

Statistics & Statistical Machine Learning

Bài 4: Đánh Giá Hệ Thống ML



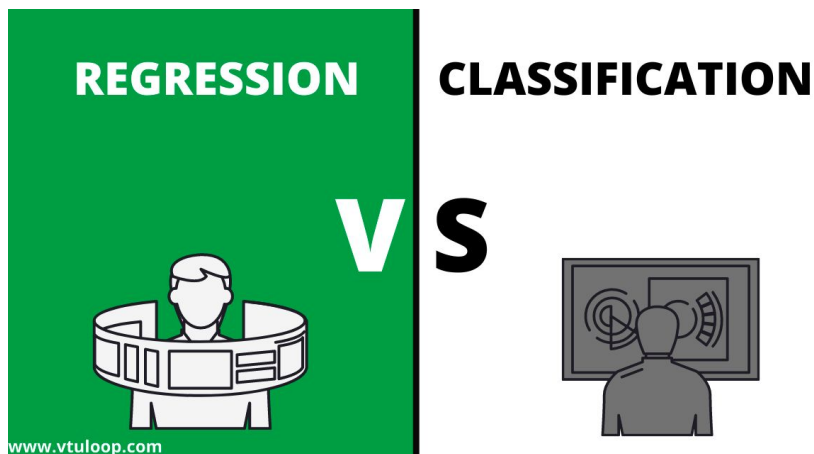
Quang-Khai Tran, Ph.D
CyberLab, 10/2022



(Ảnh: Internet)

Nội dung

1. Đánh giá các mô hình hồi quy
2. Đánh giá các mô hình phân lớp



Đánh giá mô hình trong ML

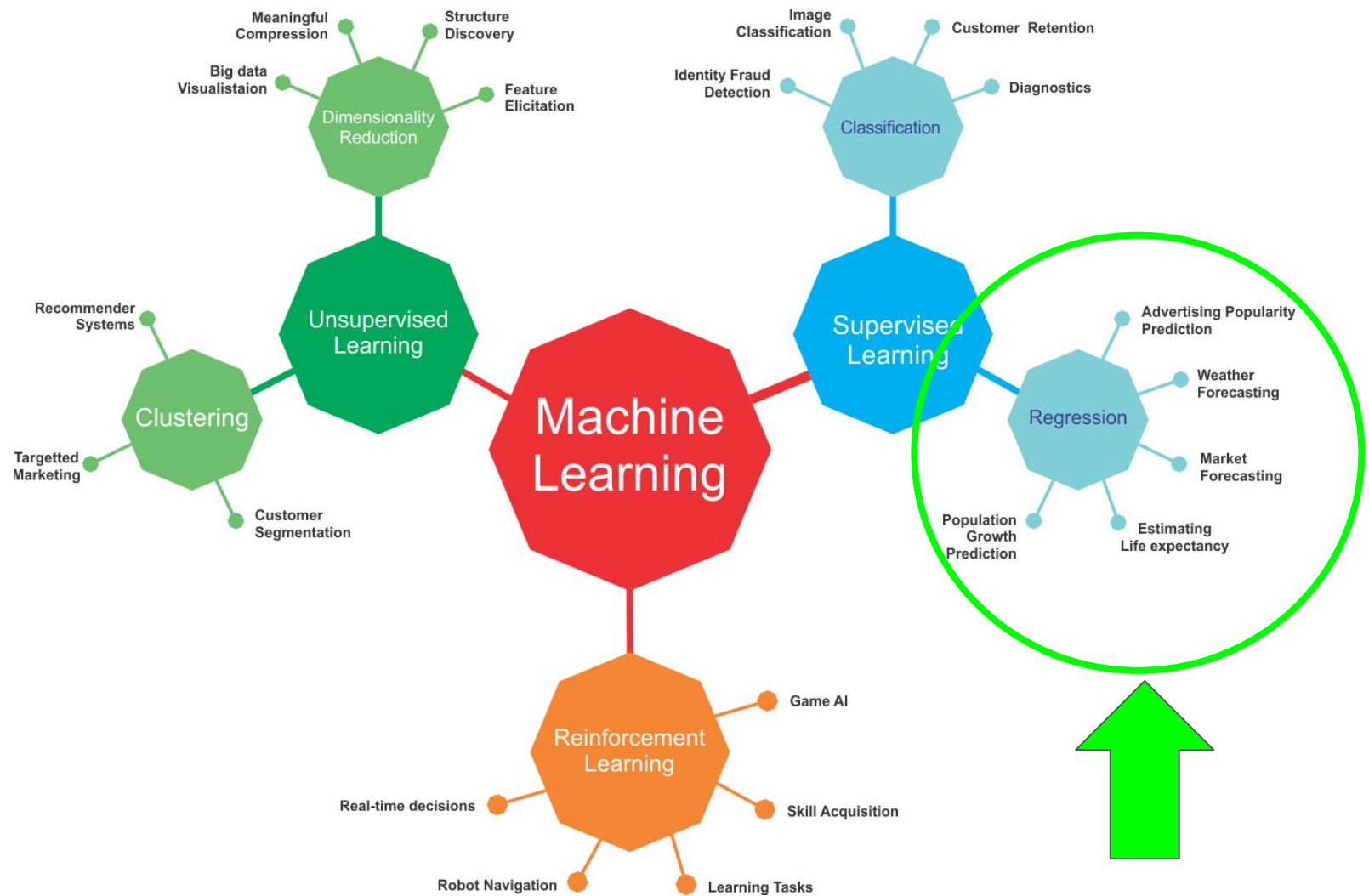
Regression	Classification	Recommender System
<ul style="list-style-type: none">• Mean Absolute Error (MAE)• Root Mean Squared Error (RMSE)• R-Squared and Adjusted R-Squared	<ul style="list-style-type: none">• Recall• Precision• F1-Score• Accuracy• Area Under the Curve (AUC)	<ul style="list-style-type: none">• Mean Reciprocal Rank• Root Mean Squared Error (RMSE)



Phần 1.

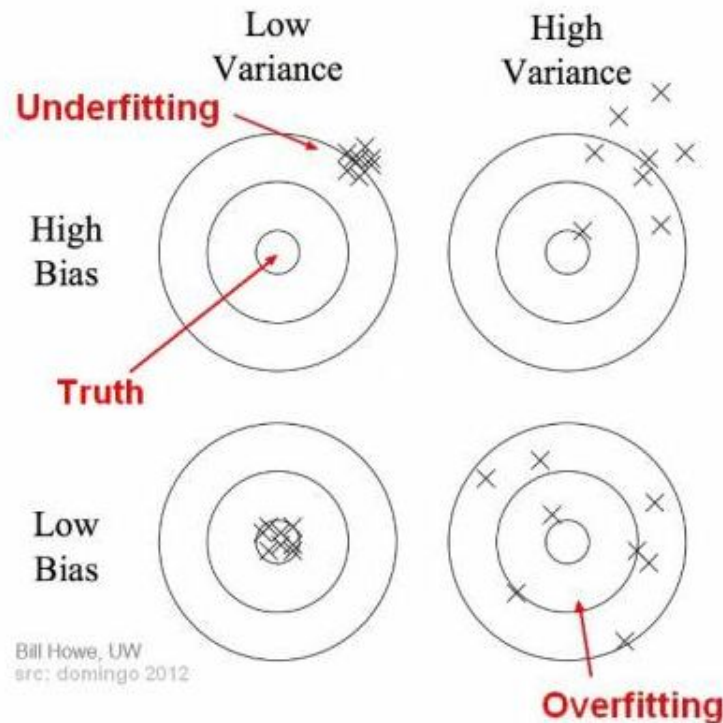
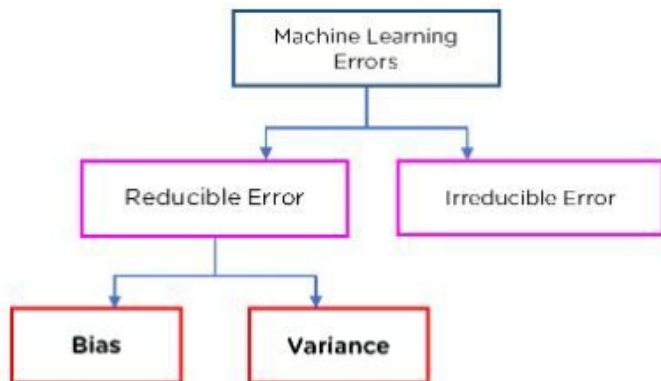
Các mô hình Regression

- 1.1. Bias, Variance và Irreducible Errors
- 1.2. Error metrics:
MAE, MSE, RMSE, MAPE
- 1.3. Đánh giá chất lượng:
R-squared, Adjusted R-squared



1.1 Bias, Variance và Irreducible Errors

- ❖ Bias: độ lệch
- ❖ Variance: phương sai
- ❖ Reducible Error: lỗi có thể giảm
- ❖ Irreducible Error (lỗi không thể giảm VD: noise trong data)

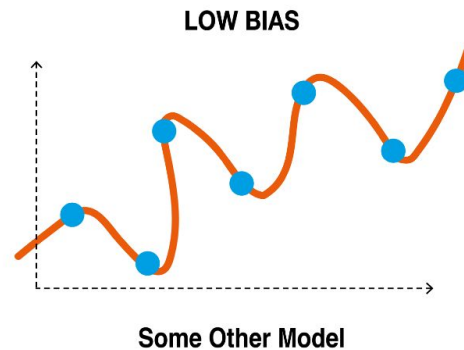
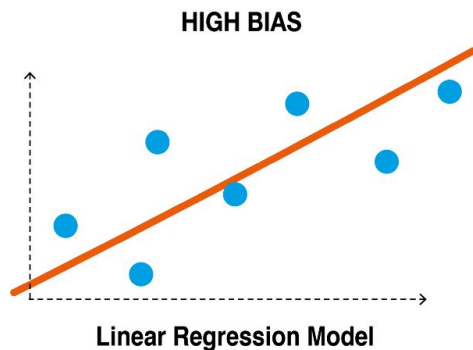


- ❖ Bias (độ lệch): là độ đo cho biết mức độ mà một phương pháp ML "không thể nắm bắt" được mối quan hệ thực sự (true relationship) của dữ liệu
 - Có thể dùng để biểu thị sự chênh lệch giữa giá trị trung bình của các dự đoán và giá trị thực tế của dữ liệu

(Công thức tham khảo)

$$\text{bias}^2 = (Y - \text{mean}(\hat{Y}))^2$$

TRAINING

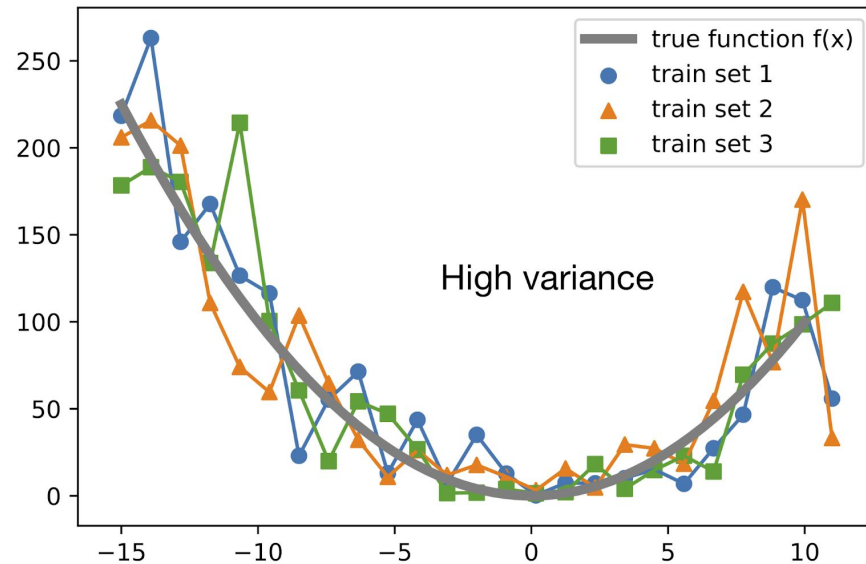


1.1 Bias, Variance và Irreducible Errors

- ❖ Variance (phương sai): là độ đo cho biết mức độ thay đổi của dự đoán khi mô hình được train với các tập training-set khác nhau
 - Có thể dùng để biểu thị sự phân tán của các giá trị mà mô hình dự đoán so với giá trị thực tế

(Công thức tham khảo)

$$\text{Variance} = \text{mean}((\hat{Y} - \text{mean}(\hat{Y}))^2)$$



1.1 Bias, Variance và Irreducible Errors

- ❖ Không có cách tính chính xác bias và variance
(Vì ta không biết được mô hình trên thực tế đã sinh ra data)
- ❖ Chỉ có cách tính ước lượng (estimate):
 - Dùng thư viện mlxtend:

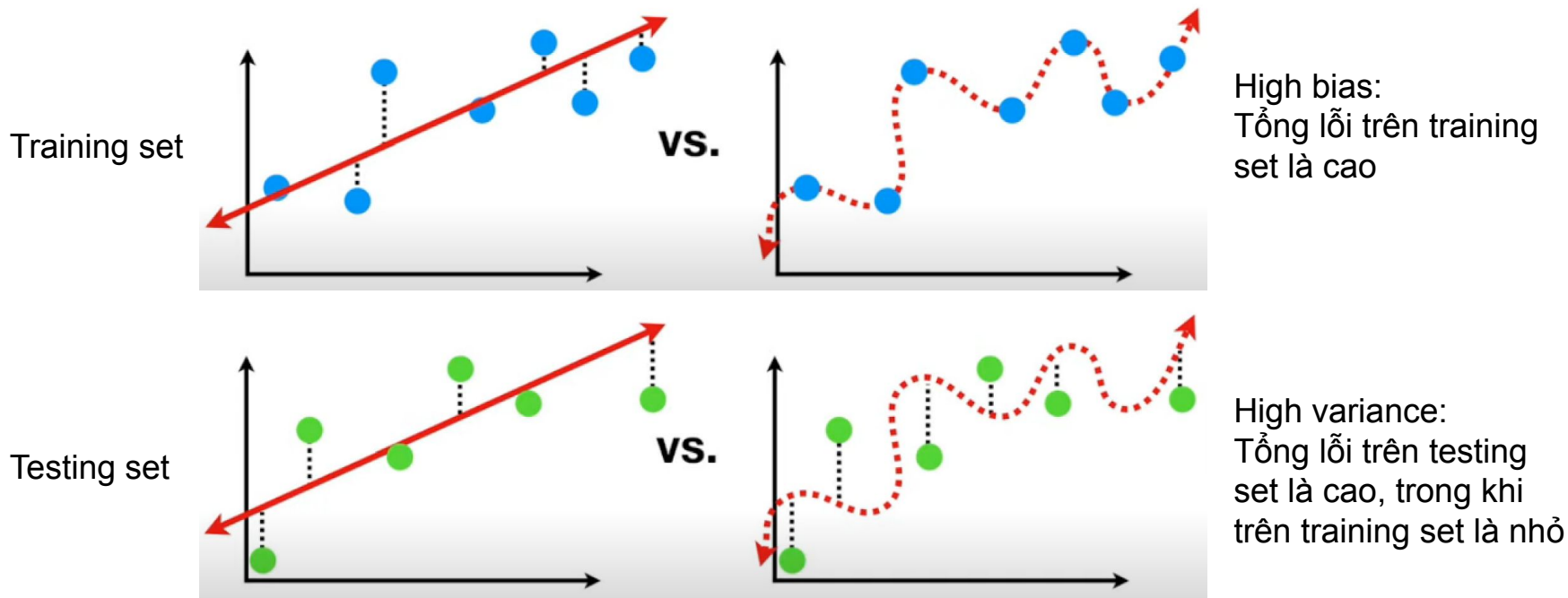
http://rasbt.github.io/mlxtend/user_guide/evaluate/bias_variance_decomp/

SNo.	Metrics	Squared Loss [MSE]
1	Single loss	$(y - \hat{y})^2$
2	Expected loss	$E[(y - \hat{y})^2]$
3	Main prediction $E[\hat{y}]$	mean (average)
4	Bias ²	$(y - E[\hat{y}])^2$
5	Variance	$E[(E[\hat{y}] - \hat{y})^2]$

Tham khảo: <https://medium.com/analytics-vidhya/calculation-of-bias-variance-in-python-8f96463c8942>

1.1 Bias, Variance và Irreducible Errors

- ❖ Làm sao để biết mô hình bị high bias, high variance?



Tham khảo: <https://www.youtube.com/watch?v=EuBBz3bl-aA>

1.2 Các loại lỗi (phổ biến)

Đánh giá mô hình regression bằng các loại lỗi (error)

Mean squared error

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2$$

Root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$$

Mean absolute error

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t|$$

Mean absolute percentage error

$$\text{MAPE} = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{e_t}{y_t} \right|$$

1.2 Đánh giá chất lượng

R-squared:

- ❖ Coefficient of determination: hệ số xác định, hệ số tương quan bội
- ❖ Là độ đo cho biết sự biến thiên trên một biến giải thích tốt hay không sự biến thiên trên một biến phụ thuộc khác
- ❖ Ví dụ: sự biến thiên nhiệt độ (mùa hè) giải thích sự biến thiên của tiền điện



R^2

$$SSE = \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (\text{Sum of Squared Errors - Residual})$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (\text{Sum of Squared Regression})$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{Sum of Squared Totals})$$

$$SST = SSE + SSR$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

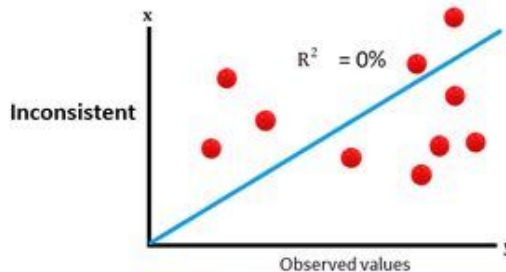
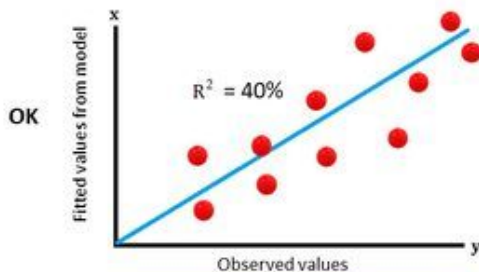
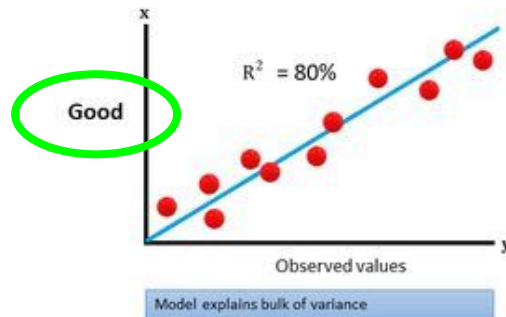
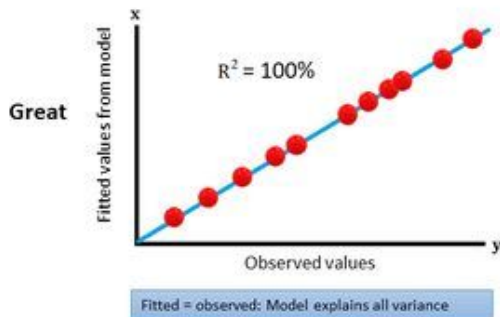
1.2 Đánh giá chất lượng

R-squared trong Machine Learning:

❖ Ví dụ: R-squared = 0.81

⇒ Các biến input giải thích được 81% sự biến thiên của output

⇒ Phần còn lại 19% được giải thích bởi các biến ngoài mô hình và sai số ngẫu nhiên



1.2 Đánh giá chất lượng

R-squared trong Machine Learning:

- ❖ Là độ đo cho biết một mô hình khớp vào data tốt thế nào (tức giải thích được sự biến thiên ở output)
- ❖ Giá trị: $0 \leq R^2 \leq 1$
 - Gần bằng 0: không giải thích được sự biến thiên (variability) của output
 - Càng tiến về 1: mô hình giải thích được sự biến thiên của output (càng tốt)

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

\hat{y} – predicted value of y

\bar{y} – mean value of y

R-squared khác R:

- ❖ Với simple linear regression: cả R và R-squared đều có thể đo sự tương quan của dữ liệu đầu vào (X, chỉ có 1 giá trị) với output của mô hình

$$R = \text{Cor}(X, Y)$$

$$R^2 = \text{Cor}(X, Y)^2$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- ❖ Với multiple linear regression: không thể đo được

⇒ Do đó cần dùng R-squared để đo tỷ lệ giữa \hat{Y} và Y

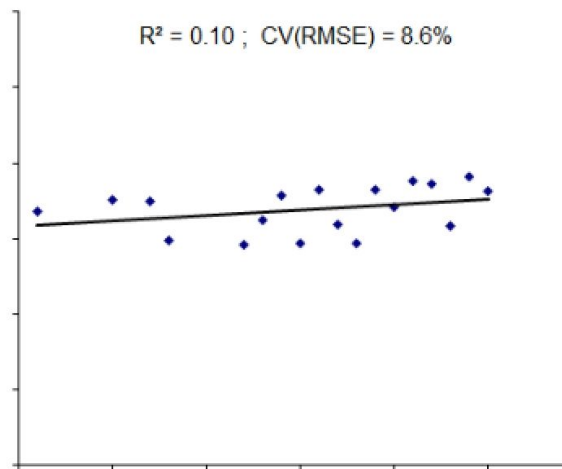
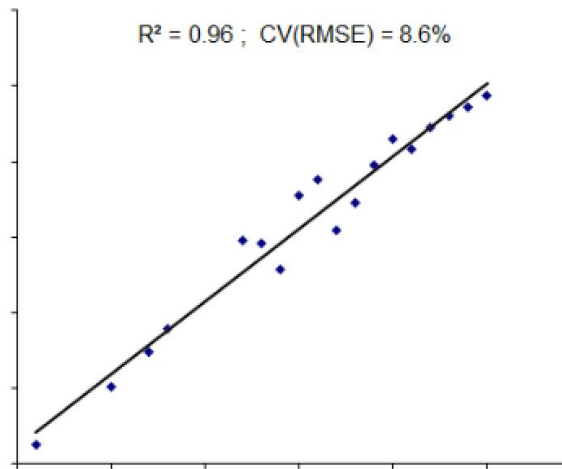
⇒ Thay x bằng \hat{y} trong công thức

(Lưu ý: vẫn có thể tính $\text{Cor}(Y, \hat{Y})$, nhưng từ “tương quan” có ý nghĩa khác)

1.2 Đánh giá chất lượng

R và R-squared:

- ❖ Có thể giúp phát hiện trường hợp một dữ liệu đầu vào không đóng góp nhiều cho dự đoán, dù lỗi của mô hình là tốt



Vấn đề với R-squared:

- ❖ R-squared là hàm không giảm theo số biến độc lập được đưa vào mô hình
- ❖ Càng đưa thêm biến độc lập vào mô hình thì R-squared càng tăng
- ❖ Không phải phương trình càng có nhiều biến thì càng tốt hơn

⇒ **Adjusted R-squared**: R bình phương hiệu chỉnh

- ❖ Phản ánh sát hơn mức độ phù hợp của mô hình hồi quy tuyến tính đa biến
- ❖ Không nhất thiết tăng lên khi đưa thêm các biến độc lập vào mô hình

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

Where

R^2 Sample R-Squared

N Total Sample Size

p Number of independent
variable



Phần 2.

Các mô hình Classification

2.1. Accuracy

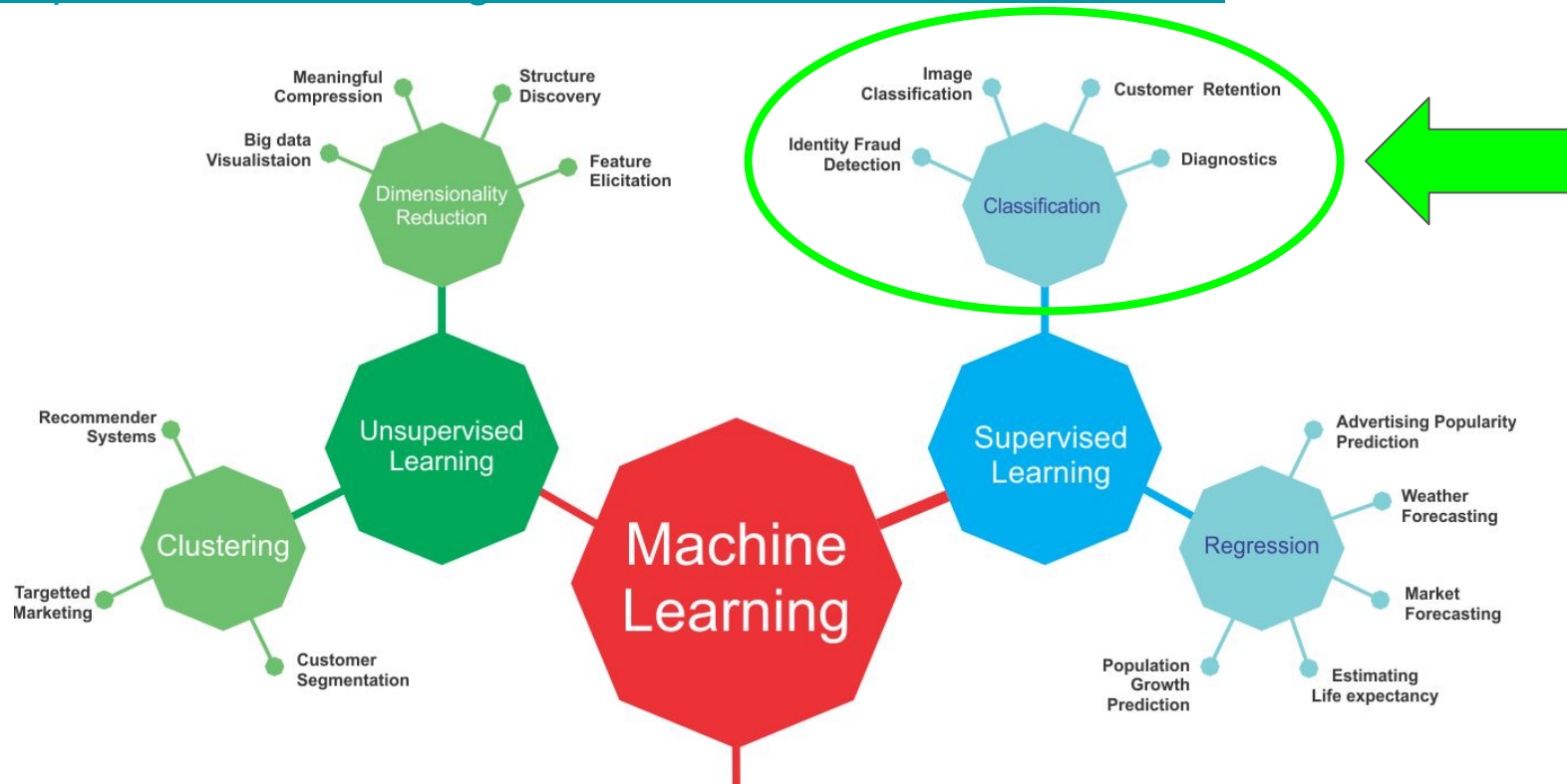
2.2. Confusion matrix

2.3. Precision, Recall, Specificity,
F1-score

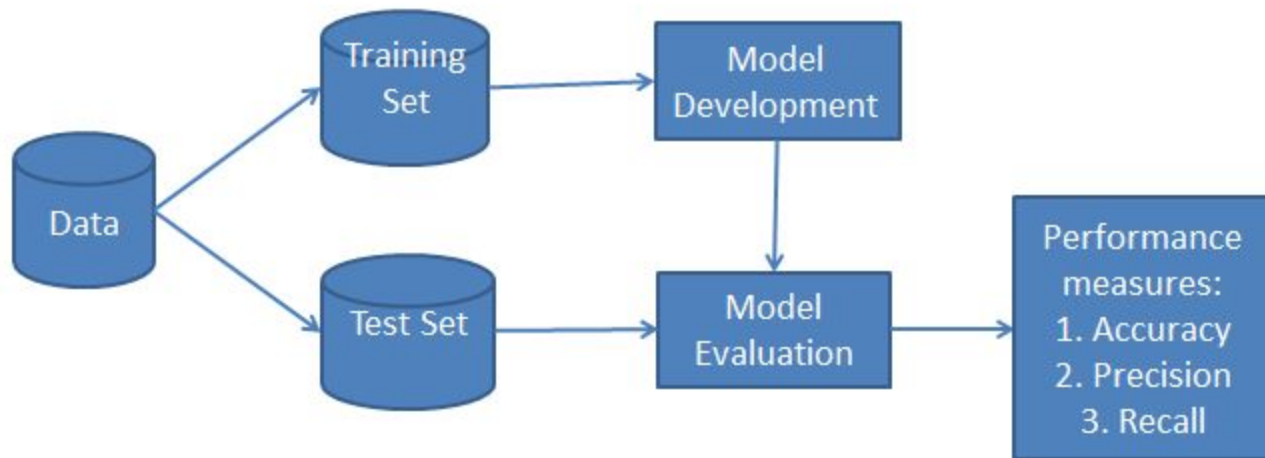
2.4. ROC và AUC

Tham khảo: "**Bài 33: Các phương pháp đánh giá một hệ thống phân lớp**"

Link: <https://machinelearningcoban.com/2017/08/31/evaluation/>



Các bước thực hiện



Tham khảo: "**Methods to Check the Performance of Classification Models**"

<https://www.enjoyalgorithms.com/blog/evaluation-metrics-classification-models>

Confusion Matrix

		Predicted Class	
		Positive	Negative
Actual Class	Positive	True Positive	False Negative
	Negative	False Positive	True Negative



2.1 Accuracy

Accuracy (sự chính xác, độ chính xác)

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of Records}}$$



2.2 Confusion Matrix

Đánh giá mô hình phân loại (2 classes)

		Predicted	
		Positive (+)	Negative (-)
Actual	Positive (+)	True Positive (TP)	False Negative (FN)
	Negative (-)	False Positive (FP)	True Negative (TN)

True Positives (TP): đoán = 1, thật = 1

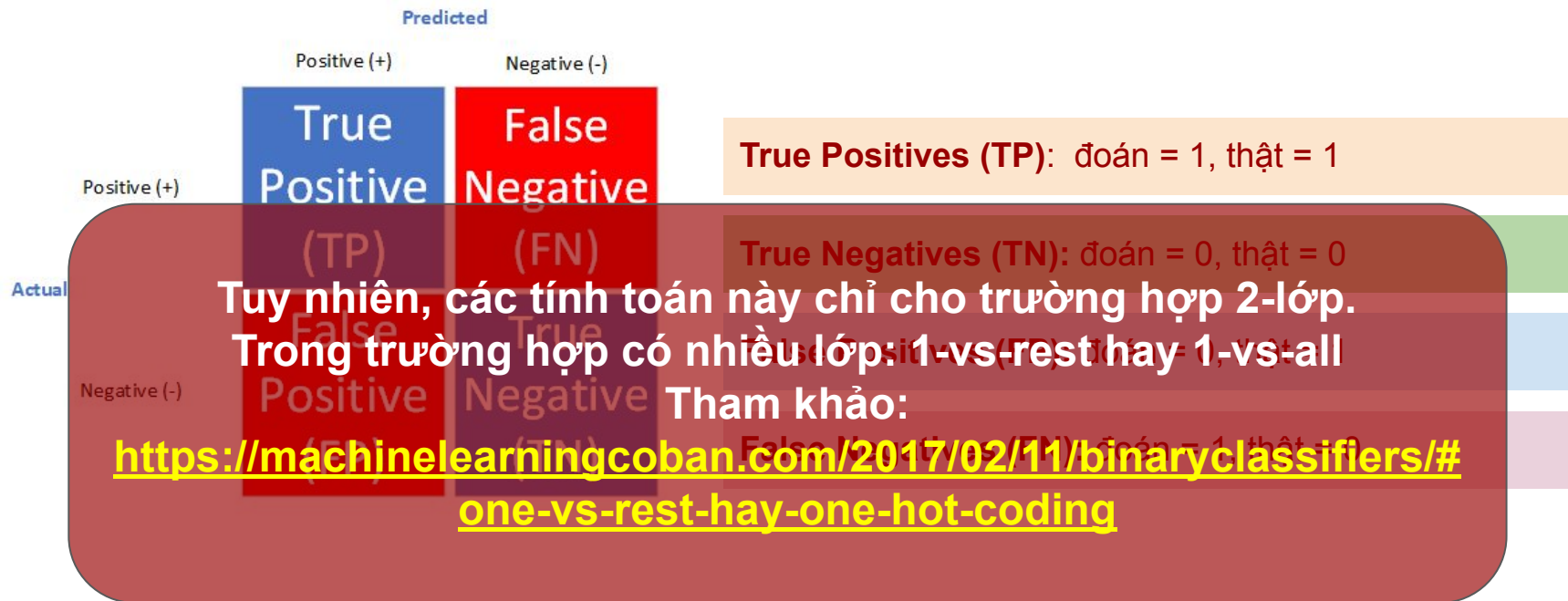
False Positives (FP): đoán = 1, thật = 0

True Negatives (TN): đoán = 0, thật = 0

False Negatives (FN): đoán = 0, thật = 1

2.2 Confusion Matrix

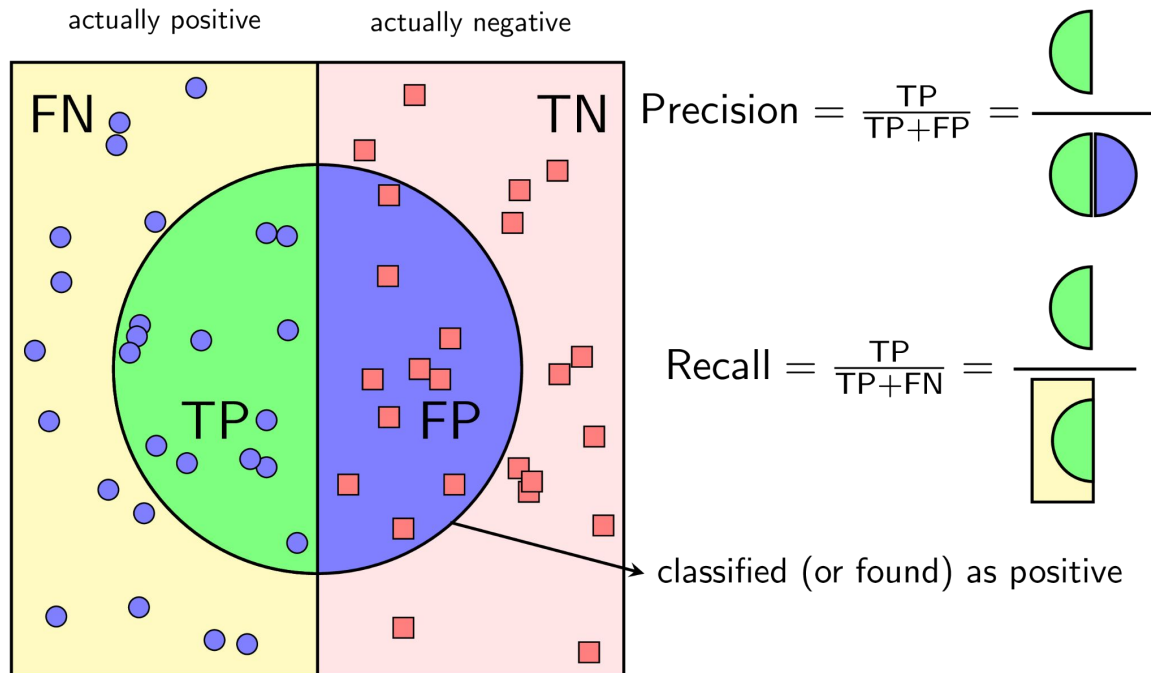
Đánh giá mô hình phân loại (nhiều classes)



2.2 Confusion Matrix

Đánh giá mô hình phân loại (2 classes)

⇒ 2 khái niệm quan trọng: Precision vs. Recall



2.2 Confusion Matrix

Đánh giá mô hình phân loại (2 classes)

		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP + FN)}$ Recall
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	Negative Predictive Value $\frac{TN}{(TN + FN)}$	Accuracy $\frac{TP + TN}{(TP + TN + FP + FN)}$

2.3 Precision, Recall, Specificity, F1-score

Đánh giá mô hình phân loại (2 classes)

Metric	Formula	Interpretation
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
Precision	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
Recall Sensitivity	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
Specificity	$\frac{TN}{TN + FP}$	Coverage of actual negative sample
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes

2.3 Precision, Recall, Specificity, F1-score

Đánh giá mô hình phân loại (2 classes): ví dụ

n=165		Predicted: NO	Predicted: YES	
Actual: NO		TN = 50	FP = 10	60
Actual: YES		FN = 5	TP = 100	105
		55	110	

Recall = 100/105

Precision = 100/110

2.3 Precision, Recall, Specificity, F1-score

Đánh giá mô hình phân loại (2 classes): ví dụ

		Predicted Class	
		Spam	Non-Spam
Actual Class	Spam	TP=45	FN=20
	Non-Spam	FP=5	TN=30

Recall = ???

Precision = ???

2.3 Precision, Recall, Specificity, F1-score


Đánh giá mô hình phân loại (nhiều classes): ví dụ




		<i>labels</i>			
		urgent	normal	spam	
<i>system output</i>	urgent	8	10	1	$\text{precision}_u = \frac{8}{8+10+1}$
	normal	5	60	50	$\text{precision}_n = \frac{60}{5+60+50}$
	spam	3	30	200	$\text{precision}_s = \frac{200}{3+30+200}$
		$\text{recall}_u = \frac{8}{8+5+3}$	$\text{recall}_n = \frac{60}{10+60+30}$	$\text{recall}_s = \frac{200}{1+50+200}$	

2.3 Precision, Recall, Specificity, F1-score

F1-score (F-measure):

- ❖ Trường hợp có nhiều classes
- ❖ Số lượng phần tử dữ liệu trong các classes không đều nhau









	precision	recall	f1-score	support
Aeroplane 	0.67	0.67	0.67	3
Boat 	0.25	1.00	0.40	1
Car 	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

2.3 Precision, Recall, Specificity, F1-score

F1-score (F-measure): Ví dụ

(tham khảo: <https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f>)

No	Actual	Predicted	Match
1	Airplane	Airplane	✓
2	Car	Boat	✗
3	Car	Car	✓
4	Car	Car	✓
5	Car	Boat	✗
6	Airplane	Boat	✗
7	Boat	Boat	✓
8	Car	Airplane	✗
9	Airplane	Airplane	✓
10	Car	Car	✓

		Predicted		
		 Airplane	 Boat	 Car
Actual	 Airplane	2	1	0
	 Boat	0	1	0
	 Car	1	2	3

2.3 Precision, Recall, Specificity, F1-score

F1-score (F-measure):

- ❖ Có thể đánh giá tốt hơn nếu kết hợp cả Recall và Precision để tóm tắt hiệu suất của một mô hình

$$\text{F1 Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{F1 Score} = \frac{\text{TP}}{\text{TP} + \frac{1}{2} (\text{FP} + \text{FN})}$$

- ❖ Có thể chia ra 3 loại:
 - Micro F1 (accuracy)
 - Macro F1 (tính trung bình không quan tâm tỷ lệ)
 - Weighted Average F1 (tính trung bình với tỷ lệ)

2.3 Precision, Recall, Specificity, F1-score

F1-score (F-measure): Ví dụ (tt.)

(tham khảo: <https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f>)

- ❖ Micro F1 (accuracy): dataset là cân bằng
- ❖ Macro F1: nếu dataset là không cân bằng và các classes quan trọng như nhau
- ❖ Weighted Avg F1: nếu dataset cũng không cân bằng, nhưng muốn đề cao các classes có nhiều phần tử hơn

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

Per-Class F1 scores

Average F1 scores




2.3 Precision, Recall, Specificity, F1-score

F1-score (F-measure): các mức kết luận

(tham khảo: <https://stephenallwright.com/good-f1-score/>)

F1 score	Kết luận
> 0.9	Rất tốt
0.8 - 0.9	Tốt
0.5 - 0.8	OK
< 0.5	Không tốt



	precision	recall	f1-score	support
Aeroplane 	0.67	0.67	0.67	3
Boat 	0.25	1.00	0.40	1
Car 	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

2.3 Precision, Recall, Specificity, F1-score

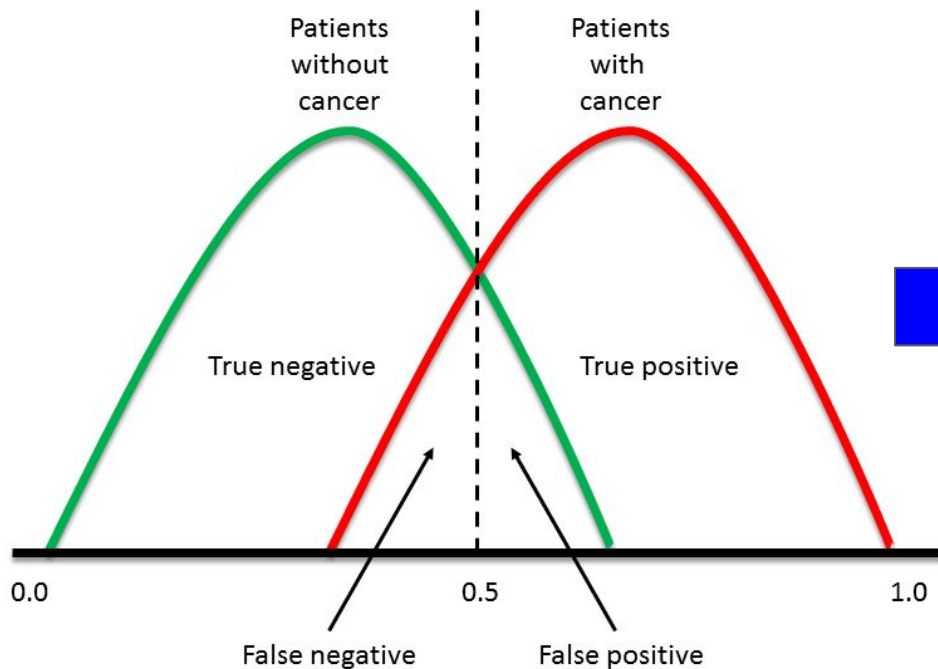
Một số thuật ngữ liên quan:

- ❖ TPR (True Positive Rate): Recall (đoán đúng việc cần đoán)
- ❖ FPR (False Positive Rate): False Alarm Rate (FAR, tỉ lệ báo động nhầm)
- ❖ FNR (False Negative Rate): Miss Detection Rate (MDR, tỉ lệ bỏ sót)
- ❖ TNR (True Negative Rate): Specificity

	Predicted as Positive	Predicted as Negative
Actual: Positive	$TPR = TP / (TP + FN)$	$FNR = FN / (TP + FN)$
Actual: Negative	$FPR = FP / (FP + TN)$	$TNR = TN / (FP + TN)$

2.4 ROC và AUC

- ❖ Ví dụ: khi ngưỡng để xác định bệnh ung thư thay đổi
⇒ Trường hợp này cần ưu tiên cảnh báo ung thư hơn là cảnh báo sai



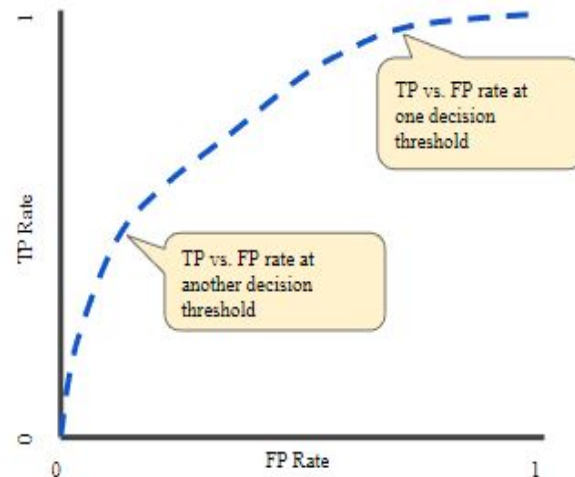
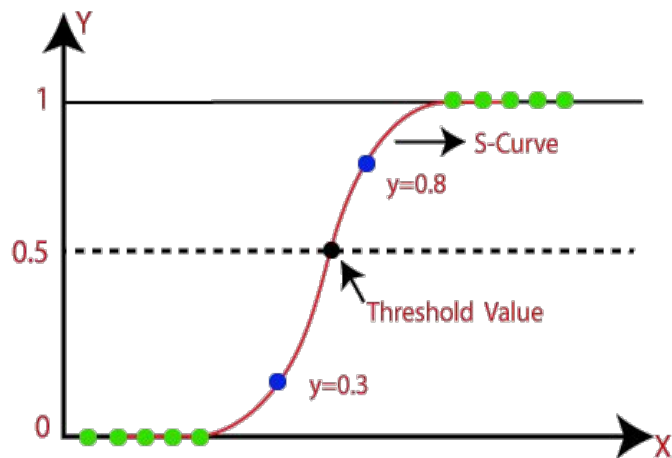
	Actual Cancer = Yes	Actual Cancer = No
Predicted Cancer = Yes	True Positive 57	False Positive 14
Predicted Cancer = No	False Negative 23	True Negative 171

2.4 ROC và AUC

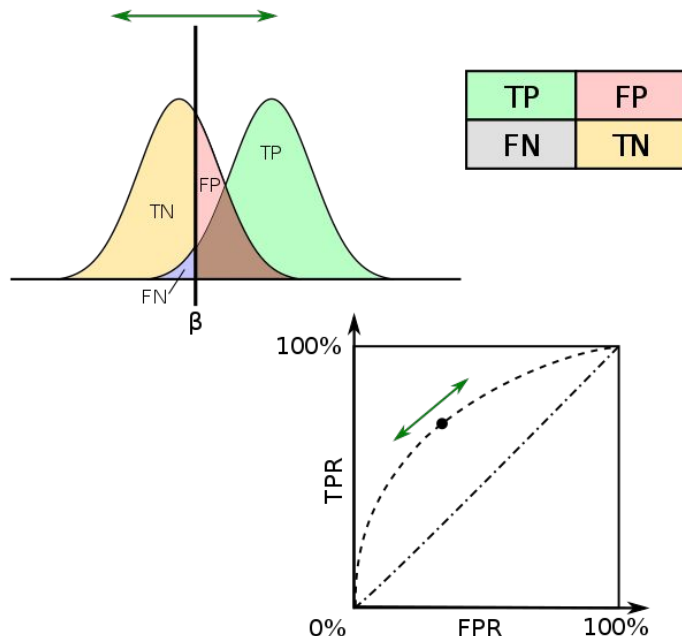
- ❖ Trong nhiều lĩnh vực thực tế, người ta quan tâm nhiều đến TPR và FPR

$$\text{TPR} = \text{Recall} = \text{Sensitivity} = \frac{TP}{TP + FN}$$

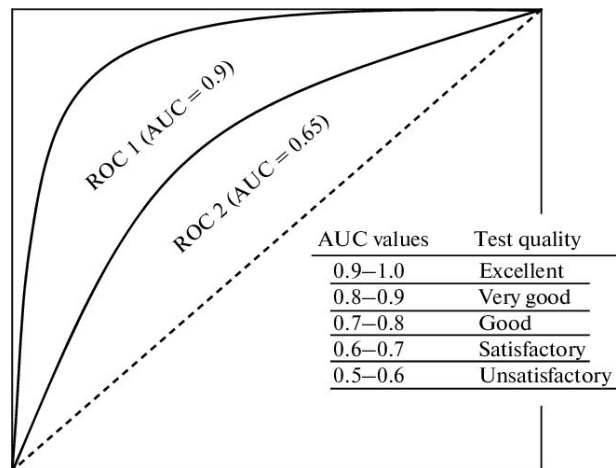
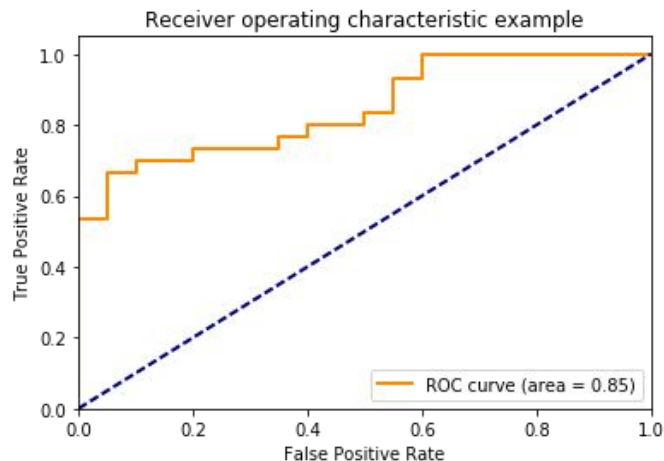
$$\text{False Positive Rate (FPR)} = 1 - \text{specificity} = 1 - \frac{TN}{TN + FP} = \frac{FP}{FP + TN}$$



- ❖ Receiver Operating Characteristic: đường cong ROC (curve)
 - Khi ngưỡng để xác định class thay đổi
 - Cho thấy sự thay đổi của TPR so với FPR



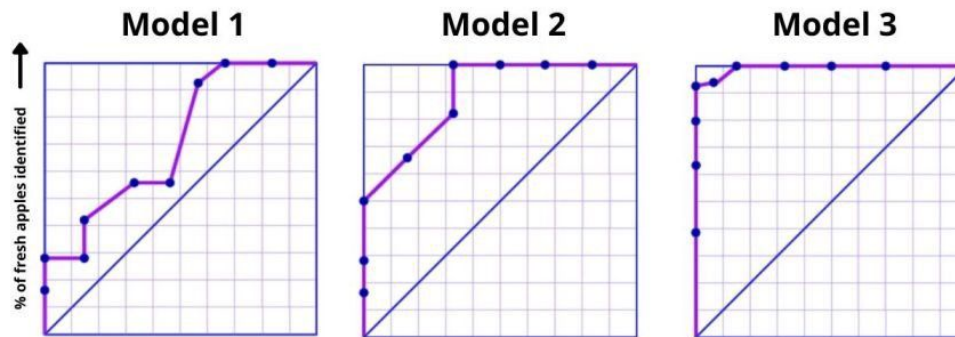
- ❖ Receiver Operating Characteristic: đường cong ROC (curve)
 - Khi ngưỡng để xác định class thay đổi
 - Cho thấy sự thay đổi của TPR so với FPR
- ❖ Area Under the Curve
Diện tích nằm dưới đường màu cam



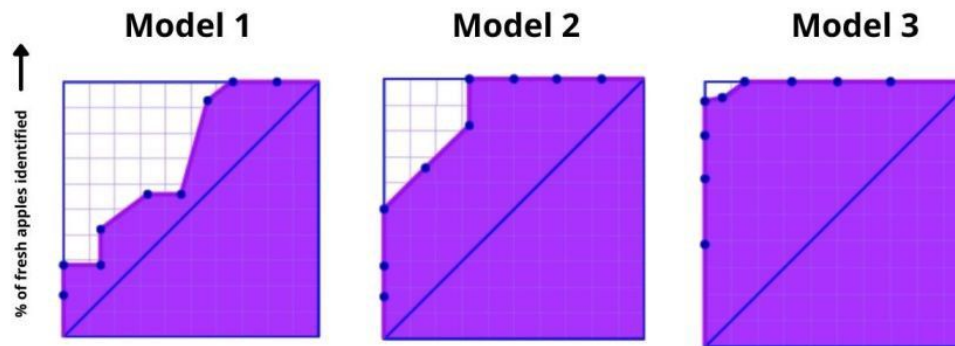
2.4 ROC và AUC

❖ Ví dụ:

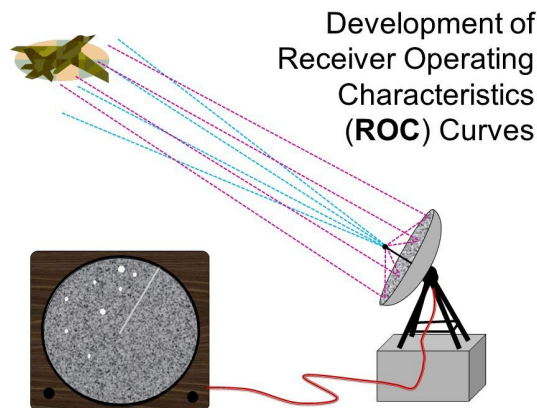
ROC Curves For Three Models



AUC For Three Models

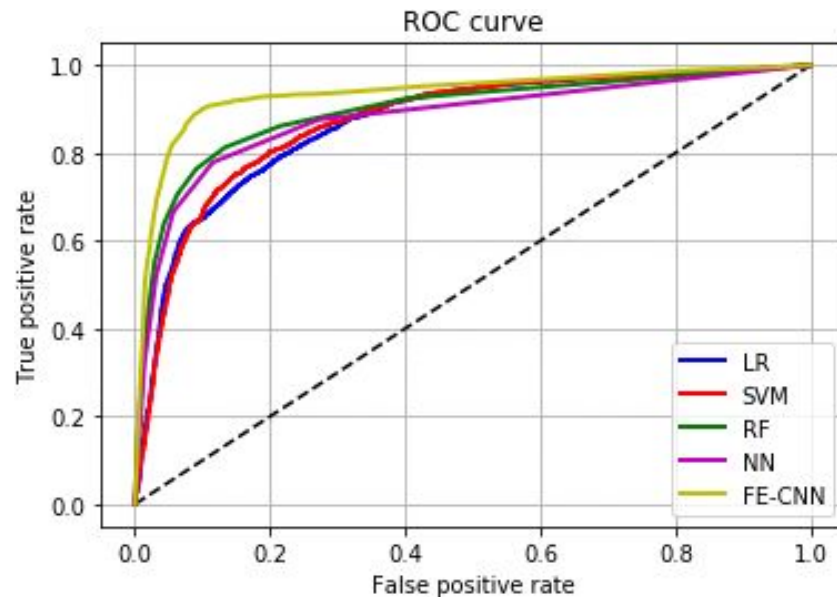
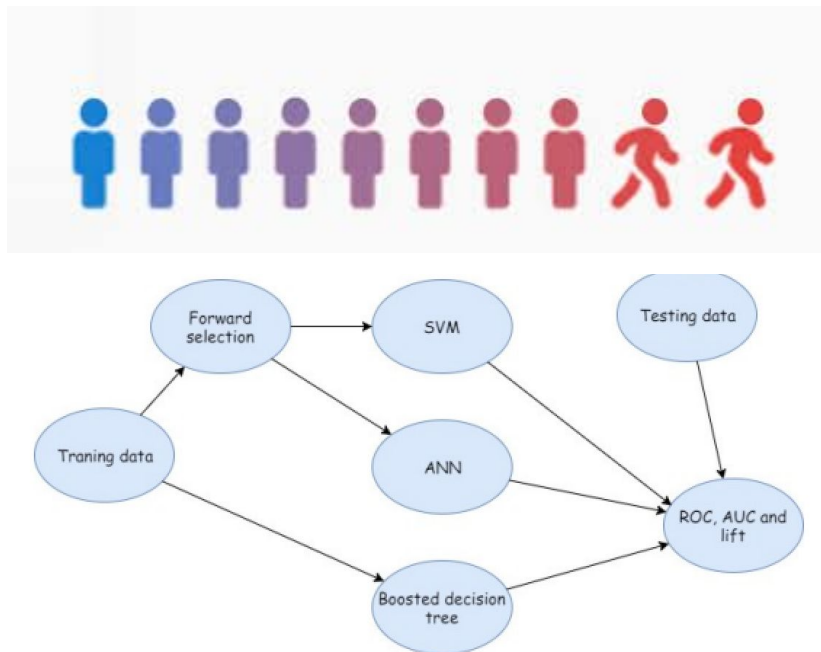


- ❖ Receiver Operating Characteristic: đường cong ROC (curve)
 - Trong lý thuyết thu phát tín hiệu: đường cong đặc trưng hoạt động của bộ thu nhận – để xác định là có tín hiệu hay chỉ là do nhiễu
 - Ứng dụng đầu tiên là cho việc nghiên cứu các hệ thống nhận diện trong việc phát hiện các tín hiệu radio khi có sự hiện diện của nhiễu vào thập niên 1940
 - Một trục là độ nhạy
 - Một trục là giá trị đánh giá một hệ thống phân loại khi ngưỡng thay đổi



2.4 ROC và AUC

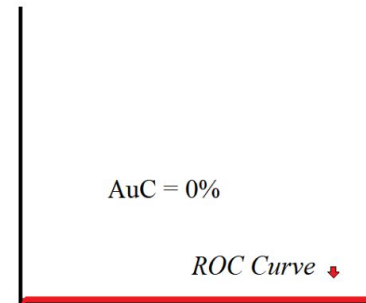
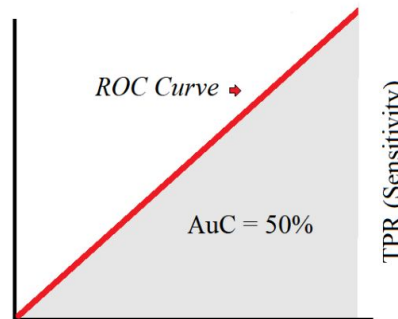
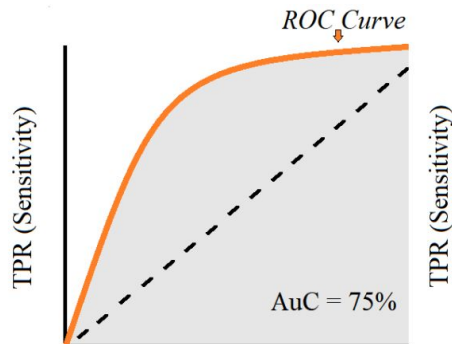
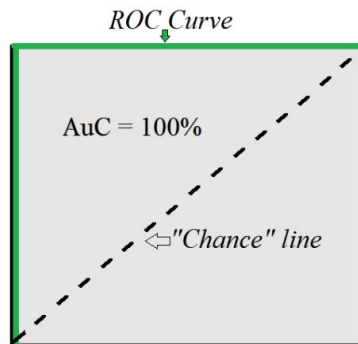
Ví dụ: dùng ROC để tìm mô hình tốt nhất trong customer churn prediction



2.4 ROC và AUC

❖ Mức của AUC

AUC	Ý nghĩa
>0.90	Rất tốt (Excellent)
0.80 - 0.90	Tốt (Good)
0.70 - 0.80	Trung bình (Fair)
0.60 - 0.70	Không tốt (Poor)
0.50 - 0.60	Vô dụng (Fail)



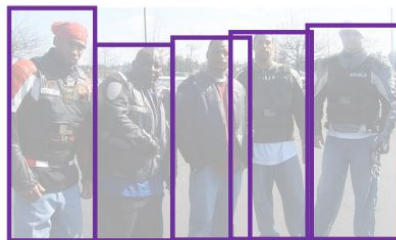
Sau này có thể dùng vào việc gì?

Ví dụ: mean Average Precision (nhận dạng đối tượng, phân tách đối tượng)

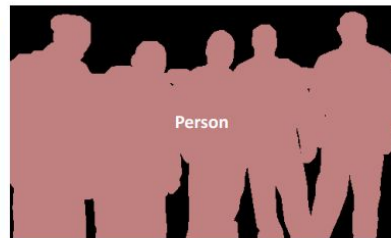
$$mAP = \frac{1}{n} \sum_{k=1}^{k=n} AP_k$$

AP_k = the AP of class k

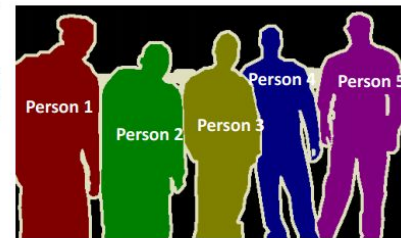
n = the number of classes



Object Detection



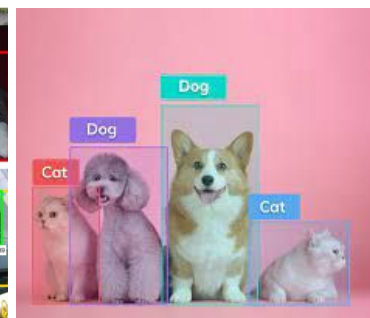
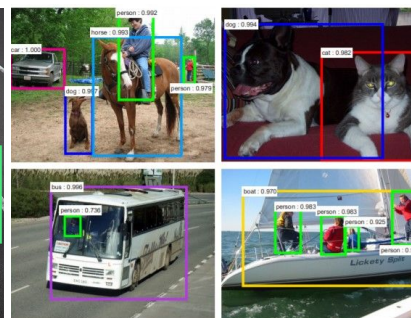
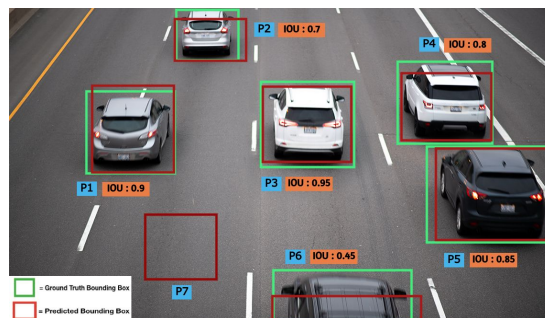
Semantic Segmentation



Instance Segmentation

$$AP = \sum_{k=1}^N p(k) \Delta r(k)$$

(Precision-Recall Curve)



THANK YOU!

