**HW1**

2019/3/21

林佳緯

107598012

1. In the lecture, we mentioned that different algorithms have different problems. Use your own words to explain the shortcomings of each of the following methods:

* Neural networks (particularly CNN)

缺點:模型容易出現梯度消散問題、需要大樣本量，物理含義不明確、平移不變性、Pooling 層會丟失大量的有價值信息。

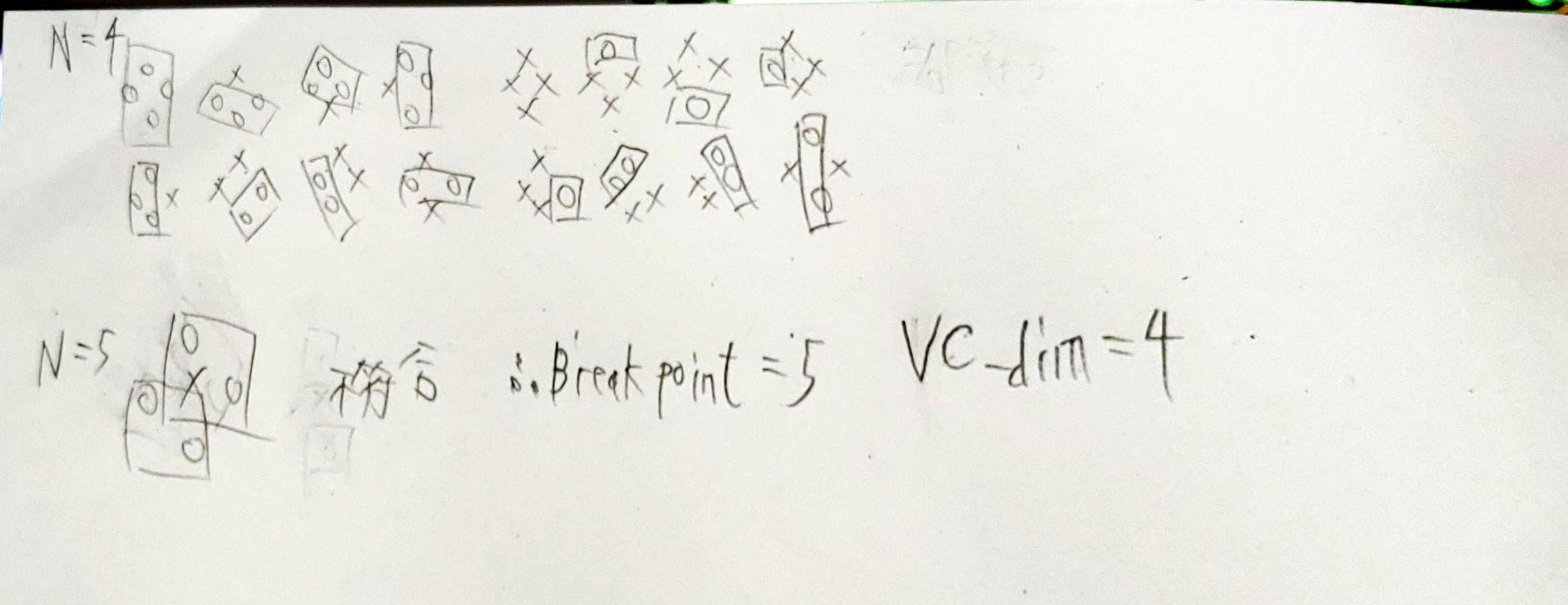
* C4.5 decision tree

缺點: 很容易在訓練資料中產生複雜的樹狀結構，造成overfitting的問題。減少樹枝節點可以緩解overfitting的負作用，常用方法是限制樹的高度、葉子節點中的最少樣本數量。決策樹模型無法表示XOR，相乘的概念。

* Adaboost

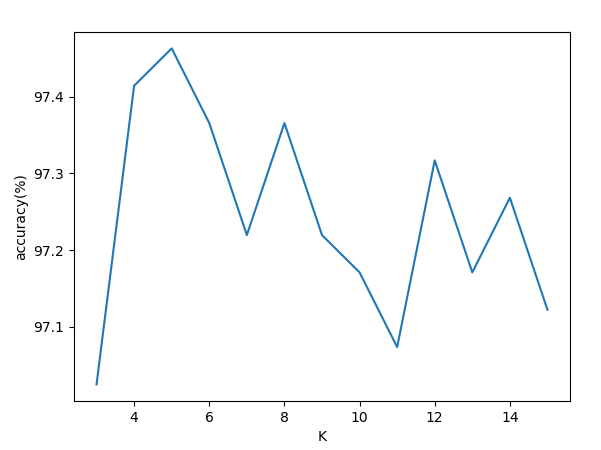
缺點: AdaBoost疊代次數也就是弱分類器數目不太好設定，可以使用交叉驗證來進行確定，數據不平衡導致分類精度下降，訓練比較耗時。

1. The textbook claims that the set of rectangles in 𝑅 2 has a VC dimension of 4. Why is it the case? You may draw plots to show it .



1. UC Irvine has a large repository for various kinds of data. In this problem, you are asked to use the dataset of “Breast Cancer Wisconsin (Original) Data Set” (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original %29) to perform the experiments. To simplify the problem, you just need to classify whether the subject is benign or malignant (attribute 11). Implement the k-NN classifier for the classification task. To begin one experiment, randomly draw 70 % of the instances from each class for training and the rest are for testing. Repeat the drawing and the k-NN classification 10 times and compute the average accuracy. Then, plot the curve of k versus accuracy for k = 3, …, 15. For simplicity, use the Euclidean distance in your computation.

* Source code連結: <https://github.com/linitachi/machine-learning>



k=3的平均正確率: 97.02439024390243%

k=4的平均正確率: 97.41463414634146%

k=5的平均正確率: 97.46341463414635%

k=6的平均正確率: 97.36585365853658%

k=7的平均正確率: 97.21951219512195%

k=8的平均正確率: 97.36585365853658%

k=9的平均正確率: 97.21951219512195%

k=10的平均正確率: 97.17073170731707%

k=11的平均正確率: 97.07317073170731%

k=12的平均正確率: 97.3170731707317%

k=13的平均正確率: 97.17073170731707%

k=14的平均正確率: 97.26829268292683%

k=15的平均正確率: 97.1219512195122%

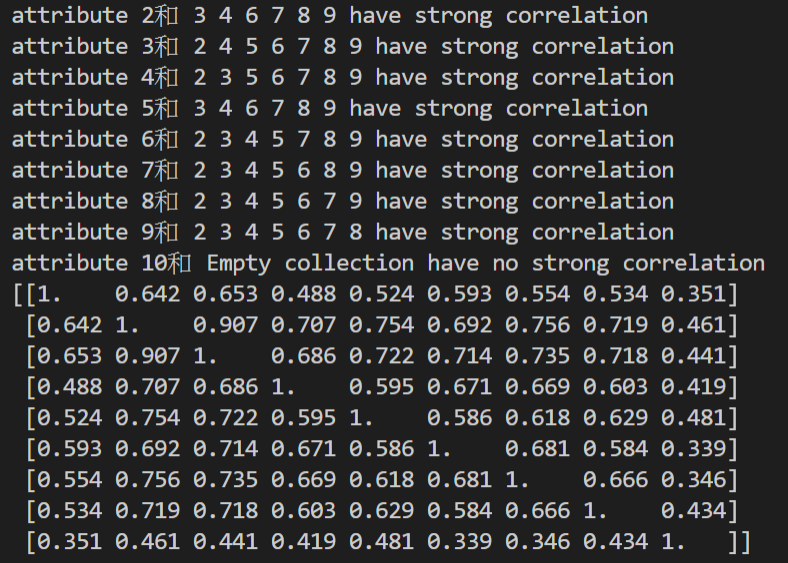
* Source code:





1. Following problem 3, compute the covariance matrix of the dataset. The matrix is of size 9 × 9 (attribute 2 – 10). Do you see strong correlation between any two attributes?

* attribute 2和 3 4 6 7 8 9 have strong correlation
* attribute 3和 2 4 5 6 7 8 9 have strong correlation
* attribute 4和 2 3 5 6 7 8 9 have strong correlation
* attribute 5和 3 4 6 7 8 9 have strong correlation
* attribute 6和 2 3 4 5 7 8 9 have strong correlation
* attribute 7和 2 3 4 5 6 8 9 have strong correlation
* attribute 8和 2 3 4 5 6 7 9 have strong correlation
* attribute 9和 2 3 4 5 6 7 8 have strong correlation
* attribute 10沒有和其他attribute 有strong correlation

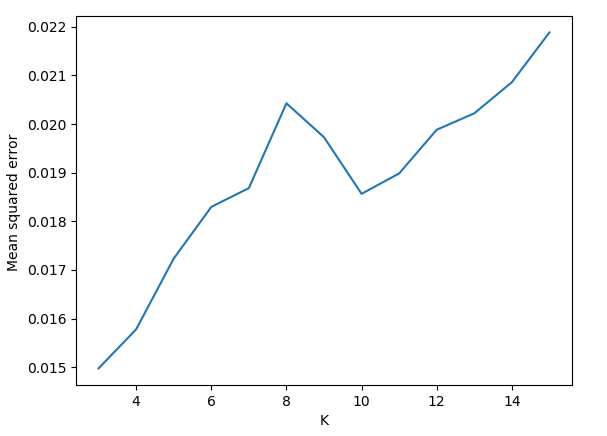


* Source code:

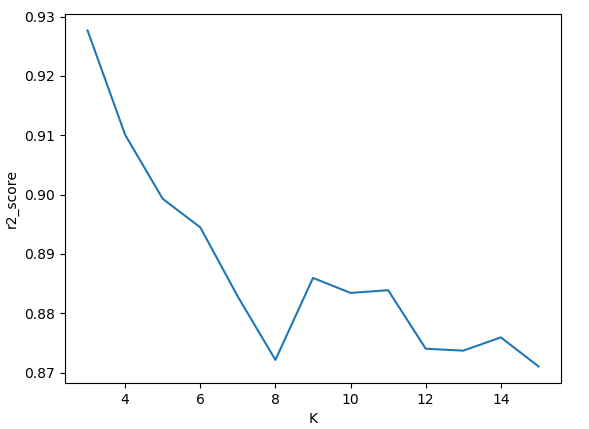


1. Consult any paper to learn how to extend the k-NN approach to perform regression. Based on your findings, implement the program and test it on “Computer Hardware Data Set,” which is available at https://archive.ics.uci.edu/ml/datasets/Computer+Hardware for downloading.

* 首先先把dataset分成70%的trainset跟30%的testset，接著把attribute 9當成要預測的值，透過KNN regression預測是否正確，總共跑K =3~15 ，每個K跑10次，取MSE以及R2 score。



MSE圖



R2\_score圖

* Source code:



