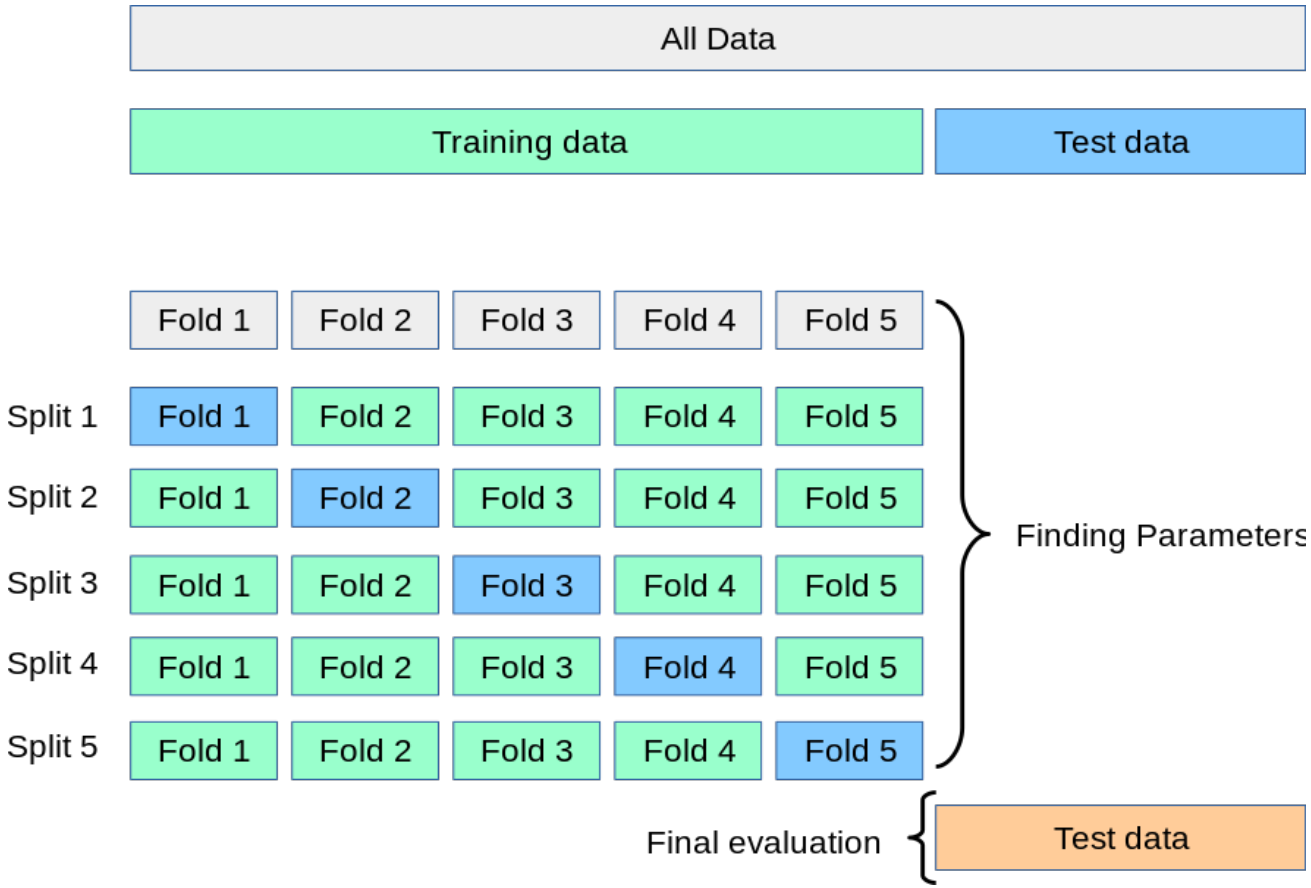
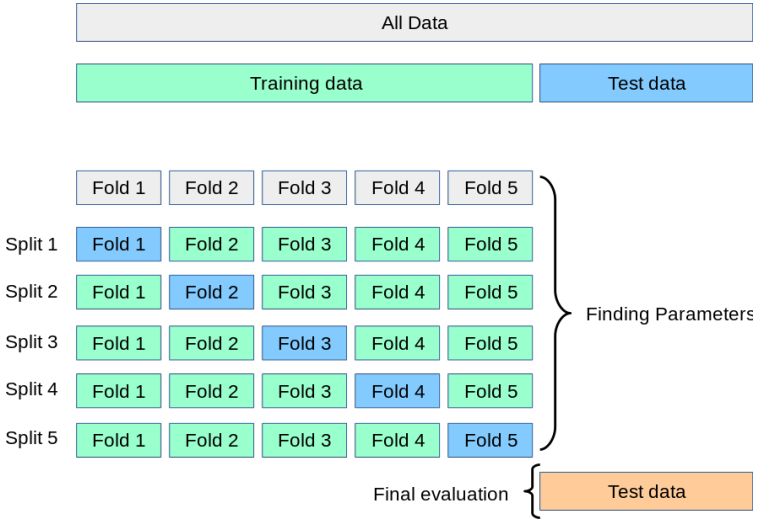
Classification Models Training and Testing Methodology

K-fold cross validation



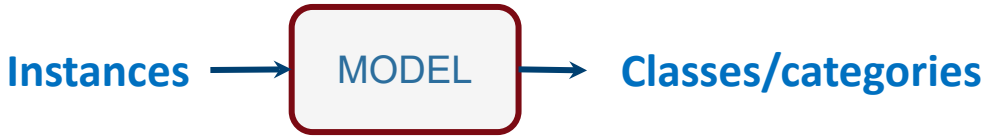
Why K-fold cross validation?



- Avoid inappropriate random [train, test, split] division. Cross-validation can let each sample appear at least once in the validation set
- Through multiple fold divisions, it will provide us the sensitivity information of our model to the training set (worst/best-case performance)
- Higher data utilization

Evaluation Metrics

- How to measure model performance?



- Let's look at an example:
  - 8 instances, 4 positive, 4 negative
  - 2 positives classified correctly: True Positives (TP),
  - 3 negatives classified correctly True Negatives (TN),
  - 2 negatives classified as positives: False Positives (FP),
  - 1 positive classified as negative: False Negative (FN).

Instance	Target	Classified
1	Good	Good
2	Good	Good
3	Good	Bad
4	Bad	Bad
5	Bad	Good
6	Bad	Good
7	Bad	Bad
8	bad	Bad

Evaluation Metrics cont.

- This can be shown as a matrix, called confusion matrix or contingency table.

		True values	
		Positives	Negatives
Predicted values	Positives	TP=2	FP=2
	Negatives	FN=1	TN=3

Instance	Target	Classified	
1	Good	Good	**TP**
2	Good	Good	**TP**
3	Good	Bad	**FN**
4	Bad	Bad	**TN**
5	Bad	Good	**FP**
6	Bad	Good	**FP**
7	Bad	Bad	**TN**
8	bad	Bad	**TN**

What error types (Type I, Type II) these correspond to?

Evaluation Metrics cont.

		True values	
		Positives	Negatives
Predicted values	Positives	TP=2	FP=2 Type II error
	Negatives	FN=1 Type I error	TN=3

Instance	Target	Classified	
1	Good	Good	**TP**
2	Good	Good	**TP**
3	Good	Bad	**FN**
4	Bad	Bad	**TN**
5	Bad	Good	**FP**
6	Bad	Good	**FP**
7	Bad	Bad	**TN**
8	bad	Bad	**TN**

Evaluation Metrics cont.

- The following metrics are derived from the confusion matrix:

Recall or **Sensitivity** or **True Positive Rate**:

$$TPR = TP/(TP + FN)$$

Precision or **Positive Predictive Value**:

$$PPV = TP/(TP + FP)$$

F1 score or harmonic mean of TPR and PPV:

$$F1 = 2 * TPR * \frac{PPV}{TPR+PPV}$$

False Positive Rate:

$$FPR = FP/(FP + TN)$$

		True values	
		Positives	Negatives
Predicted values	Positives	TP=2	FP=2
	Negatives	FN=1	TN=3

Multi class classification

- So far: single class, or binary classification.
- Set of classes may contain more than two, then it is multi-class classification
  - E.g. news articles: Politics, Finance, Sport, each article is classified into ONE of the three.
- It is also possible that an article is classified into more than one category, then it is multi-label categorisation.
  - E.g. news articles: Politics, Finance, Sport, an article is classified into Politics and Finannce.

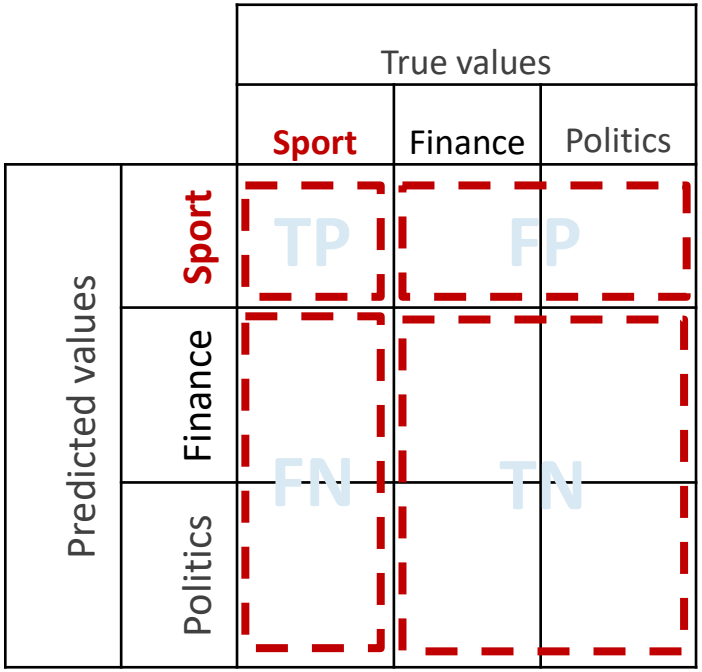
Multi-class Metrics

- How to calculate metrics in this case?
- There are two type of metrics
- 1) Calculate metrics for each class, then average each metric over all classes. This is macro-metrics calculation. E.g.,

$$\text{TPR} = \frac{1}{K}\sum_{i \in K} \frac{TP_i}{TP_i + FN_i}$$

- 2) Add up values for all classes, creating a new confusion matrix 2 x 2, then calculate all metrics as usual. This is called micro-metrics calculation. E.g.,

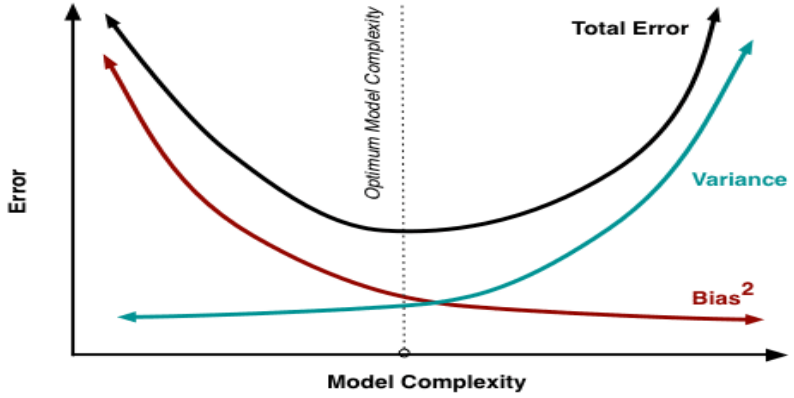
$$TP = \text{TP} = \sum_{i \in K} TP_i$$

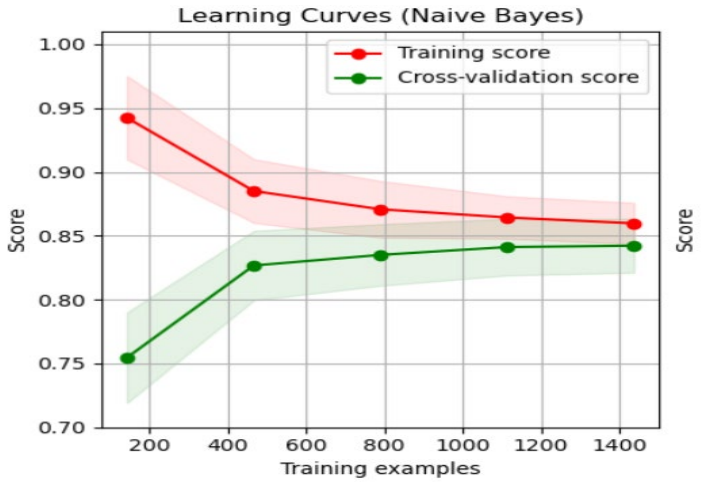


Metric Interpretation

- How to choose and explain measurement results?
- Consider two scenarios:
  - 1) A journalist J is looking for articles on a rare topic among thousands. J can view many articles but does not want to miss any important. ML model presents a number of articles. Which metric we should apply? **Recall or TPV** is relevant here, since it measures proportion of good instances currently classified.
  - 2) An investment manager M is looking for a good stock to invest. ML model delivers recommended stocks. Which metric to apply? Here is better to have fewer good stocks rather than many bad. Therefore, **Precision or PPV** is the right metric.

Methodology: Training/Validation





Model complexity grows with more training examples or more complex layout