠」或「不夠」都能實作 NLP 的應用。

範例一：人物研究

1. 步驟：
   a. 盡可能從不同類別和資料來源收集所有關於目標人物的原始資料文本。
   b. 利用「機器學習模型」粗略地分類文本類別：
      i. 新聞類：信任度最高
      ii. PTT、DCard...次之
      iii. 其它來源…再次之
   c. 對不同類別的文本設計不同的資訊擷取邏輯，並利用「非機器學習」的工具，實作這些邏輯成為程式。

範例一：金融應用

1. 步驟：
    a. 將行情的「漲」和「跌」視為兩種不同的類別。
    b. 利用「機器學習模型」粗略地分類文本類別。
    c. 對「漲」和「跌」兩個類別的文本設計不同的資訊擷取邏輯，並利用「非機器學習」的工具，實作這些邏輯成為程式。
    d. 進一步確認「漲」的理由是否合理；「跌」的理由是否合理。
    e. 依合理的資訊進行操作

Quiz:

課堂中說明了機器學習方法最基礎的原理，請思考以下問題。

1. 為什麼「大量數據」是機器學習的必備條件？
2. 類神經網路就是智慧嗎？智慧就是類神經網路嗎？
3. 回到課程最原始的起點，請說明以下概念：
    a. 什麼是「字」、什麼是「詞」？電腦是否能區辨兩者的差別？
    b. 詞頻能否解決前述的「字/詞」之分，或是呈現語言的內部結構？
    c. 向量能否解決前述的「字/詞」之分，或是呈現語言的內部結構？
    d. 機器學習能否解決前述的「字/詞」之分，或是呈現語言的內部結構？
    e. Articut 藉由呈現語言的內部結構，可解決哪些問題？

Assignment: 小組作業，每組繳一份至你們的「組名目錄」即可

1. 從課程 github repo 中把課程中提供的 week14 的目錄 git pull 下來。
2. 把 Project.ods 改名為 **Project_分組隊名.ods**
3. 利用前述的簡報範本，完成期末專題的第一次提案。