الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

مقرر تعلم الآلة

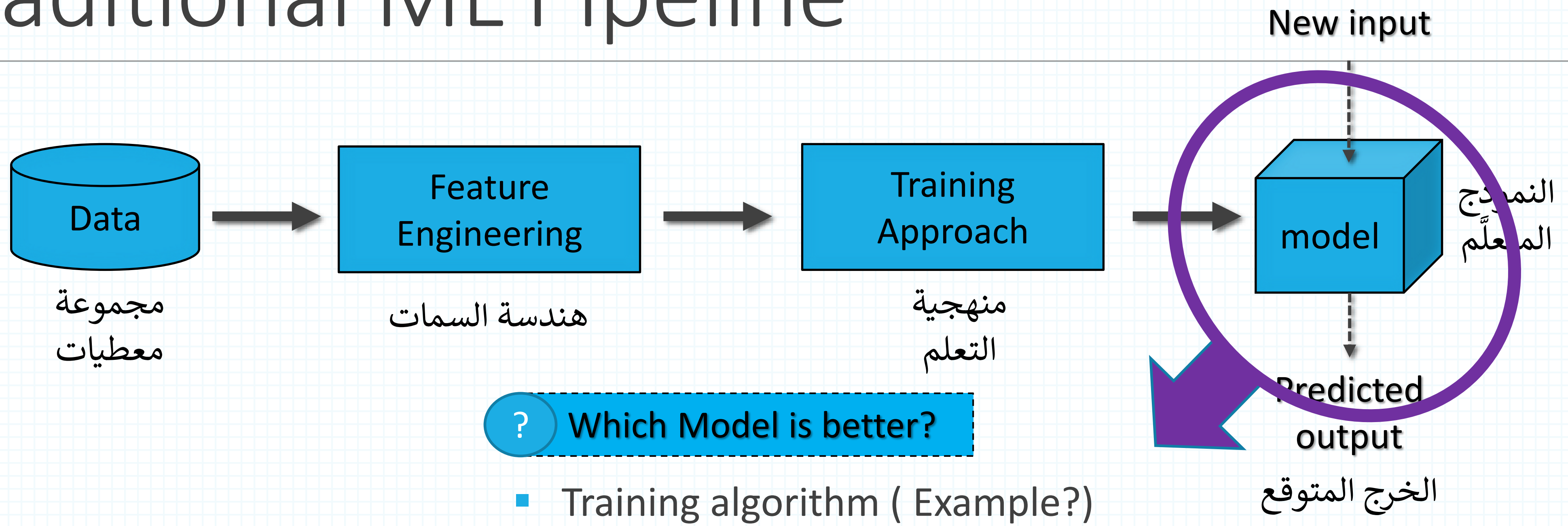
كلية الهندسة المعلوماتية

المحاضرة الخامسة

معايير تقييم جودة النماذج المدرَّبة

د. رياض سنبل

Traditional ML Pipeline



? Which Model is better?

- Training algorithm (Example?)
- Parameters (Hyperparameters)?
- Handle unseen cases

Estimation Strategy

Evaluation Metrics

Metrics

- It is extremely important to use quantitative metrics for evaluating a machine learning model.
- These metrics can be used to better evaluate and understand the model
- For classification: Accuracy/Precision/Recall/F1-score, ROC curves,…
- For regression: Mean Absolute Error (NMAE),…
- In this lecture, we will focus on classification tasks.

Accuracy

- Accuracy is a measure of how close a given set of guessing from our model are closed to their true value:

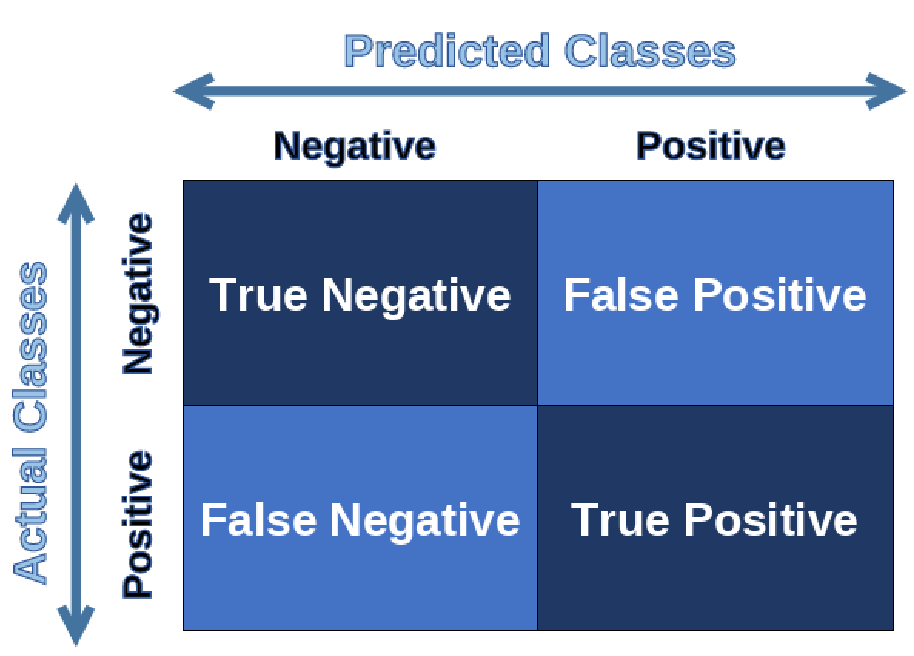
$$\text{Accuracy} = \frac{\#\ \text{Correct classifications}}{\#\ \text{All classifications}}$$

- If a classifier make 10 predictions and 8 of them are correct, the accuracy is 80%.

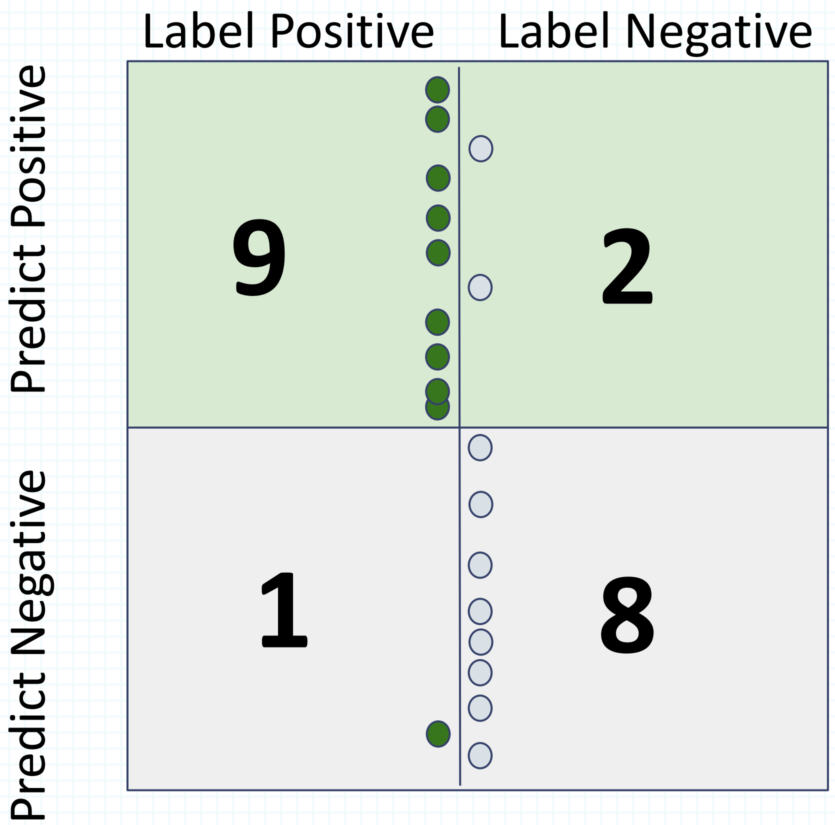
Confusion Matrix

- Confusion Matrix is a performance measurement for the machine learning classification problems where the output can be two or more classes. It is a table with combinations of predicted and actual values.

- **Example (Binary Classification)**
  - True Positive: We predicted positive and it’s true.
  - True Negative: We predicted negative and it’s true.
  - False Positive (Type 1 Error): We predicted positive and it’s false.
  - False Negative (Type 2 Error): We predicted negative and it’s false.



Example



TP	TN	FP	FN	Acc
9	8	2	1	0.85

Limitation of Accuracy

- Consider a 2-class problem
  - Number of Class 0 examples = 9990
  - Number of Class 1 examples = 10
- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
- Accuracy is misleading because model does not detect any class 1 example
- Solution ☺
  - Use precision, recall, F1 measure, etc

Precision

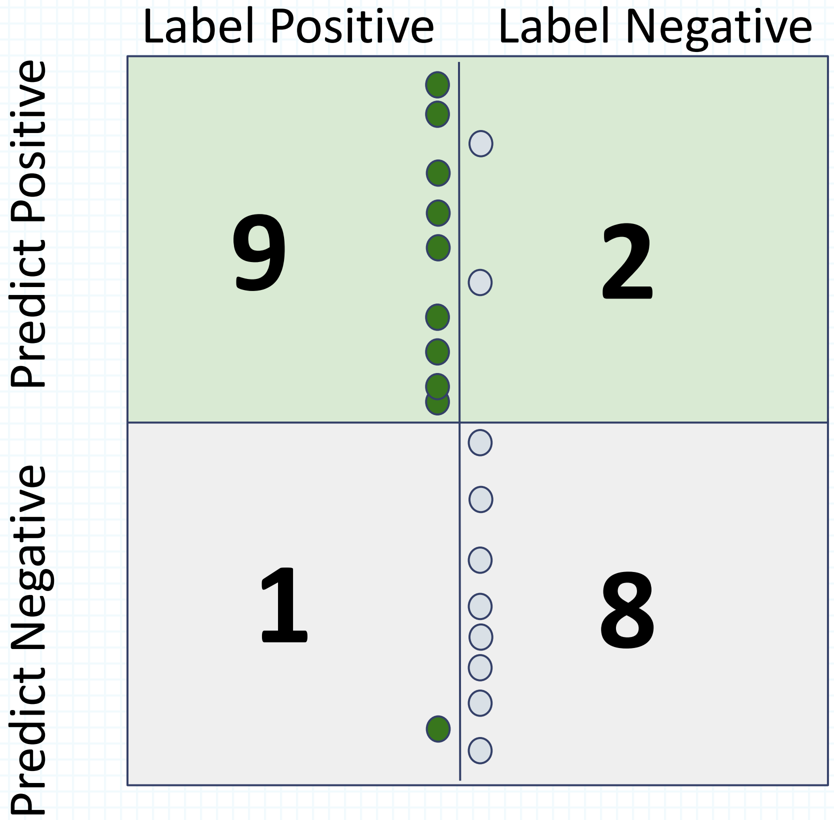
- Precision is defined as the ratio of True Positives count to total True Positive count made by the model.
- **Precision = TP/(TP+FP)**
- It explains how many of the correctly predicted cases actually turned out to be positive.
- Precision is useful in the cases where **False Positive is a higher concern** than False Negatives.
  - Ex: music or video recommendation systems

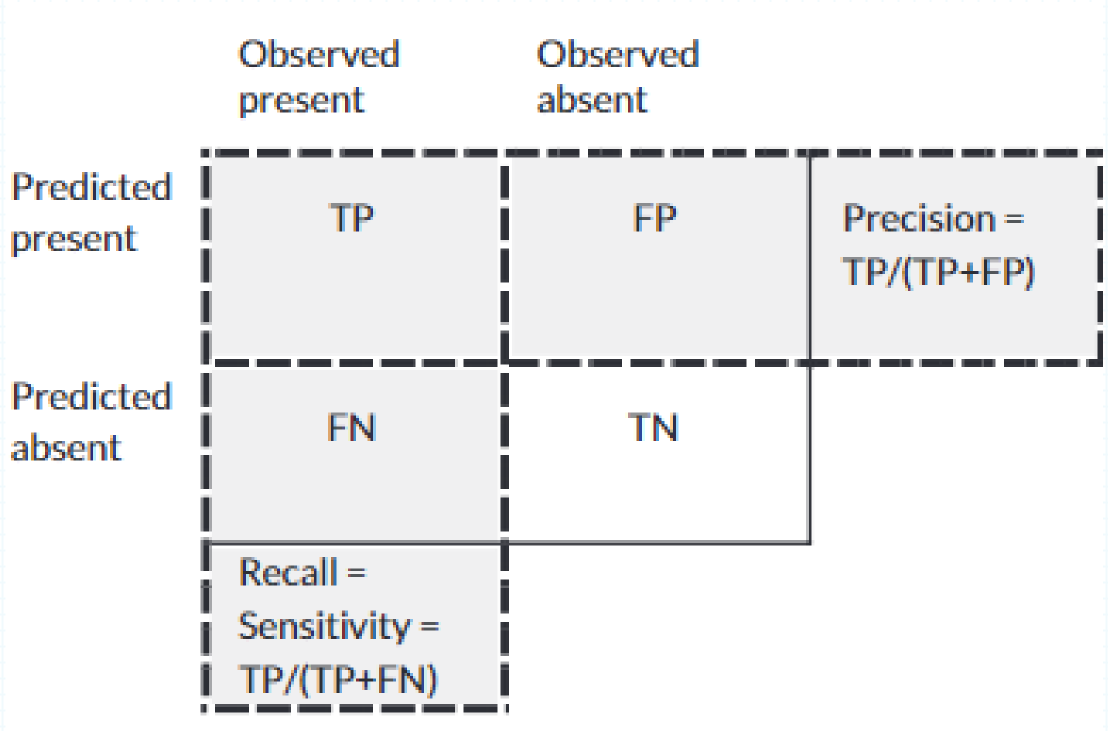
Recall

- Recall is defined as the ratio of True Positives count to the total Actual Positive count. It is also called “True Positive Rate”/ “sensitivity”.
- **Recall = TP/(TP+FN)**
- It explains how many of the actual positive cases we were able to predict correctly with our model.
- Recall is a useful metric in cases where **False Negative is of higher concern** than False Positive.
  - Ex: medical cases

Example

TP	TN	FP	FN	Acc	P	R	F1
9	8	2	1	0.85	0.81	0.90	0.857





F1-score

- It is usually better to compare models by means of **one number only**.
- The F1–score can be used to combine precision and recall

	Precision(P)	Recall (R)	Average	F_1 Score
Algorithm 1	0.5	0.4	0.45	$\boxed{0.444}$
Algorithm 2	0.7	0.1	0.4	0.175
Algorithm 3	0.02	1.0	0.51	0.0392

The best is Algorithm 1

Algorithm 3 predict always 1

Average says not correctly that Algorithm 3 is the best

$$\text{Average} = \frac{P+R}{2} \qquad F_1\text{score} = 2\frac{PR}{P+R}$$

- $P = 0 \text{ or } R = 0 \Rightarrow F_1\text{score} = 0$
- $P = 1 \text{ and } R = 1 \Rightarrow F_1\text{score} = 1$

Multiclass Classifier

- Having m classes, confusion matrix is a table of size m×m , where, element at (i, j) indicates the number of instances of class i but classified as class j.

Class	C_1	C_2	C_3	C_4	C_5	C_6
C_1	52	10	7	0	0	1
C_2	15	50	6	2	1	2
C_3	5	6	6	0	0	0
C_4	0	2	0	10	0	1
C_5	0	1	0	0	7	1
C_6	1	3	0	1	0	24

- What is the accuracy?

Multiclass Classifier

- Precision, Recall, and F1-score are calculated for each class.
  - A large confusion matrix of m*m can be considered into 2*2 matrix.

Class	C_1	C_2	C_3	C_4	C_5	C_6
C_1	52	10	7	0	0	1
C_2	15	50	6	2	1	2
C_3	5	6	6	0	0	0
C_4	0	2	0	10	0	1
C_5	0	1	0	0	7	1
C_6	1	3	0	1	0	24

C1	+	-
+	52	18
-	21	123

- Finally, we can merge overall scores (ex: using weighted average)
  - What is the overall F1-score in the previous example?