時間序列資料切分

在機器學習中, 無論其目的是分類, 預測或是分析用途
1.都會Random的預留10%左右的資料作為最後的測試用, 此也是最後模型預測精確率, recall等指標的來源
2. 剩下90%的資料也Random分成8(訓練) : 2(驗證)用, 或是用K-Fold來作訓練及驗證更佳
-
如果是股票分析, 天氣預測等有 (時間, 前/後) 順序的資料作預測時
問題1: 應該是只能用資料最後段的10%來作最後的測試用, 對吧?
問題2: 那剩下90%的資料也需 (前8成)作訓練 : (後2成)作驗證用嗎?
或是此8:2可以Random分呢?
問題3: 有 (時間, 前/後) 順序的資料, 作訓練/驗證時, 應該不適合用K-Fold對吧?

你的理解在**一般機器學習（非時間序列）**是正確的 👍
但只要資料「有時間順序」（例如股票、天氣、病人追蹤、IoT感測資料），**切分方式就完全不同**。

下面我幫你系統化整理。

一、一般機器學習（無時間順序）

例如：

- 影像分類
- 醫療影像 segmentation（像你做蛀牙）
- 表格資料分類

✅ 作法：

1. Random 留 10~20% 作 test set
2. 剩下資料 random 分成 train / validation
3. 或使用 K-Fold Cross Validation

這是標準流程 ✔️

二、時間序列資料（股票、天氣、需求預測）

例如：

- 股票價格預測
- 天氣預測
- 用電量預測
- 病人未來復發預測

⚠️ 核心原則：

> **絕對不能讓未來資料出現在訓練資料中**

這叫做：

> 🚨 Data Leakage（資料洩漏）

問題1

> 應該是只能用資料最後段的10%來作最後的測試用, 對吧?

✅ 對！完全正確。

時間序列一定是：

```
過去 ----------------------> 未來
[ Train ][ Validation ][ Test ]
```

例如：

時間範圍	用途
2015–2021	Train
2022	Validation
2023	Test

Test 必須是「最新、最後那一段」

因為你真正的應用情境是：

> 用過去資料預測未來

問題2

> 剩下90%的資料也需 (前8成)作訓練 : (後2成)作驗證用嗎?
> 或是此8:2可以Random分呢?

❌ 不能 random 分

時間序列永遠要保持順序。

正確方式：

```
前 70%  → Train
中間 20% → Validation
最後 10% → Test
```

或

```
前 80% → Train
後 20% → Validation
最後再保留 Test
```

但都必須是「前→後」切。

🚫 不能：

```
隨機抽 20% 當 validation
```

那會讓模型偷看到未來。

問題3

> 有時間順序的資料, 作訓練/驗證時, 應該不適合用K-Fold對吧?

✅ 一般K-Fold 不適合

因為一般 K-Fold 是：

```
Fold1: random
Fold2: random
Fold3: random
```

這會打亂時間順序。

但可以用「Time Series Cross Validation」

這種方法叫：

> Rolling Window / Expanding Window

例如：

方法1：Expanding Window

```
Train: 2015
Valid: 2016

Train: 2015–2016
Valid: 2017

Train: 2015–2017
Valid: 2018
```

訓練資料會越來越多。

方法2：Sliding Window

```
Train: 2015–2017
Valid: 2018

Train: 2016–2018
Valid: 2019
```

訓練資料固定長度往前滑動。

在 sklearn 裡有：

```
TimeSeriesSplit
```

這就是為時間序列設計的 K-Fold。

總結對照表

問題	正確答案
Test 是否一定是最後一段？	✅ 是
Validation 可以 random 嗎？	❌ 不行
一般 K-Fold 可以用嗎？	❌ 不行
有替代方法嗎？	✅ TimeSeriesSplit

一句話核心觀念

> 有時間順序的資料，切分一定是「前訓練、後驗證、最後測試」，絕對不能 random。

如果你願意，我可以幫你畫一張：

- 一般 ML 切分圖
- 時間序列切分圖
- Rolling Window 示意圖

讓你在上課或論文寫 methodology 時直接用 👌