

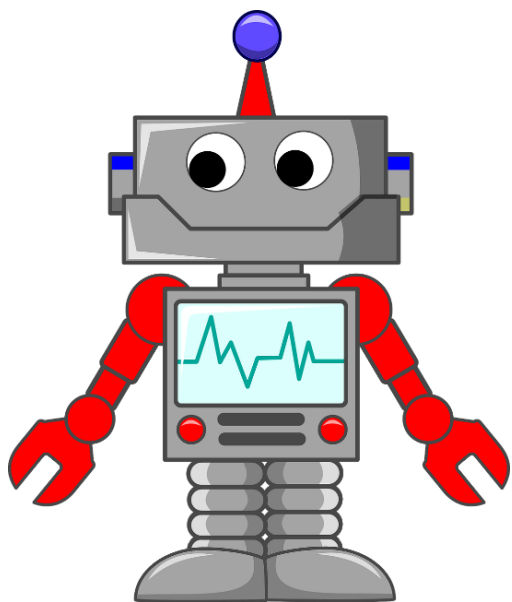


CS116 – LẬP TRÌNH PYTHON CHO MÁY HỌC

BÀI 05

ĐÁNH GIÁ MÔ HÌNH

TS. Nguyễn Vinh Tiệp





NỘI DUNG

1. TẠI SAO CẦN PHẢI ĐÁNH GIÁ MÔ HÌNH

2. ĐỘ ĐO CHO BÀI TOÁN HỒI QUY

3. ĐỘ ĐO CHO BÀI TOÁN PHÂN LỚP

4. QUY TRÌNH ĐÁNH GIÁ KHÁCH QUAN
protocol



Tại sao cần đánh giá mô hình

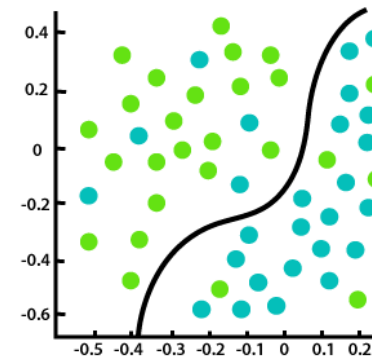
đánh giá bài toán trước khi chọn mô hình nào

- **Xác thực tính hiệu quả :**
 - Mô hình thực sự đoán chính xác hay chỉ nhớ dữ liệu train?
 - Có bị hiện tượng **overfitting**?
- **Chọn lựa mô hình:** cho phép so sánh các mô hình với nhau, từ đó chọn ra mô hình tốt nhất
định lượng
- **Sự tin cậy:** mô hình của mình có ổn định và đáng tin cậy khi triển khai thực tế không? khách hàng tin tưởng khi mô hình hoạt động tốt
- **Định hướng cải thiện mô hình:** các độ đo đánh giá giúp ta biết cần cải tiến mô hình theo hướng nào?

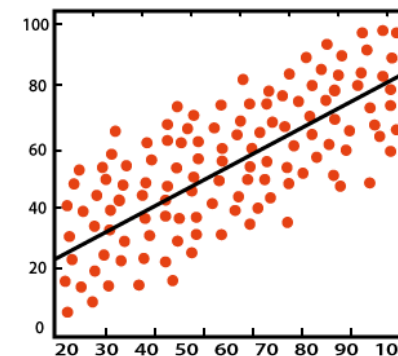


Phân biệt bài toán học máy

- Hai loại bài toán chính:
 - Hồi quy – Regression
 - Phân lớp – Classification
- Cách phân biệt “**đơn giản**” dựa trên tính chất của biến kết quả (output)
 - Hồi quy: đầu ra là giá trị **liên tục**
 - Phân lớp: đầu ra là giá trị **rời rạc**
- Câu hỏi:** Bài toán “dự đoán tuổi” là Hồi quy hay Phân lớp?

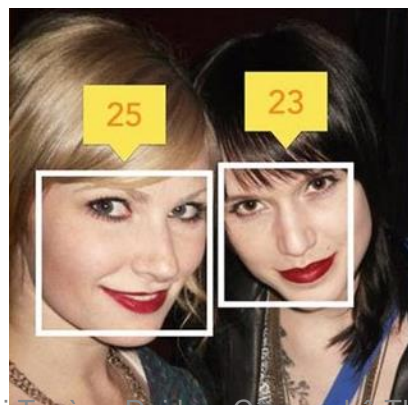


Classification



Regression

hồi quy vì giá trị là liên tục 0 -> 150, phân lớp vì tuổi nằm trong tập hợp {0,1,2,...150}



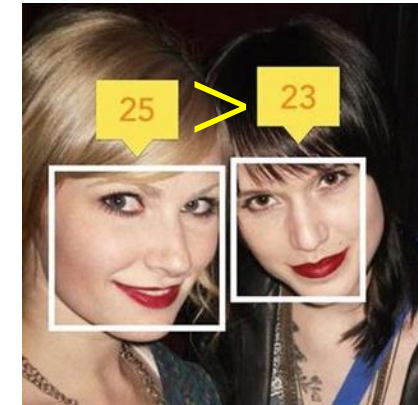
Có 2 luồng ý kiến trái chiều



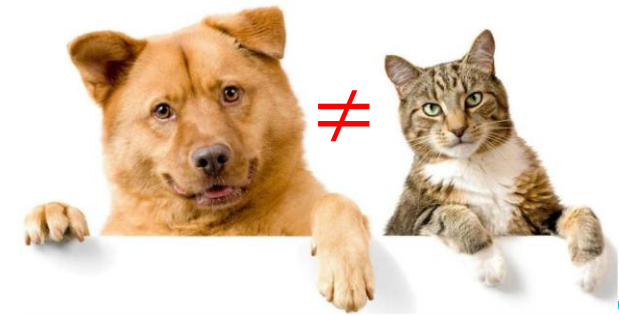
Phân biệt bài toán học máy

- Cách phân biệt “rõ ràng” hơn:
 - Hồi quy: 2 giá trị output bất kỳ có thể thực hiện tất cả so sánh $>$, $<$, $=$, \neq
 - Phân lớp: 2 giá trị output bất kỳ **chỉ có thể** thực hiện được so sánh $=$ hoặc \neq

- **Đáp án:** Bài toán “dự đoán tuổi” là **bài toán hồi quy**



- Tương tự, bài toán đoán đối tượng: “chó”, “mèo” là bài toán phân loại?
 $\text{chó} > \text{mèo}$: không được vì vậy là phân lớp





Tại sao cần Phân biệt đúng bài toán

- Để chọn lựa **hàm mục tiêu huấn luyện phù hợp**
 - Hàm loss/hàm độ lỗi đúng việc huấn luyện nhanh hơn
/hàm mất mát
- Để chọn lựa **độ đo đánh giá phù hợp**
 - Đánh giá khách quan hơn



NỘI DUNG

1. TẠI SAO CẦN PHẢI ĐÁNH GIÁ MÔ HÌNH

độ đo hàm lỗi -> dùng trong huấn luyện

2. ĐỘ ĐO CHO BÀI TOÁN HỒI QUY

độ đo đánh giá -> sau khi huấn luyện

3. ĐỘ ĐO CHO BÀI TOÁN PHÂN LỚP

4. QUY TRÌNH ĐÁNH GIÁ KHÁCH QUAN



Bài toán hồi quy: chọn hàm độ lỗi (hàm loss)

hàm dự đoán cho kết quả gần xấp xỉ thực tế

- Xét ví dụ sau:

Biến đầu vào					Biến đầu ra
	R&D Spend	Administration	Marketing Spend	State	Profit
1	165349.2	136897.8	471784.1	New York	192261.83
2	162597.7	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.9	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.6

train

test

Giá trị thực tế

Mô hình
máy học
 $f_{\theta}(x)$

Dự đoán
167187.9
155991.1
152122.5
159752.6

y

\tilde{y}

hồi quy thường dùng
hàm lỗi Mean Square
Error

Hàm lỗi MSE

$$\frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$



Bài toán hồi quy: độ đo đánh giá

- Độ đo đánh giá của bài toán hồi quy cũng tương tự hàm lỗi

	R&D Spend	Administration	Marketing Spend	State	Profit
1	165349.2	136897.8	471784.1	New York	192261.83
2	162597.7	151377.59	443898.53	California	191792.06
3	153441.51	101145.55	407934.54	Florida	191050.39
4	144372.41	118671.85	383199.62	New York	182901.99
5	142107.34	91391.77	366168.42	Florida	166187.94
6	131876.9	99814.71	362861.36	New York	156991.12
7	134615.46	147198.87	127716.82	California	156122.51
8	130298.13	145530.06	323876.68	Florida	155752.6

train

test

Giá trị thực tế

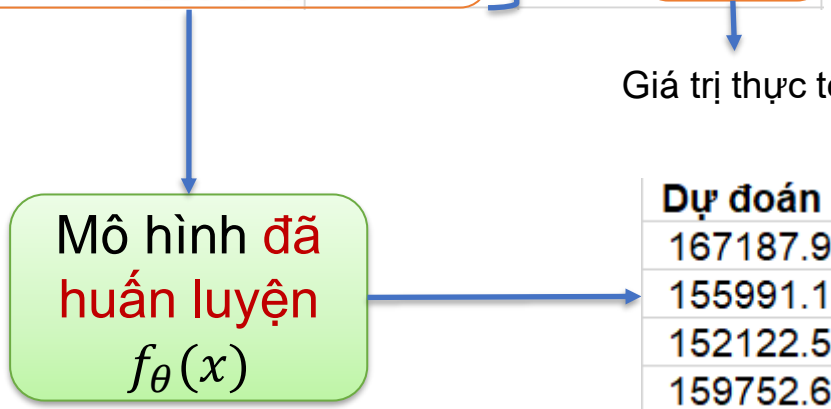
y

Độ đo đánh giá:

- MAE
- MSE
- RMSE
- ...

\tilde{y}

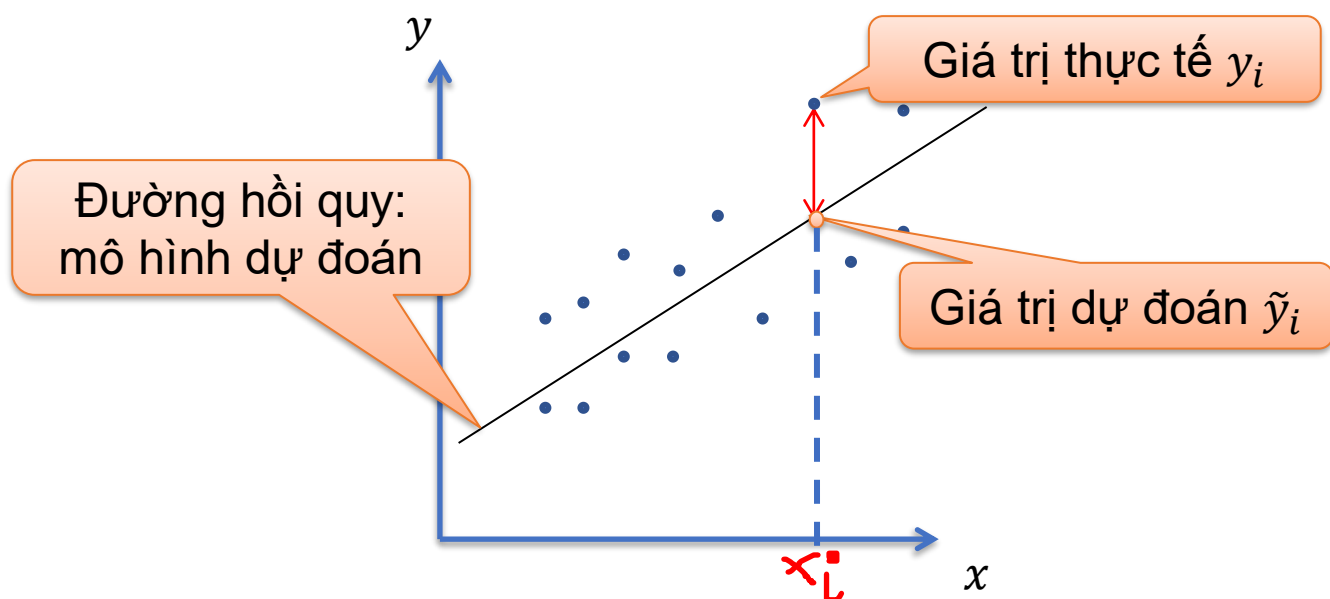
test





Bài toán hồi quy: độ đo đánh giá

- Các độ đo đánh giá cho bài toán hồi quy



Độ đo: đo lường sai số giữa dữ liệu thực với đường hồi quy:

absolute

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \tilde{y}_i|$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

Root

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

relative

$$RAE = \frac{\sum_{i=1}^n |y_i - \tilde{y}_i|}{\sum_{i=1}^n |y_i - \bar{y}|}$$

relative to

Square

$$RSE = \frac{\sum_{i=1}^n (y_i - \tilde{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$



NỘI DUNG

1. TẠI SAO CẦN PHẢI ĐÁNH GIÁ MÔ HÌNH

2. ĐỘ ĐO CHO BÀI TOÁN HỒI QUY

3. ĐỘ ĐO CHO BÀI TOÁN PHÂN LỚP

4. QUY TRÌNH ĐÁNH GIÁ KHÁCH QUAN

Bài toán phân lớp: chọn hàm độ lỗi (hàm loss)

- Xét ví dụ sau:

Biến đầu vào									Biến phân loại đầu ra
tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
11	33	7	136	5	5	0	1	1	1
33	33	12	33	2	0	0	0	0	1
23	30	9	30	1	2	0	0	0	0
38	35	5	76	2	10	1	1	1	0
7	35	14	80	2	15	0	1	0	1
11	33	7	136	5	5	0	1	1	1
33	33	12	33	2	0	0	0	0	1
23	30	9	30	1	2	0	0	0	0
38	35	5	76	2	10	1	1	1	0
7	35	14	80	2	15	0	1	0	1

Train

Test

Giá trị thực tế

y

\tilde{y}

Mô hình
máy học
 $f_{\theta}(x)$

predict
1
1
1
0
1

Hàm lỗi MSE? có thể

$$\frac{1}{n} \sum_{i=1}^n (y_i - \tilde{y}_i)^2$$

val: spam / not spam

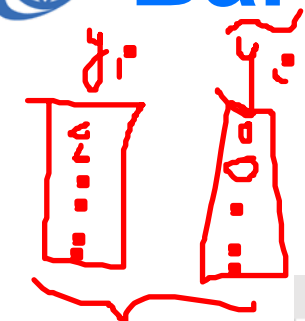
tuy nhiên làm quá trình
huấn luyện chậm vì y_i
- y_i thì từ 0->1 rất bé
nên chạy chậm



Bài toán phân lớp: hàm độ lỗi log loss

thường dùng cho bài toán k class dùng log loss

cho 1 vector



1 mẫu

tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
11	33	7	136	5	5	0	1	1	1
33	33	12	33	2	0	0	0	0	1
23	30	9	30	1	2	0	0	0	0
38	35	5	76	2	10	1	1	1	0
7	35	14	80	2	15	0	1	0	1

Train

predict prob	Log loss
0.83	0.08
0.75	0.12
0.96	1.40
0.32	0.17
0.91	0.04

= - [0.log0.96 + 1log(1-0.96)]

Giá trị thực tế y

Xác suất dự đoán \tilde{y}

Log loss: 0.0 0.1 0.2 1.0

Độ chính xác càng cao

Lỗi mẫu thứ i:

$$-[y \log \tilde{y} + (1 - y) \log(1 - \tilde{y})]$$

Log loss:

$$-\frac{1}{n} \sum_{i=1}^n y_i \log \tilde{y}_i + (1 - y_i) \log(1 - \tilde{y}_i)$$

Trung bình cộng



Bài toán phân lớp: độ đo đánh giá (1)

- Tại sao không sử dụng log loss làm độ đo đánh giá?

→ không nên vì không hằng

- Sử dụng Jaccard index

- Y : Tập giá trị thực tế
- \tilde{Y} : Tập giá trị dự đoán

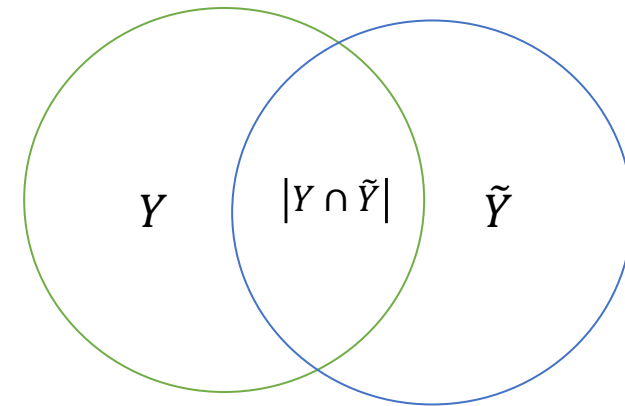
$$Y = \tilde{Y} \Rightarrow J = 1$$

$$J(Y, \tilde{Y}) = \frac{|Y \cap \tilde{Y}|}{|Y \cup \tilde{Y}|} = \frac{|Y \cap \tilde{Y}|}{|Y| + |\tilde{Y}| - |Y \cap \tilde{Y}|}$$

$$Y: [1, 1, 0, 0, 1, 0, 1]$$

$$\tilde{Y}: [1, 1, 1, 0, 1, 0, 0]$$

$$J(Y, \tilde{Y}) = \frac{5}{7 + 7 - 5} = 0.41$$





Bài toán phân lớp: độ đo đánh giá (2)

- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1-score = $2 (Pr \times Rec) / (Pr + Rec)$

Trung bình điều hòa

F1-score: 0.0 0.1 0.2 1.0



Độ chính xác càng cao

	Precision	Recall	F1-score
Churn=0	$24 / (24 + 1)$	$24 / 25$	$24 / 25$
Churn=1	$6 / 7$	$6 / 15$	

Trung bình: ?

$$\frac{c + b}{2}$$

Confusion matrix

	Churn=1	Churn=0
Churn=1	6 TP	9 FN
Churn=0	1 FP	24 TN

Nhãn thực tế y

Nhãn dự đoán y

Precision

Recall



Nhận xét

$churn = 0 \rightarrow$ weigh (lớn)
 $churn = 1 \rightarrow$ nhỏ

$0.8 F_1$ $churn = 0$ $40.2 F_1$
không nhất thiết là tính trung bình

- Các độ đo đánh giá, hàm độ lỗi trên là cách tính trung bình, thực tế có thể có **tính thiên lệch**
- Một số độ đo đánh giá bắt nguồn **yếu tố thực tiễn**: doanh thu, lợi nhuận, lợi nhuận biên, thiệt hại kinh tế,...
- Một số hàm độ lỗi lấy trực tiếp từ độ đo đánh giá

dùng vậy luôn trực tiếp thì lúc này đánh giá sẽ sát dễ đúng hơn vì học từ độ đo đánh giá đó luôn mà



NỘI DUNG

1. TẠI SAO CẦN PHẢI ĐÁNH GIÁ MÔ HÌNH

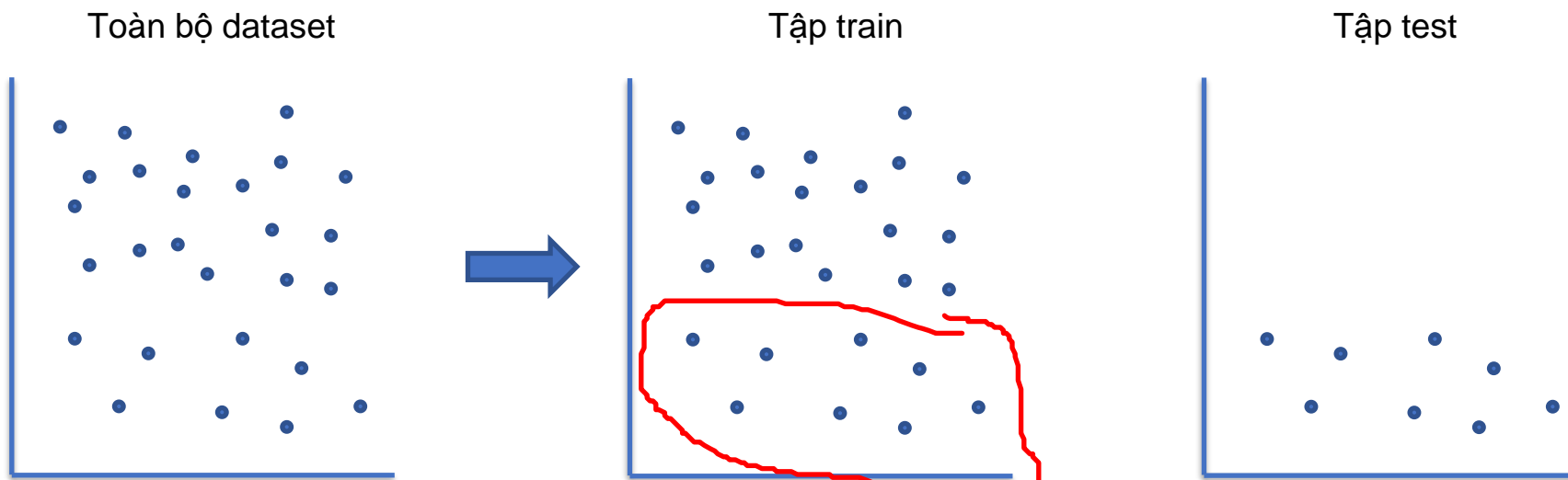
2. ĐỘ ĐO CHO BÀI TOÁN HỒI QUY

3. ĐỘ ĐO CHO BÀI TOÁN PHÂN LỚP

4. QUY TRÌNH ĐÁNH GIÁ KHÁCH QUAN

Quy trình 1: Tập test là một phần tập train

- Sử dụng một tập con của tập train làm dữ liệu test:



không khách quan,
overfit, chỉ phù hợp với
dữ liệu ít, thời gian chạy rất
chậm mới nên dùng cách này

Độ chính xác **cao** trên tập train (và test)
Độ chính xác **thấp** trên tập ngoài mẫu
(out-of-sample dataset)



Quy trình 1: Tập test là một phần tập train

- Độ chính xác trên tập train
 - Độ chính xác cao không phản ánh đúng hiệu quả của mô hình
 - Là kết quả của hiện tượng over-fitting (quá khớp): mô hình quá đầ trên tập dữ liệu dẫn đến học cả nhiễu và không có tính tổng quát
- Độ chính xác trên tập ngoài mẫu (out-of-sample)
 - Độ chính xác trên tập ngoài mẫu phản ánh hiệu quả khi áp dụng thực tế
 - Làm sao để cải thiện độ chính xác trên tập ngoài mẫu?

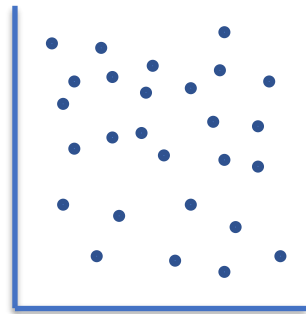


Quy trình 2: Train-Test tách biệt (train-test split)

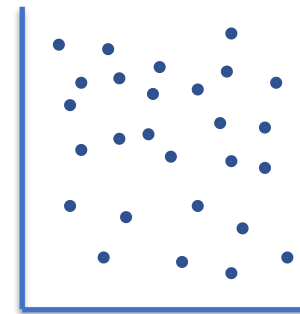
đảm bảo không trùng lặp, nhưng không đảm bảo yếu tố khách quan, vì tập test có thể quá dễ mô hình có acc cao, tập test khó thì cũng không phản ánh đúng mô hình mình làm

Train và test trên cùng dataset

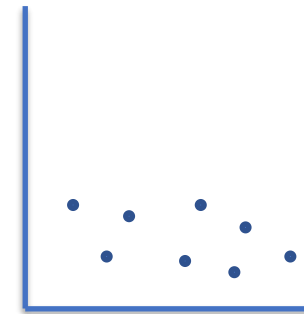
Toàn bộ dataset



Tập train

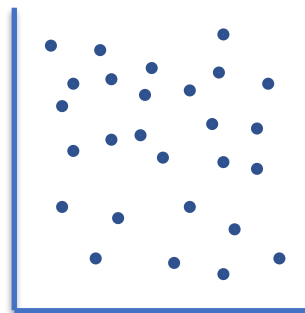


Tập test

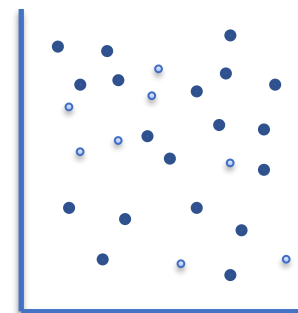


Train và test riêng biệt

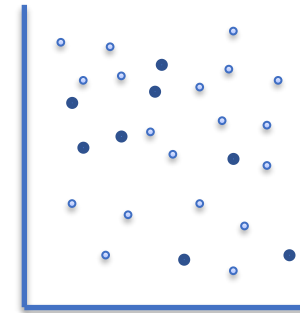
Toàn bộ dataset



Tập train



Tập test



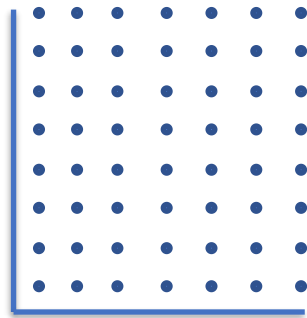
tập train là tô đậm, bên đây tập test là đậm



Quy trình 3: Kiểm định chéo – K-Fold cross validation

khung đỏ là tập test, còn lại các data trong 1 fold là để train

Toàn bộ dataset



- Sử dụng ở bước test
- Sử dụng ở bước train

độ chính xác

K=4



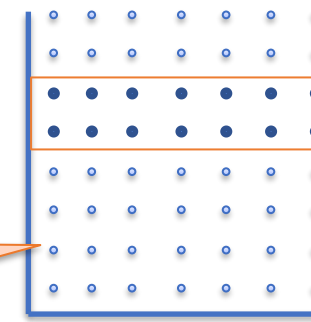
Fold 1

80%



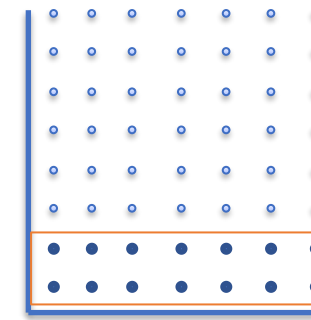
Fold 2

84%



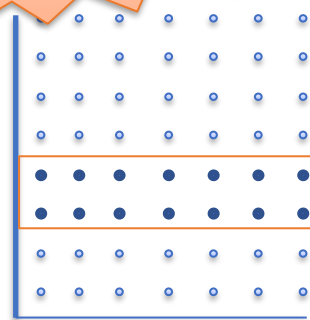
Fold 4

86%



Trung bình:
83%

82%



qua 4 fold thì đã test hết dataset, và test train độc lập nhau



So sánh các quy trình đánh giá

	Test là một phần của Train	Train-Test split	K-Fold Cross Validation
Ưu điểm	<ul style="list-style-type: none">- Đơn giản: Không cần thêm quy trình đánh giá phức tạp- Nhanh, do tận dụng chính tập train làm tập test	<ul style="list-style-type: none">- Đơn giản, dễ thực hiện- Ít tốn kém chi phí tính toán so với Cross Validation- Phù hợp khi tập dữ liệu lớn	<ul style="list-style-type: none">- Đánh giá toàn diện về tính tổng quát của mô hình- Phù hợp khi tập dữ liệu ít hoặc huấn luyện nhanh
Khuyết điểm	<ul style="list-style-type: none">- Đánh giá không khách quan- Tính tổng quát của mô hình không được đảm bảo	<ul style="list-style-type: none">- Tốn thêm thời gian- Không khách quan so với cross validation	<ul style="list-style-type: none">- Chi phí tính toán cao do huấn luyện nhiều lần



BÀI QUIZ VÀ HỎI ĐÁP