



Bài giảng môn học:

**Kỹ nghệ tri thức và học máy (4080540)**

# **CHƯƠNG 3: HỌC CÓ GIÁM SÁT - 04 (Supervised Learning)**

**Giảng viên: Đặng Văn Nam**

**Email: [dangvannam@humg.edu.vn](mailto:dangvannam@humg.edu.vn)**

# Nội dung chương 3

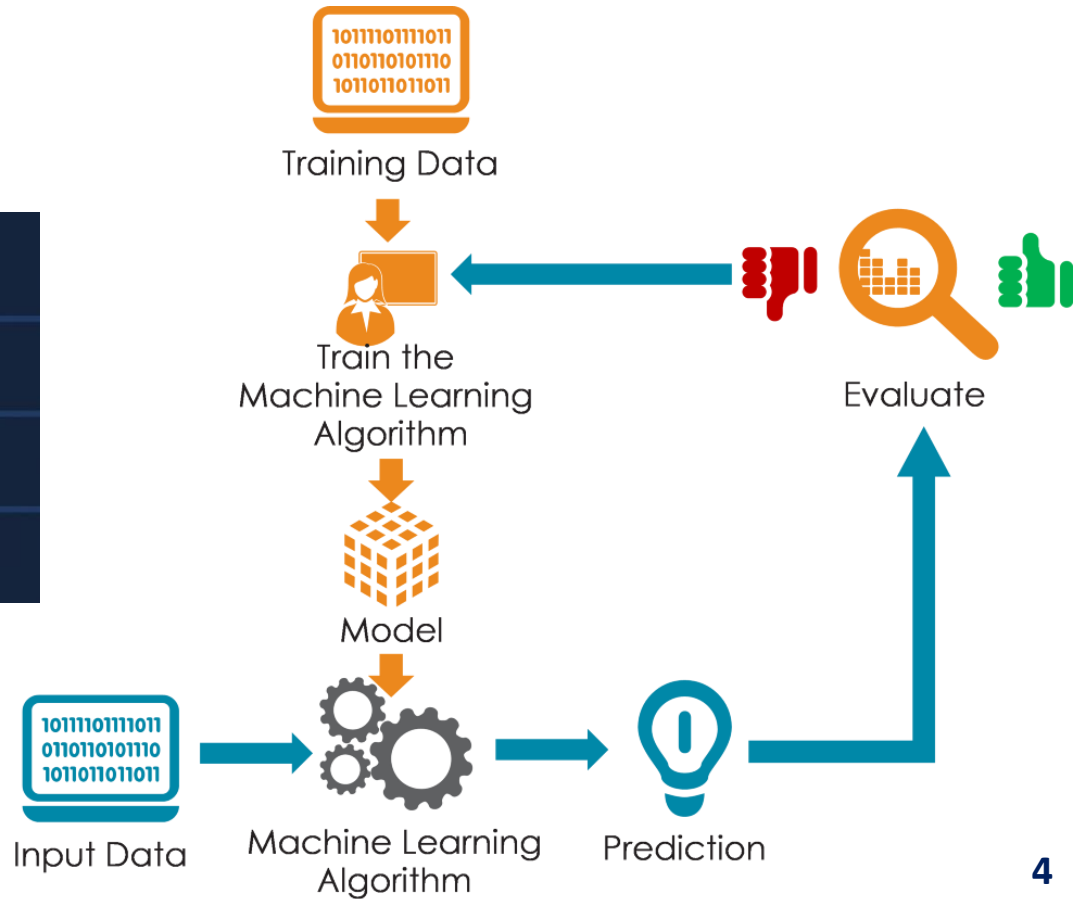
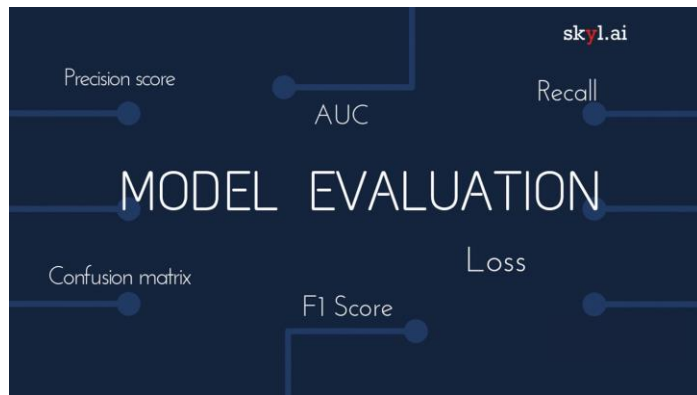
---

1. Các bước xây dựng một mô hình học máy
2. Datasets
3. Học có giám sát (Supervised Learning)
4. Phân loại học có giám sát (Classification – Regression)
5. Thuật toán phân loại (KNN, Decision Tree)
6. Thuật toán hồi quy (Linear, Polynomial regression)
7. Đánh giá độ chính xác của mô hình phân lớp, hồi quy

## **7. Đánh giá độ chính xác của một mô hình học máy**

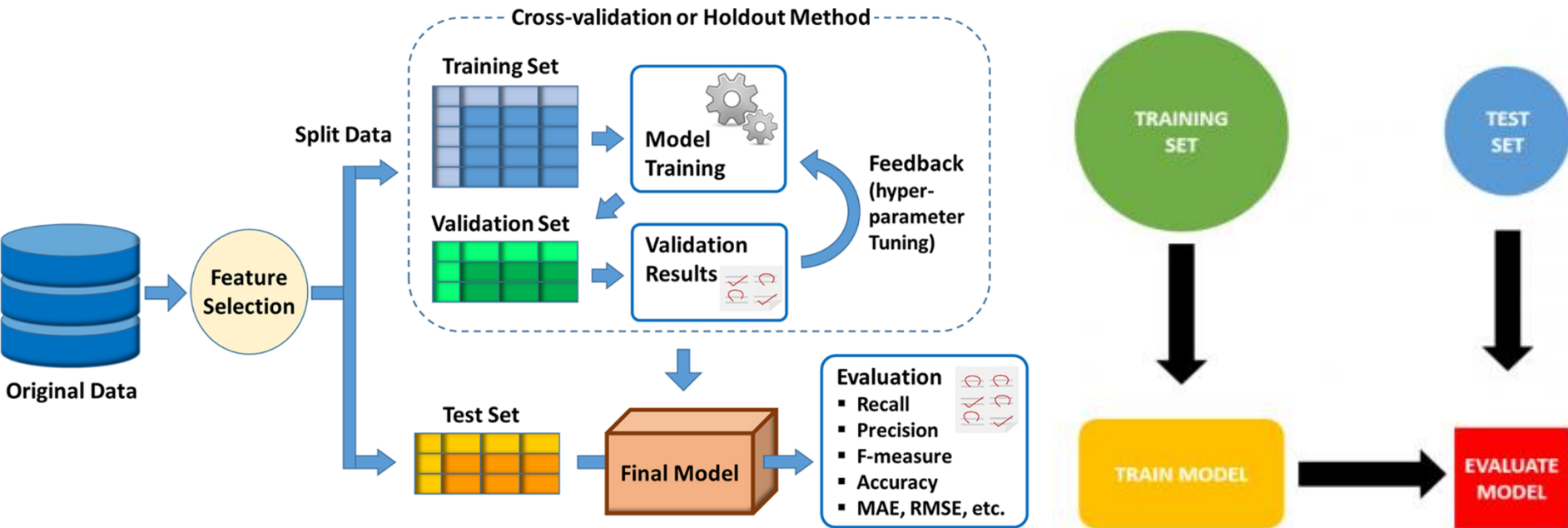
# Đánh giá mô hình học máy

Khi xây dựng một mô hình học máy, chúng ta cần một phép đánh giá để xem mô hình sử dụng có hiệu quả không và để so sánh độ chính xác của các mô hình học máy khác nhau cho bài toán cần giải quyết.



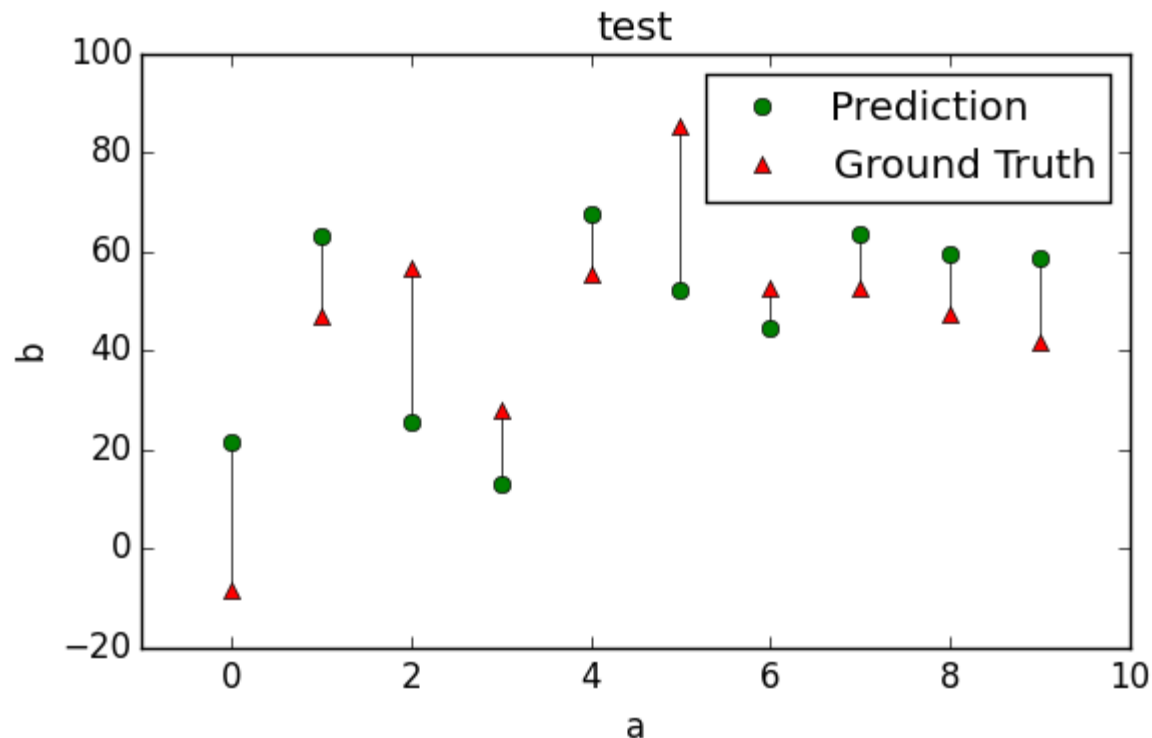
# Đánh giá mô hình học máy

Hiệu năng của một mô hình thường được đánh giá dựa trên tập dữ liệu kiểm thử (Test Data)



# Đánh giá mô hình học máy

- **y\_predict:** Kết quả dự đoán của mô hình học máy với tập dữ liệu kiểm thử (Test Data)
- **y\_groundtruth:** Nhãn đầu ra thật (target) của tập dữ liệu kiểm thử (Test Data)



# Đánh giá mô hình học máy



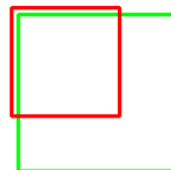
## Phân lớp (Classification)

y_predict	y_groundtruth
0	0
0	0
1	0
0	0
0	1
1	1
0	0
0	0
1	0
1	1
1	0
0	1
0	0

## Hồi quy (Regression)

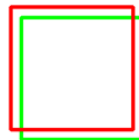
y_predict	y_groundtruth
22 890	23 432
19 120	18 850
9 590	10 500
20 231	22 567
7 498	5 235
13 675	11 563
22 453	25 005
24 645	19 214
30 654	27 087
5 643	8 675
14 087	13 675
8 000	7 465
25 986	29 875

IoU: 0.4034



Poor

IoU: 0.7330



Good

IoU: 0.9264



Excellent

# Đánh giá mô hình học máy



## Evaluation Metrics

### Classification

- *Confusion Matrix*
- *Accuracy*
- *Precision and Recall*
- *F-score*
- *AUC-ROC*
- *Log Loss*
- *Gini Coefficient*

### Regression

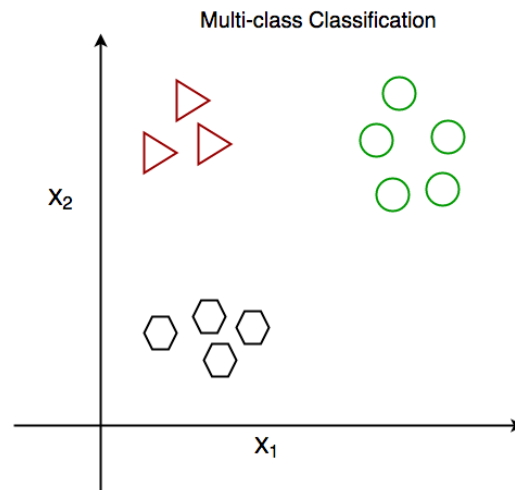
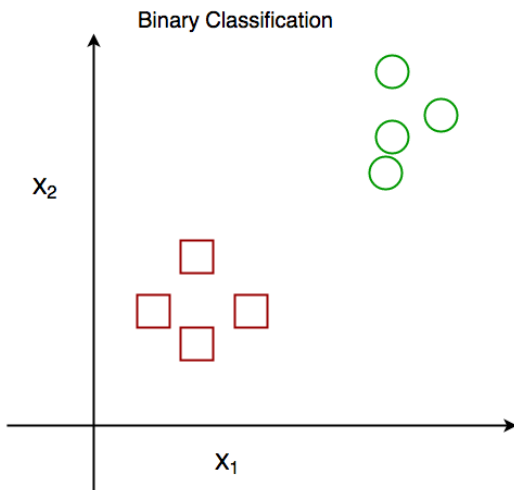
- *MAE*  
(mean abs. error)
- *MSE*  
(mean sq. error)
- *RMSE*  
(Root mean sq. error)
- *RMSLE*  
(Root mean sq. error log error)
- *$R^2$  and Adjusted  $R^2$*



## 7.1 Đánh giá mô hình phân lớp

# Đánh giá mô hình phân lớp

- Giả thiết: Có một mô hình học máy thực hiện việc dự đoán một bệnh nhân có bị đau tim hay không?
- Mô hình sau khi được huấn luyện với dữ liệu Training, thực hiện kiểm thử mô hình trên tập dữ liệu Test với số lượng 500 mẫu.



y_predict	y_groundtruth
0	0
0	0
1	0
0	0
0	1
1	1
0	0
0	0
1	0
1	1
1	0
0	1
0	0

# Đánh giá mô hình phân lớp

- Để đánh giá một hệ thống phân lớp thường sử dụng các tham số:
  - Accuracy
  - Confusion matrix
  - Precision, Recall
  - F1-score

## Evaluation Metrics for Classification Models

Accuracy, Confusion Matrix, Precision, Recall, Specificity, F1-Score

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

# 1. Accuracy (Độ chính xác)

- Cách đơn giản và hay được sử dụng nhất là Accuracy: Tính tỷ lệ giữa số điểm được dự đoán đúng và tổng số điểm trong tập dữ liệu kiểm thử.

y_predict	y_groundtruth
0	0
0	0
1	0
0	0
0	1
1	1
0	0
0	0
1	0
1	1
1	0
0	1
0	0

```
1 #Độ chính xác theo tỷ lệ số mẫu dự đoán đúng / tổng số mẫu của tập test
2 from sklearn.metrics import accuracy_score
3 acc = accuracy_score(y_groundtruth,y_predict)
4 print('Độ chính xác của mô hình phân lớp:', acc)
5 print('Độ chính xác theo %:', round(acc*100,2))
```

Độ chính xác của mô hình phân lớp: 0.792  
Độ chính xác theo %: 79.2

```
1 # Số liệu cho biết tổng số mẫu dự đoán đúng trên tập Test
2 acc2 = accuracy_score(y_groundtruth,y_predict,normalize=False)
3 print('Tổng số mẫu dự đoán đúng:', acc2)
```

Tổng số mẫu dự đoán đúng: 396

## 2. Confusion matrix

- Cách tính sử dụng accuracy như ở trên chỉ cho chúng ta biết được bao nhiêu phần trăm lượng dữ liệu được phân loại đúng (hoặc tổng có bao nhiêu mẫu phân loại đúng) mà không chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác. Để có thể đánh giá được các giá trị này, chúng ta sử dụng một ma trận được gọi là *confusion matrix*.

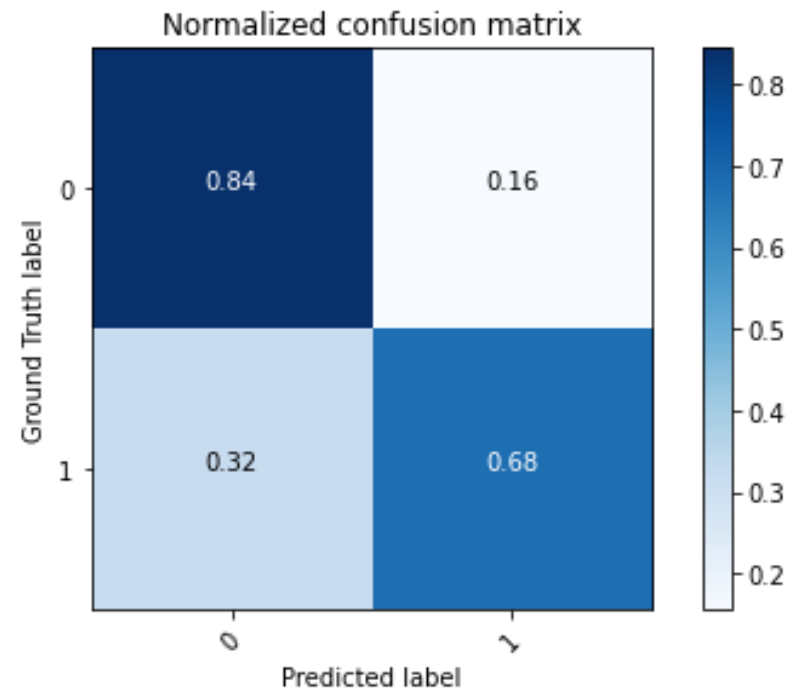
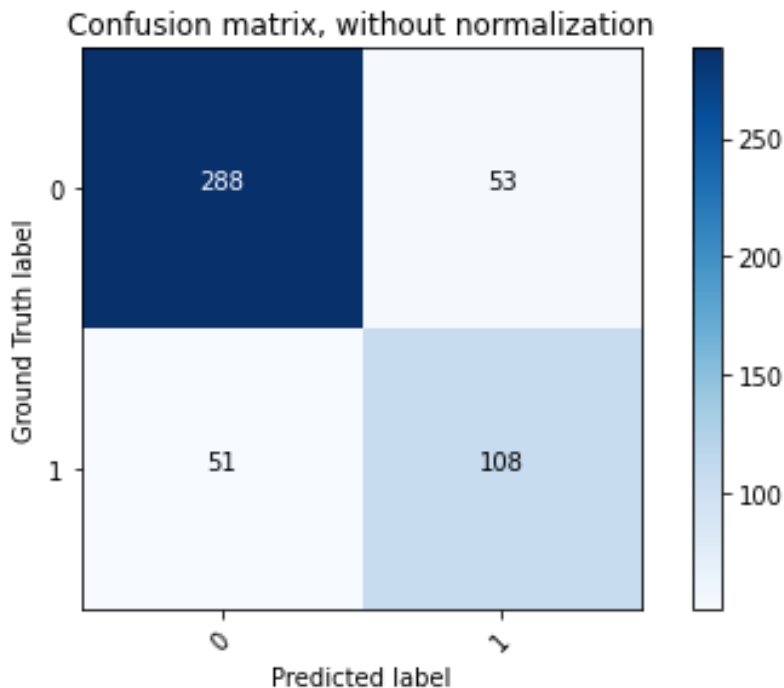
### Making sense of the confusion matrix

	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100



## 2. Confusion matrix

- Confusion matrix là một ma trận vuông với kích thước mỗi chiều bằng số lượng lớp dữ liệu. Giá trị tại hàng  $i$ , cột  $j$  là số lượng điểm lẽ ra thuộc vào lớp  $i$  nhưng lại được dự đoán là thuộc vào lớp  $j$ .



- Tổng số phần tử của toàn ma trận chính là số điểm trong tập kiểm thử. Các phần tử trên đường chéo của ma trận là số điểm được phân loại đúng của mỗi lớp dữ liệu.

### 3. Precision và Recall

- Precision được định nghĩa là tỷ lệ số điểm True positive trong số những điểm được phân loại là Positive (TP+FP)
- Recall được định nghĩa tỷ lệ số điểm True Positive trong số những điểm thực sự là Positive (TP+FN)

		Assigned class		
		Positive	Negative	
Real class	Positive	TP	FN	Recall $\frac{TP}{TP+FN}$
	Negative	FP	TN	False positive rate $\frac{FP}{TN+FP}$
		Precision $\frac{TP}{TP+FP}$	Specificity $\frac{TN}{TN+FN}$	Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

### 3. Precision và Recall

	Predicted class POSITIVE (spam 📧 )	Predicted class NEGATIVE (normal 📧 )	
Actual class POSITIVE (spam 📧 )	TRUE POSITIVE (TP) 📧 📧 320	FALSE NEGATIVE (FN) 📧 📧 43	$\text{Recall} = \frac{TP}{TP + FN} = \frac{320}{320 + 43} = 0.882$
Actual class NEGATIVE (normal 📧 )	FALSE POSITIVE (FP) 📧 📧 20	TRUE NEGATIVE (TN) 📧 📧 538	
	$\text{Precision} = \frac{TP}{TP + FP} = \frac{320}{320 + 20} = 0.941$		

Precision và Recall đều là các số không âm nhỏ hơn 1.

Một mô hình phân lớp tốt là một mô hình có cả Precision và Recall đều cao (càng gần 1 càng tốt)



### 3. Precision và Recall

```
1 from sklearn.metrics import precision_score
2
3 #Tính chỉ số precision cho tất cả các lớp
4 precision = precision_score(y_groundtruth,
5                             y_predict,
6                             average=None)
7 print('Chỉ số Precision [0,1] = ',precision)
```

Chỉ số Precision [0,1] = [0.84955752 0.67080745]

```
1 from sklearn.metrics import recall_score
2
3 #Xác định chỉ số recall cho tất cả các lớp:
4 recall = recall_score(y_groundtruth,
5                       y_predict,
6                       average=None)
7 print('Chỉ số recall[0,1] = ',recall)
```

Chỉ số recall[0,1] = [0.84457478 0.67924528]

## 4. F1 score

Nếu chỉ có **Precision** hay chỉ có **Recall** thì không đánh giá được chất lượng mô hình.

- Chỉ dùng Precision, mô hình chỉ đưa ra dự đoán cho một điểm mà nó chắc chắn nhất. Khi đó Precision = 1, tuy nhiên ta không thể nói là mô hình này tốt.
- Chỉ dùng Recall, nếu mô hình dự đoán tất cả các điểm đều là positive. Khi đó Recall = 1, tuy nhiên ta cũng không thể nói đây là mô hình tốt.

Khi đó **F1-score** được sử dụng. F1-score là trung bình điều hòa (harmonic mean) của precision và recall (giả sử hai đại lượng này khác 0). F1-score được tính theo công thức:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

## 4. F1 score

Ví dụ mối quan hệ giữa **Precision**, **Recall** và **F1**

precision	recall	$F_1$
1	1	1
0.1	0.1	0.1
0.5	0.5	0.5
1	0.1	0.182
0.3	0.8	0.36

**F1** nằm trong nửa khoảng (0,1], F1 càng cao, bộ phân lớp càng tốt.

```

1 #Chỉ số F1:
2 from sklearn.metrics import f1_score
3
4 #Xác định chỉ số F1 cho tất cả các lớp:
5 f1 = f1_score(y_groundtruth,
6               y_predict,
7               average=None)
8 print('Chỉ số F1-Score[0,1] = ',f1)

```

Chỉ số F1-Score[0,1] = [0.84705882 0.675 ]

```

1 from sklearn.metrics import classification_report
2 target_names = ['class 0', 'class 1']
3 #Các thông số đánh giá bộ phân lớp:
4 print(classification_report(y_groundtruth,
5                             y_predict,
6                             target_names=target_names))

```

	precision	recall	f1-score	support
class 0	0.85	0.84	0.85	341
class 1	0.67	0.68	0.67	159
accuracy			0.79	500
macro avg	0.76	0.76	0.76	500
weighted avg	0.79	0.79	0.79	500

# THỰC HÀNH 3.5

# Yêu cầu.



Sinh viên xem lại bài thực hành 3.2 trong Chương 3 – phần 02 về phân loại rườu với thuật toán Decsion tree.

**Thực hiện xác định các tham số Accuracy, confusion matrix, Precision, Recall, F1-score với 2 trường hợp đã làm trên tập Test**

- Sử dụng thuật toán Cây quyết định trong 2 trường hợp:
  - Sử dụng độ đo Entropy: Xác định độ chính xác của mô hình trên tập Test (25%).
  - Sử dụng độ đo Gini: Xác định độ chính xác của mô hình trên tập Test (25%).

## Evaluation Metrics for Classification Models

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
Actual: Yes	5	100

Accuracy, Confusion Matrix, Precision, Recall, Specificity, F1-Score

## 7.2 Đánh giá mô hình hồi quy

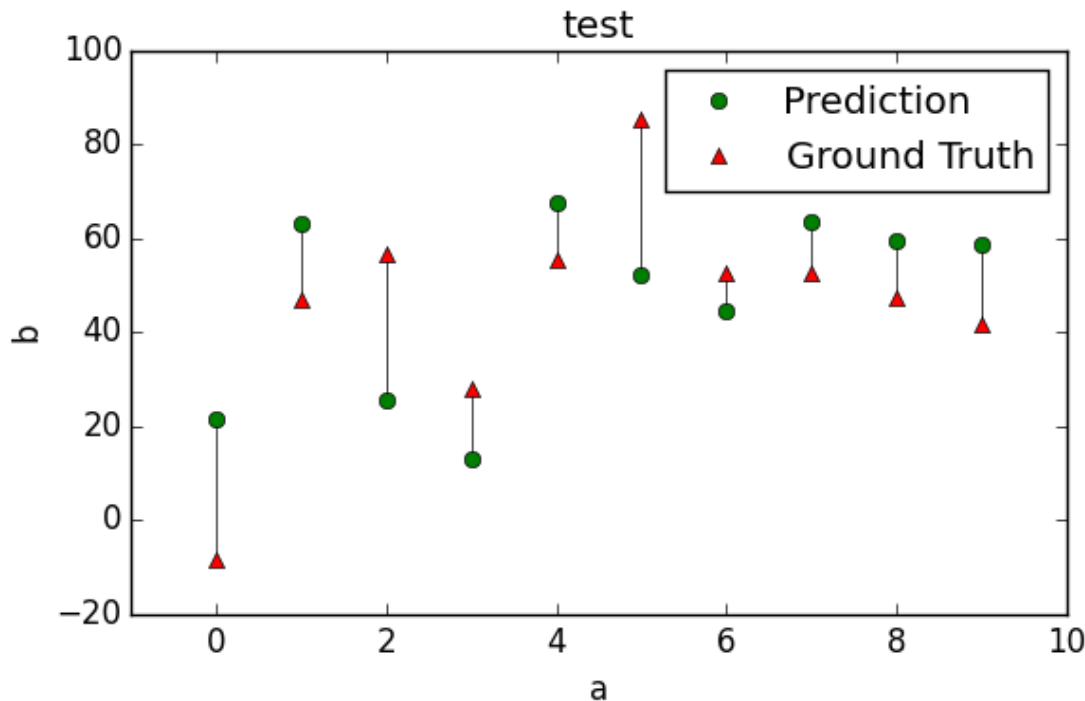
# Đánh giá mô hình hồi quy

- Giả thiết: Có một mô hình học máy thực hiện việc dự đoán giá nhà tại một khu vực?
- Mô hình sau khi được huấn luyện với dữ liệu Training, thực hiện kiểm thử mô hình trên tập dữ liệu Test với số lượng 500 mẫu.

y_predict	y_groundtruth
22 890	23 432
19 120	18 850
9 590	10 500
20 231	22 567
7 498	5 235
13 675	11 563
22 453	25 005
24 645	19 214
30 654	27 087
5 643	8 675
14 087	13 675
8 000	7 465
25 986	29 875

# Đánh giá mô hình hồi quy

- Các chỉ số cơ bản để đánh giá độ chính xác của mô hình hồi quy:



## Regression

- *MAE*  
(mean abs. error)
- *MSE*  
(mean sq. error)
- *RMSE*  
(Root mean sq. error)
- *RMSLE*  
(Root mean sq. error log error)
- $R^2$  and Adjusted  $R^2$



# 1. Sai số MAE

- Sai số tuyệt đối trung bình (MAE – Mean Absolute Error) nằm trong khoảng  $(0, +\infty)$ . MAE biểu thị biên độ trung bình của sai số mô hình nhưng không nói lên xu hướng lệch của giá trị dự đoán (predicted) và giá trị thực (Actual). Khi  $MAE = 0$ , các giá trị dự đoán hoàn toàn trùng khớp với các giá trị thực, khi đó mô hình được xem là “lý tưởng”

$$MAE = \frac{1}{n} \sum \left| \underset{\substack{\text{Actual output value}}}{y} - \underset{\substack{\text{Predicted output value}}}{\hat{y}} \right|$$

Divide by the total number of data points

Sum of

The absolute value of the residual

## 2. Sai số MSE

- Sai số bình phương trung bình (MSE) nằm trong khoảng  $(0, +\infty)$ , MSE phản ánh mức độ dao động giữa giá trị dự đoán với giá trị thực.

$$MSE = \frac{1}{n} \sum \underbrace{\left( y - \hat{y} \right)^2}_{\substack{\text{The square of the difference} \\ \text{between actual and} \\ \text{predicted}}}$$

### 3. Sai số RMSE

- **Sai số bình phương trung bình quân phương (RMSE)** là một trong những đại lượng cơ bản và thường được sử dụng phổ biến trong đánh giá độ tin cậy của mô hình hồi quy. Người ta thường hay sử dụng RMSE biểu thị độ lớn trung bình của sai số. Đặc biệt RMSE rất nhạy với những giá trị sai số lớn. Giống như MAE, RMSE không chỉ ra độ lệch giữa giá trị dự báo và giá trị thực. Giá trị của RMSE nằm trong khoảng  $(0, +\infty)$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

## 4. Hệ số $R^2$



**$R^2$ : Đánh giá tỷ lệ giải thích của mô hình ước lượng, hệ số này nằm giữa 0 và 1, càng gần 1 tỷ lệ giải thích được của mô hình càng tốt.**

- Giá trị  $R$  bình phương dao động từ 0 đến 1.  $R$  bình phương càng gần 1 thì mô hình đã xây dựng càng phù hợp với bộ dữ liệu dùng chạy hồi quy.  $R$  bình phương càng gần 0 thì mô hình đã xây dựng càng kém phù hợp với bộ dữ liệu dùng chạy hồi quy. Trường hợp đặc biệt, phương trình hồi quy đơn biến (chỉ có 1 biến độc lập) thì  $R^2$  chính là bình phương của hệ số tương quan  $r$  giữa hai biến đó.

$$\text{Coefficient of Determination} \rightarrow R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$\text{Sum of Squares Total} \rightarrow SST = \sum (y - \bar{y})^2$$

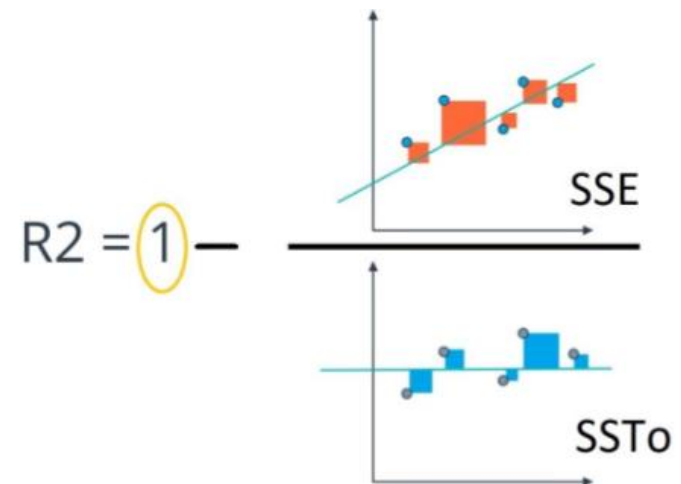
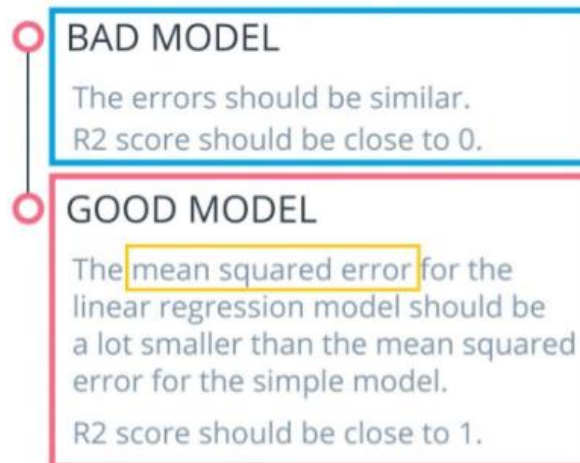
$$\text{Sum of Squares Regression} \rightarrow SSR = \sum (y' - \bar{y}')^2$$

$$\text{Sum of Squares Error} \rightarrow SSE = \sum (y - y')^2$$

# 4. Hệ số $R^2$



**Ý nghĩa  $R$  bình phương:** Giả sử  $R$  bình phương là 0.60, thì mô hình hồi quy tuyến tính này phù hợp với tập dữ liệu ở mức 60%. Nói cách khác, 60% biến thiên của biến phụ thuộc được giải thích bởi các biến độc lập. (còn 40% còn lại ở đâu, dĩ nhiên là do sai số đo lường, do cách thu thập dữ liệu, do có thể có biến độc lập khác giải thích cho biến phụ thuộc mà chưa được đưa vào mô hình nghiên cứu...vv). Thông thường, ngưỡng của  $R^2$  phải trên 50%, vì như thế mô hình mới phù hợp. Tuy nhiên tùy vào dạng nghiên cứu, như các mô hình về tài chính, không phải tất cả các hệ số  $R^2$  đều bắt buộc phải thỏa mãn lớn hơn 50%. (do rất khó để dự đoán giá vàng, giá cổ phiếu mà chỉ đơn thuần dựa vào các biến độc lập ví dụ GDP, ROA, ROE....)



# THỰC HÀNH 3.6

# Yêu cầu.

---



- Sinh viên xem lại bài thực hành 3.3 trong Chương 3 – phần 03 về bài toán hồi quy với tập dữ liệu Diabetes Dataset .
- Xây dựng mô hình hồi quy tuyến tính với tất các thuộc tính đầu vào (input). Xác định sai số MAE, MSE, RMSE và  $R^2$  trên tập Train và Test.

Evaluating Machine Learning Models



# Thank you!