**📘 Lecture Notes: N-Fold Cross-Validation**

**1. Problem: Small datasets**

* **Standard approach: split dataset into training, validation, and test sets.**
* **Issue: With small datasets:**
  + **Splitting reduces the amount of training data → model may not learn well.**
  + **Validation set might also be too small → unreliable performance check.**

**2. Solution: N-Fold Cross-Validation**

* **Combines training and validation steps cleverly while still keeping a separate test set.**
* **Goal: Use as much data as possible for training while still validating.**

**3. How It Works (Example: 10-Fold Cross-Validation)**

* **Suppose dataset has 11,000 observations:**
  + **Reserve 1,000 for test → 10,000 left.**
* **Split remaining 10,000 into 10 subsets of 1,000 each.**
* **Process:**
  + **Treat one subset as validation and the other 9 as training.**
  + **Train the model and calculate validation loss.**
  + **Rotate: next subset becomes validation, the rest training.**
  + **Repeat until each subset has served as validation once.**
* **Visual:**
  + **Orange = validation, Blue = training**
  + **Every fold ensures no overlap between training and validation in that round.**

**4. Benefits**

* **Uses almost the entire dataset for training.**
* **Helps small datasets get the most information possible.**

**5. Limitations**

* **Still uses validation data for training eventually, so:**
  + **Slightly increases risk of overfitting.**
* **Tradeoff: Better to have a slightly overfitted model than no model at all.**
* **Not a replacement for the standard approach:**
  + **Whenever possible, use three separate sets (train/validation/test).**
  + **Use N-Fold Cross-Validation only if the dataset is too small.**

**6. Key Takeaways**

* **N-Fold Cross-Validation = smart way to reuse small data for training and validation.**
* **Keeps a separate test set for final evaluation.**
* **Common fold values: 10 (10-Fold CV).**
* **Helps reduce wasted data, but may slightly increase overfitting risk.**

**✅ Summary:**

* **Standard approach: train / validation / test.**
* **If data is too small → use N-Fold Cross-Validation to maximize training data while still validating.**
* **Always keep a separate test set for final evaluation.**