K-FOLD CROSS-VALIDATION

交叉驗證 (Cross Validation)，有時亦稱循環估計，是一種統計學上將數據樣本切割成較小子集的實用方法。於是可以先在一個子集上做分析，而其它子集則用來做後續對此分析的確認及驗證。一開始的子集被稱為訓練集(training set)。而其它的子集則被稱為驗證集(validation set)或測試集(testing set)。

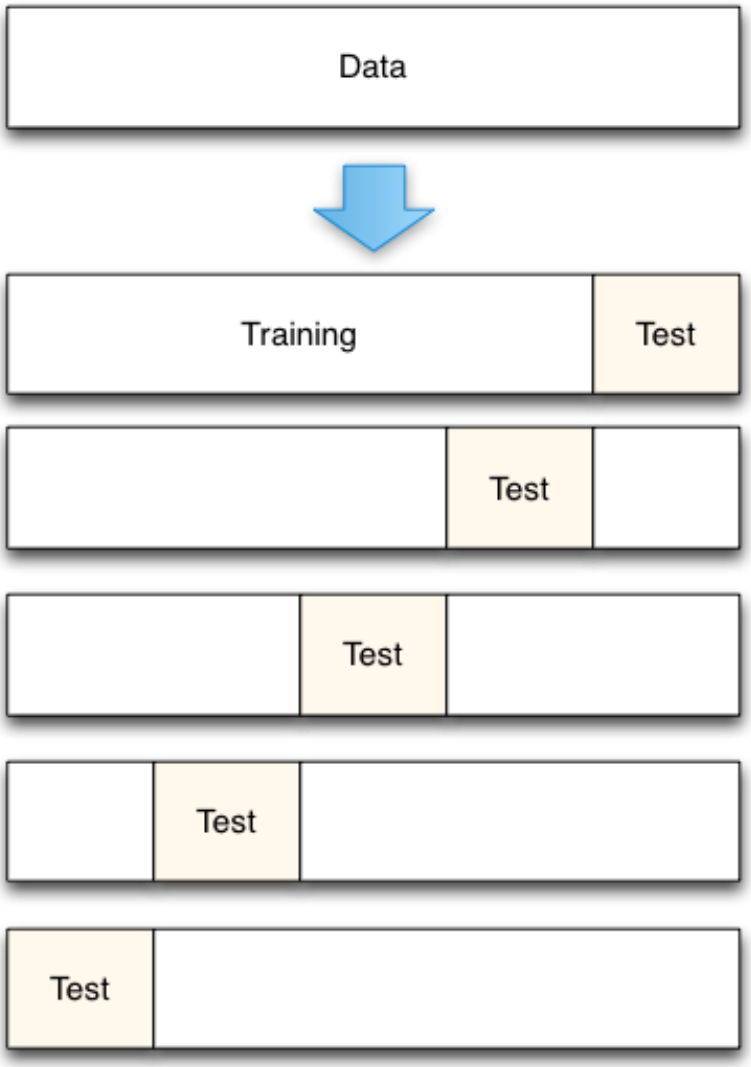
✓ **缺點：可能會選到不具代表性的訓練集與測試集組合**

● K-FOLD CROSS-VALIDATION

✧ **執行步驟**

1. 不重複抽樣將原始數據隨機分為 k 份。
2. 每一次挑選其中 1 份作為測試集，剩餘 k-1 份作為訓練集用於模型訓練。
3. 重複第二步 k 次，這樣每個子集都有一次機會作為測試集，其餘機會作為訓練集。
4. 在每個訓練集上訓練後得到一個模型，
5. 用這個模型在相應的測試集上測試，計算並保存模型的評估指標，

6. 計算 k 組測試結果的平均值作為模型精度的估計，並作為當前 k 折交叉驗證下模型的性能指標。



✓ 優點：每一個取樣資料，最後都扮演過訓練與測試的角色。

✓ 缺點：K 值選取較為困難

✧ **k 一般取 10**

之所以選擇將數據集分為 10 份，是因為通過利用大量數據集、使用不同學習技術進行的大量試驗，表明 10 折是獲得最好誤差估計的恰當選擇，而且也有一些理論根據可以證明這一點。但這並非最終診斷，爭議仍然存在，而且似乎 5 折或者 20 折與 10 折所得出的結果也相差無幾

除此之外，

1. 多次 k 折交叉驗證再求均值，例如：10 次 10 折交叉驗證，以求更精確一點。
2. 劃分時有多種方法，例如對非平衡數據可以用分層採樣，就是在每一份子集中都保持和原始數據集相同的類別比例。

Grid search

✧　先在較大範圍內進行搜索，再從結果比較好的附近區域進行精確搜索。

實驗法是指通過大量的實驗比較來確定參數，這種方法十分浪費時間，且不易尋得最優參數；Grid search 是將待搜索參數在一定的空間範圍中劃分成網格，通過遍歷網格中所有的點來尋找最優參數。

- 由於網格內多數參數組對應的分類準確率都非常低，只在一個比較小的區間內的參數組所對應的分類準確率很高，所以遍歷網格內所有參數組會相當浪費時間。
- 針對 Grid search 搜索時間長的問題提出一種改進的網格搜索法，先採用大步距大範圍粗搜，初步確定一個最優參數區間，之後在此區間內進行小步距精搜，大幅度地減少了參數尋優時間。
- 在尋得了局部最優參數之後，再在這組參數附近選擇一個小區間，採用傳統方法中的小步距進行二次精搜，找到最終的最優參數. 如果參數區間選擇合理，那麼改進的網格搜索法能夠搜索出全局最優的參數，但由於這個區間的選擇含有比較多的經驗成分，所以通常情況下只能得到局部最優的

參數。因此，這種改進的網格搜索法相當於犧牲了分類準確度，來減少大量的搜索時間。