

Hồi quy (Regression)

Nguyễn Thanh Tùng

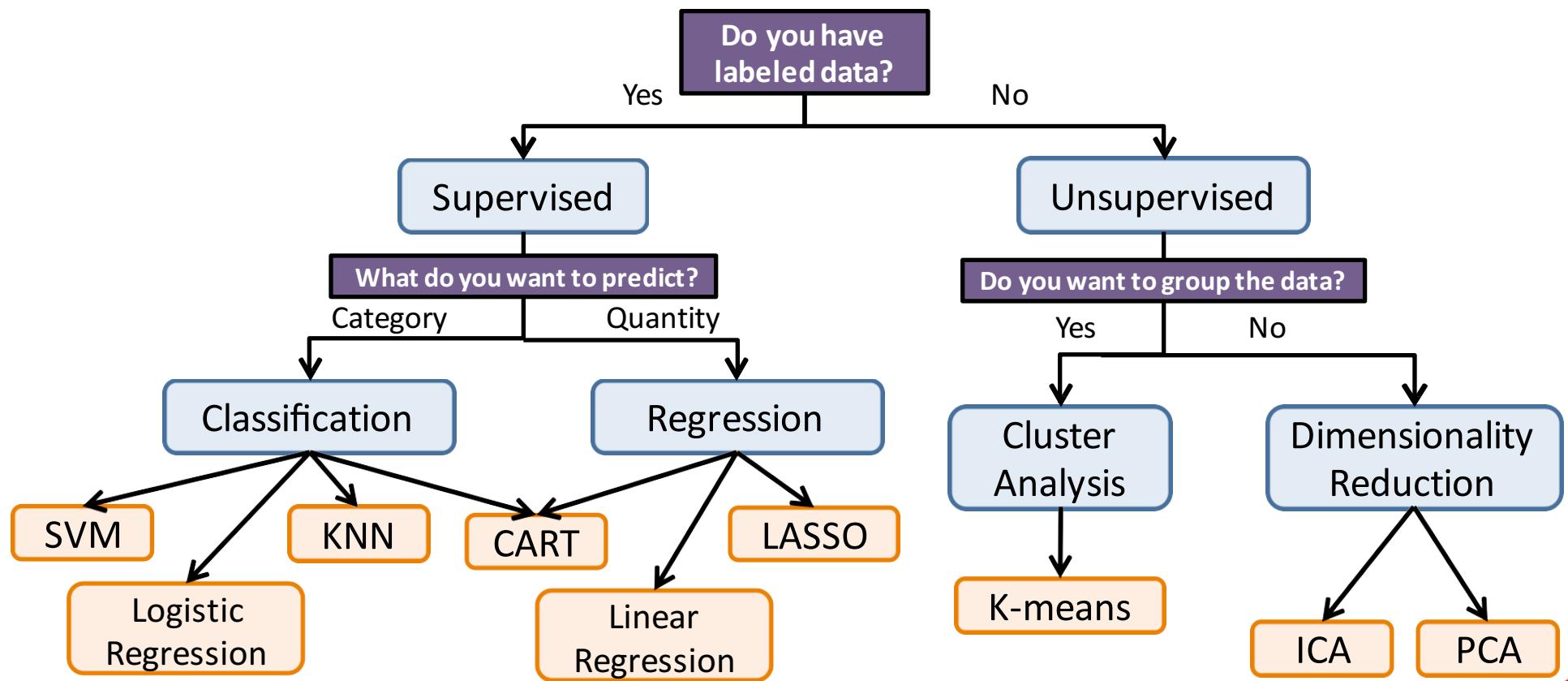
Bài giảng của DSLab

Viện nghiên cứu cao cấp về Toán (VIASM)

Nội dung

1. Giới thiệu mô hình hồi quy
2. Overfitting, kỹ thuật đánh giá chéo
3. Phân tích dữ liệu với R
4. Hồi quy tuyến tính
5. Hồi quy phi tuyến
6. Real-life problem

Các dạng giải thuật học máy



Mô hình Hồi quy

- Xét:
$$Y = f(X) + \epsilon$$
- Các phương pháp học giám sát:
 - Học bởi các ví dụ (quan sát)-“Learn by example”
 - Xây dựng mô hình \hat{f} sử dụng tập các quan sát đã được gắn nhãn
 - Y có kiểu dữ liệu liên tục

Ví dụ về Quảng cáo

- Doanh nghiệp có thể điều chỉnh chiến lược quảng cáo sản phẩm (advertising) để tăng doanh số bán hàng (sales).
- Dữ liệu: Doanh số bán hàng và ngân sách quảng cáo cho 3 phương tiện truyền thông (TV, radio, newspaper).

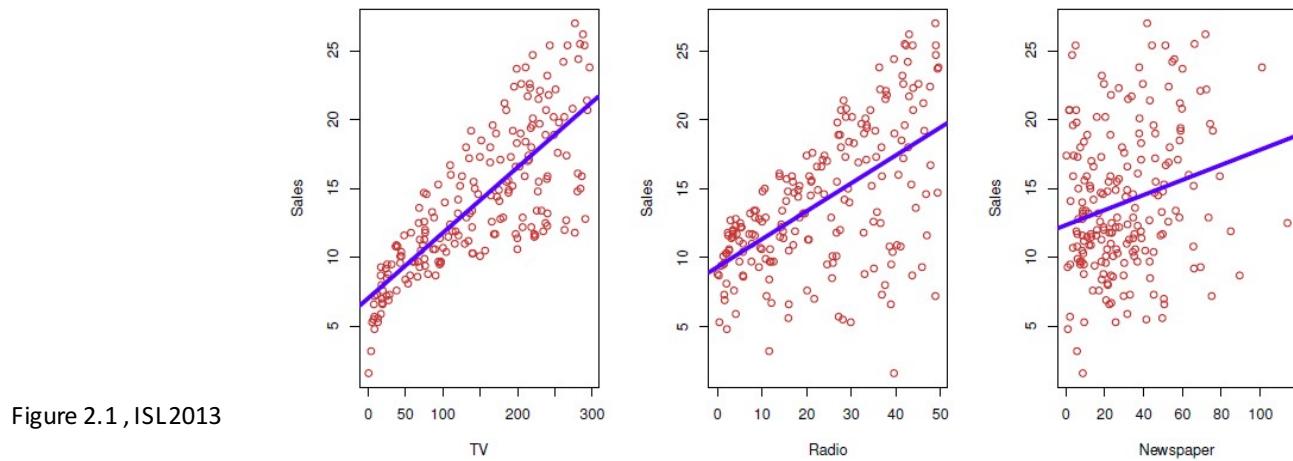


Figure 2.1 , ISL2013

Mô hình Hồi quy

- Giải thuật học
 - Lấy hàm ước lượng “tốt nhất” \hat{f} trong tập các hàm
- Ví dụ: Hồi quy tuyến tính
 - Chọn 1 ước lượng tốt nhất từ *dữ liệu học* trong tập các hàm tuyến tính

$$f(X) = \beta_0 + \beta_1 X_1 + \dots + \beta_d X_d$$

Hàm tổn thất

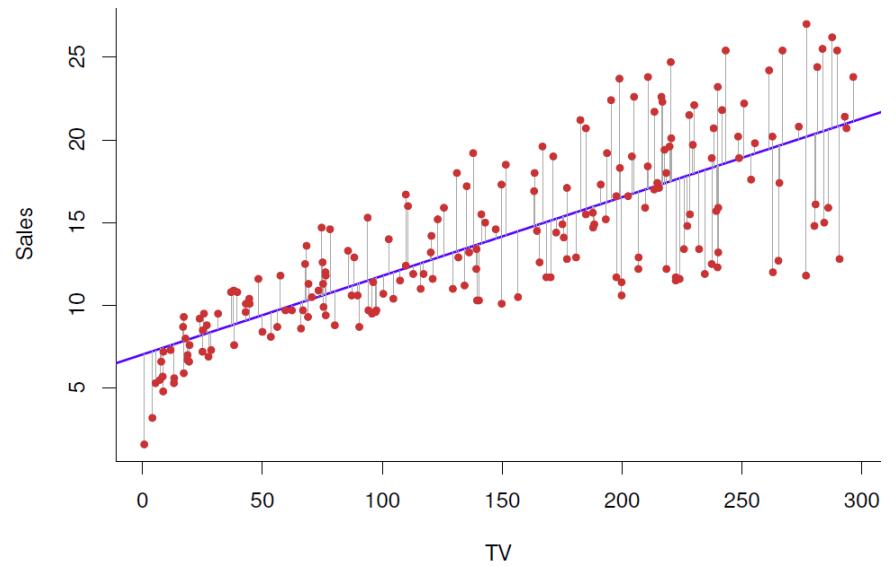
$$L(\theta_i, \hat{\theta}_i)$$

Sai số bình phương (Squared error)

$$\sum_i (\theta_i - \hat{\theta}_i)^2$$

Sai số tuyệt đối (Absolute error)

$$\sum_i |\theta_i - \hat{\theta}_i|$$



Bài toán Hồi quy

$$\hat{f} = \operatorname{argmin}_{\tilde{f}} E[L(Y, \tilde{f}(X))]$$

argument minimum: Cho giá trị nhỏ nhất của 1 hàm số trong miền xác định

Đo hiệu năng bài toán hồi quy

- Hàm tổn thất (Loss function): loại hàm dùng để đo lường sai số của mô hình
- Vd: Sai số bình phương trung bình (Mean squared error - MSE)
 - Độ đo thông dụng dùng để tính độ chính xác bài toán hồi quy

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(\hat{y}^{(i)} - y^{(i)} \right)^2$$

- Tập trung đo các sai số lớn hơn là các sai số nhỏ

Nội dung

1. Giới thiệu mô hình hồi quy
2. Overfitting, kỹ thuật đánh giá chéo
3. Phân tích dữ liệu với R
4. Hồi quy tuyến tính
5. Hồi quy phi tuyến
6. Real-life problem

Hiện tượng quá khớp

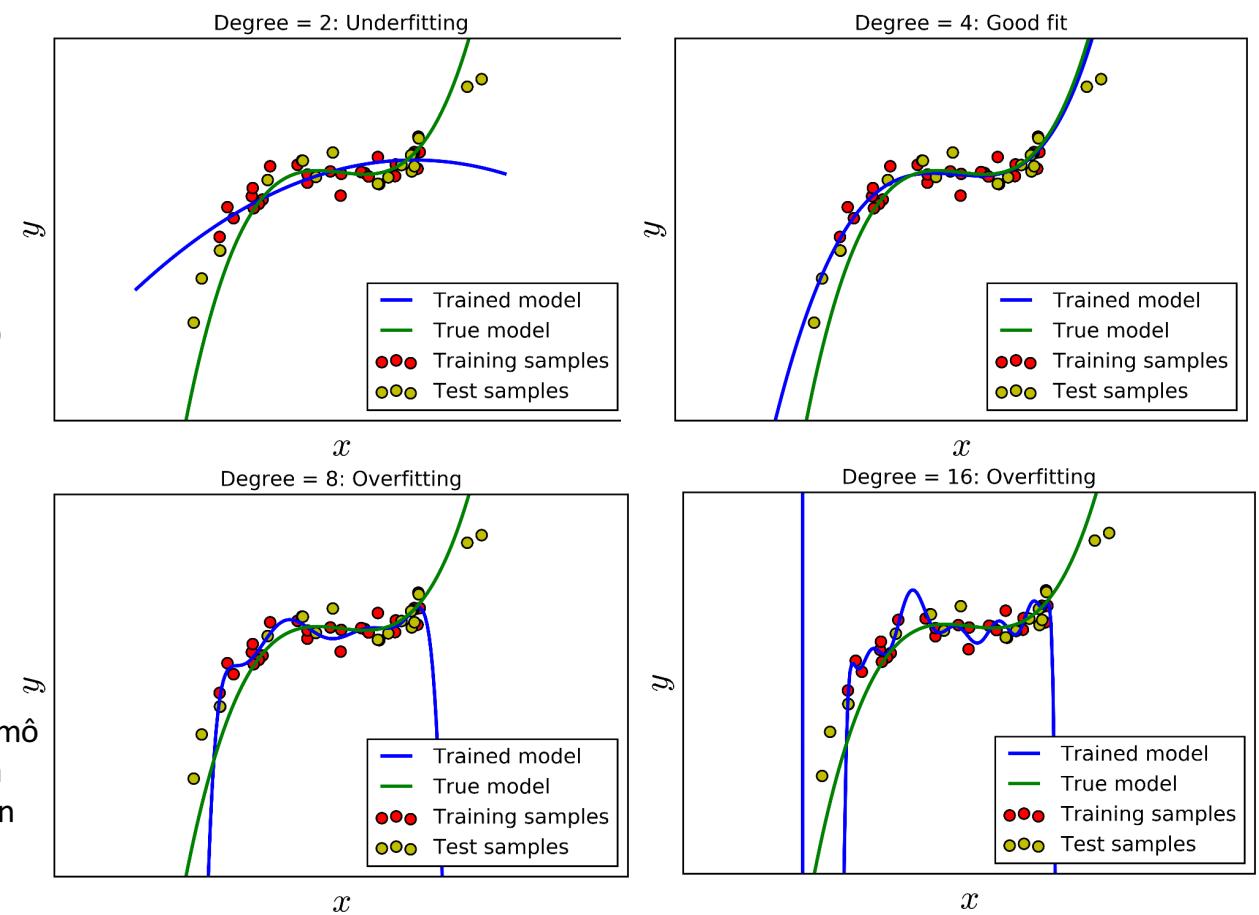
Overfitting

Vấn đề: Overfitting

- *Quá khớp (Overfitting):* Học sự biến thiên ngẫu nhiên trong dữ liệu hơn là xu hướng cơ bản
- Đặc điểm của overfitting:
 - Mô hình có hiệu năng cao trên dữ liệu học nhưng kém trên tập dữ liệu thử nghiệm.

Underfitting và Overfitting

- Có 50 điểm dữ liệu được tạo bằng một đa thức bậc ba cộng thêm nhiễu.
- Đồ thị của đa thức có màu xanh lục (true model).
- Bài toán: Giả sử ta không biết mô hình ban đầu mà chỉ biết các điểm dữ liệu, hãy tìm một mô hình “tốt” để mô tả dữ liệu đã cho?
 - Với $d=2$, mô hình không thực sự tốt vì dự đoán quá khác so với mô hình thực: *underfitting*
 - Với $d=8$ và $d=16$, với các điểm dữ liệu trong khoảng của training data, mô hình dự đoán và mô hình thực là khá giống nhau. Tuy nhiên, về phía phải, đa thức bậc 8 và 16 cho kết quả hoàn toàn ngược với xu hướng của dữ liệu: *Overfitting*.
 - $d=4$, mô hình tốt nhất.



Đánh giá hiệu năng

- Lỗi huấn luyện và lỗi kiểm thử thể hiện khác nhau
 - Tính linh hoạt của mô hình tăng lên...
 - *Lỗi huấn luyện* giảm
 - *Lỗi kiểm thử ban đầu* giảm,
Nhưng sau đó tăng lên vì overfitting → “U-shaped” lỗi kiểm thử dạng chữ U.

Đánh giá hiệu năng

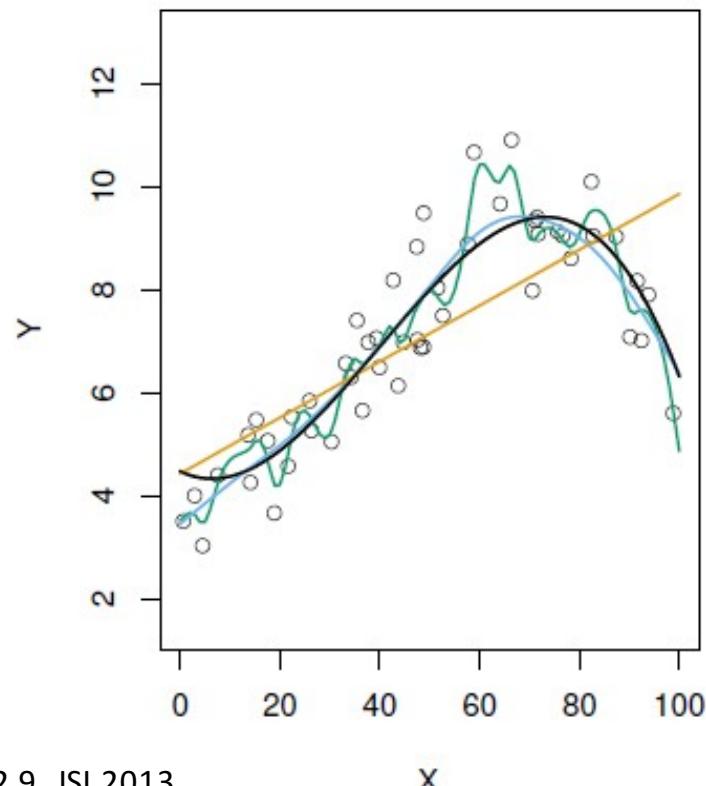
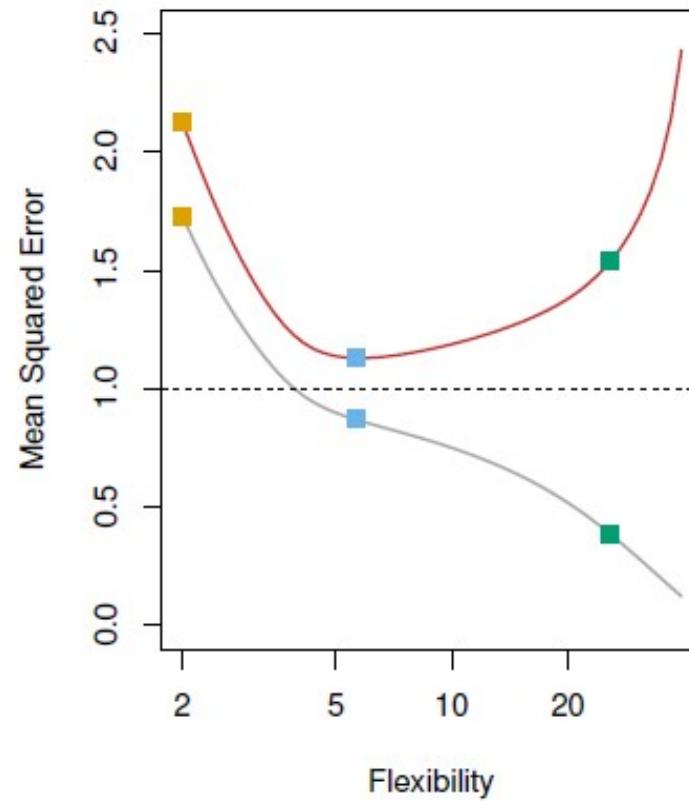


Figure 2.9 , ISL2013

15



Đánh giá hiệu năng

- Làm sao để ước lượng lượng lỗi kiểm thử để tìm một mô hình tốt?
- *Kỹ thuật đánh giá chéo (Cross-validation):*
một tập các kỹ thuật nhằm sử dụng dữ liệu huấn luyện để ước lượng lỗi tổng quát (generalization error)

Dữ liệu

- *Dữ liệu huấn luyện (Training data)*
 - Tập các quan sát (bản ghi) được sử dụng để xây dựng (học) mô hình.
- *Dữ liệu kiểm chứng (Validation data)*
 - Tập các quan sát dùng để ước lượng lỗi nhằm tìm tham số hoặc lựa chọn mô hình.
- *Dữ liệu kiểm thử (Test data)*
 - Tập các quan sát dùng để đánh giá hiệu năng trên dữ liệu chưa biết (unseen) trong tương lai.
 - Dữ liệu này không sử dụng cho giải thuật học máy trong quá trình xây dựng mô hình.

Kỹ thuật đánh giá chéo

Cross-validation

Kỹ thuật đánh giá chéo

“Dùng lỗi trên tập dữ liệu kiểm thử để ước lượng lỗi dự đoán”

$$err = E[L(Y, \hat{f}(X))]$$

Tập đánh giá (Validation)

- Thường chia tập dữ liệu ra thành training data và test data.
- Chú ý: khi xây dựng mô hình, ta không được sử dụng test data.
- Làm cách nào để biết được chất lượng của mô hình với unseen data (tức dữ liệu chưa nhìn thấy bao giờ)?

Tập đánh giá (Validation)

- Phương pháp: trích từ training data ra một tập con nhỏ và thực hiện việc đánh giá mô hình trên tập con này.
- Tập con nhỏ được trích ra từ training set này được gọi là validation set. Lúc này, training set là phần còn lại của training set ban đầu.
- Train error được tính trên training set mới này.
- Validation error: Lỗi được tính trên tập validation.

Tập đánh giá (Validation)

- Tìm mô hình sao cho cả *train error* và *validation error* đều nhỏ, qua đó có thể dự đoán được rằng *test error* cũng nhỏ.
- Phương pháp thường được sử dụng là sử dụng nhiều mô hình khác nhau. Mô hình nào cho *validation error* nhỏ nhất sẽ là mô hình tốt.

Tập đánh giá (Validation)

- Tuy nhiên, khi ta có rất hạn chế số lượng dữ liệu để xây dựng mô hình. Nếu lấy quá nhiều dữ liệu trong tập training ra làm dữ liệu validation, phần dữ liệu còn lại của tập training là không đủ để xây dựng mô hình.
- Nếu ta giữ tập validation phải thật nhỏ để có được lượng dữ liệu cho training đủ lớn. Một vấn đề khác nảy sinh, hiện tượng overfitting lại có thể xảy ra với tập training còn lại.
- **Giải pháp: Cross-validation (Kỹ thuật đánh giá chéo).**



Kỹ thuật đánh giá chéo

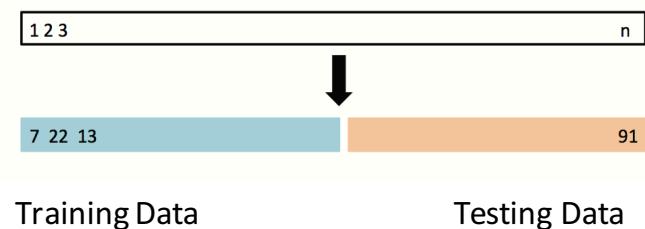
- *Cross validation* là một cải tiến của *validation* với lượng dữ liệu trong tập validation là nhỏ nhưng chất lượng mô hình được đánh giá trên nhiều tập validation khác nhau.
- Chia tập training ra k tập con không có phần tử chung, có kích thước gần bằng nhau.
- Tại mỗi lần kiểm thử, một trong số k tập con được lấy ra làm *validata set*. Mô hình sẽ được xây dựng dựa vào hợp của $k-1$ tập con còn lại.
- Mô hình cuối được xác định dựa trên trung bình của các *train error* và *validation error*.

Cách làm này còn có tên gọi là **k-fold cross validation**.

Tập huấn luyện - Training Set

Tập kiểm thử - Test Set

Tập đánh giá - Validation Set

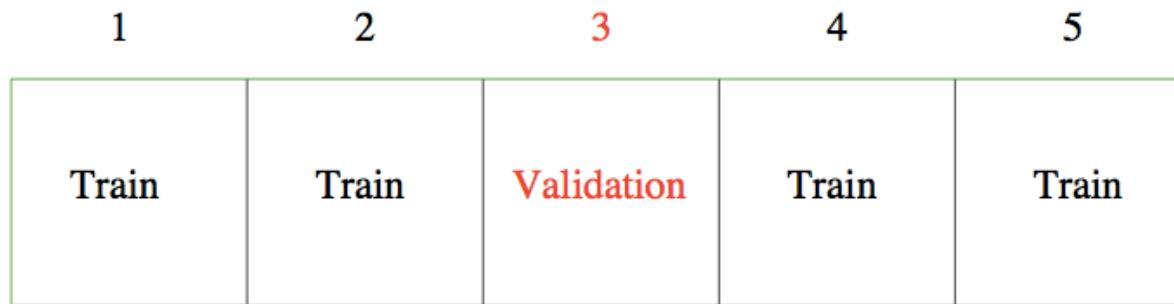


Training Data

Testing Data

Kỹ thuật đánh giá chéo K-fold

Ví dụ 5-fold

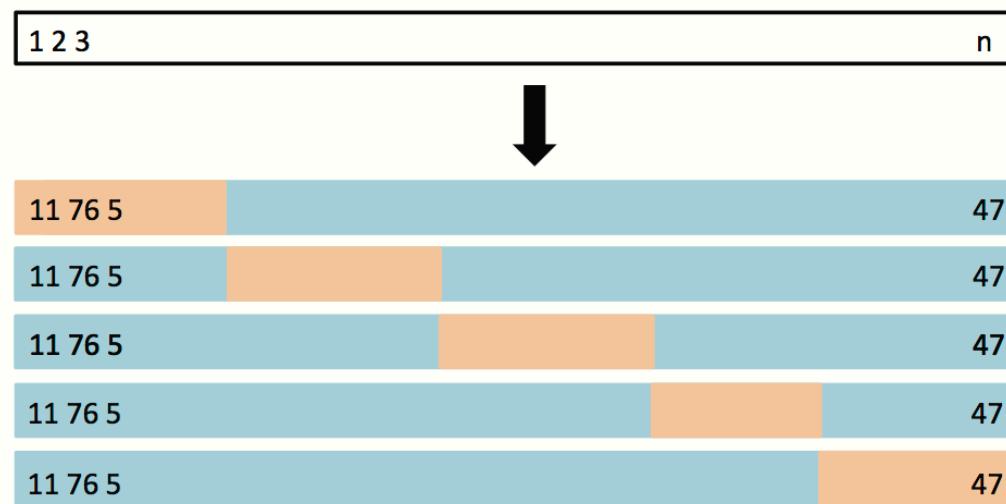


$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{f}^{-\kappa(i)}(x_i))$$

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

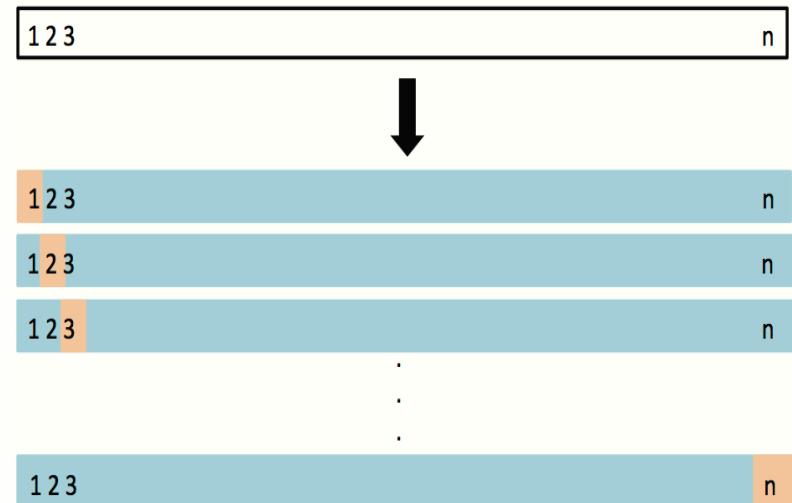
Kỹ thuật đánh giá chéo

5-fold và 10-fold thường được ưa dùng (lỗi bias cao, phương sai thấp)



Kỹ thuật đánh giá chéo

- Khi k bằng với số lượng phần tử N trong tập *training* ban đầu, tức mỗi tập con có đúng 1 phần tử, ta gọi kỹ thuật này là **leave-one-out**.
(lỗi bias thấp, phương sai cao)



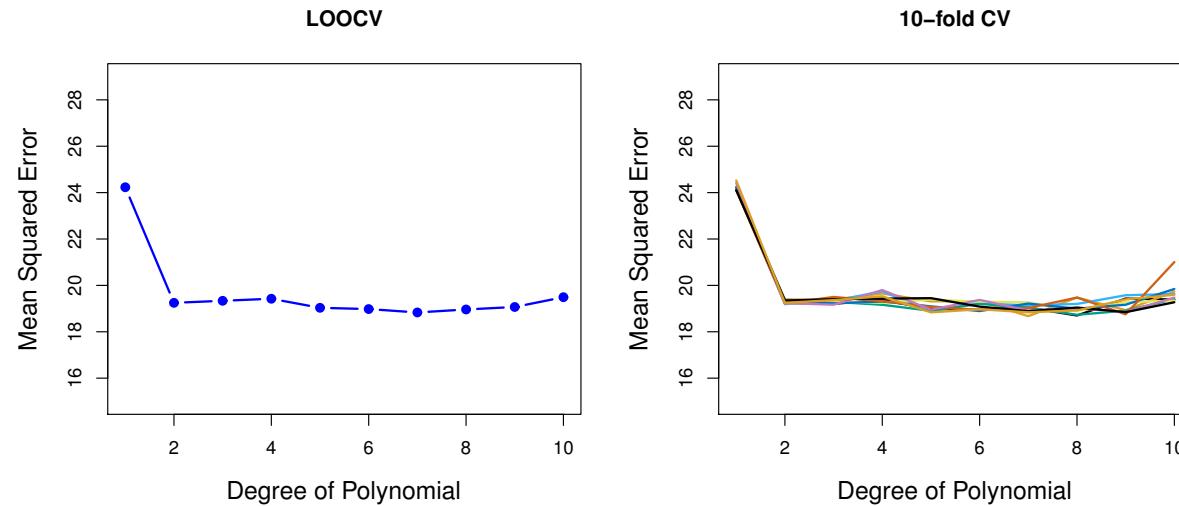
Auto Data: LOOCV vs. K-fold CV

Hình trái: Sai số LOOCV

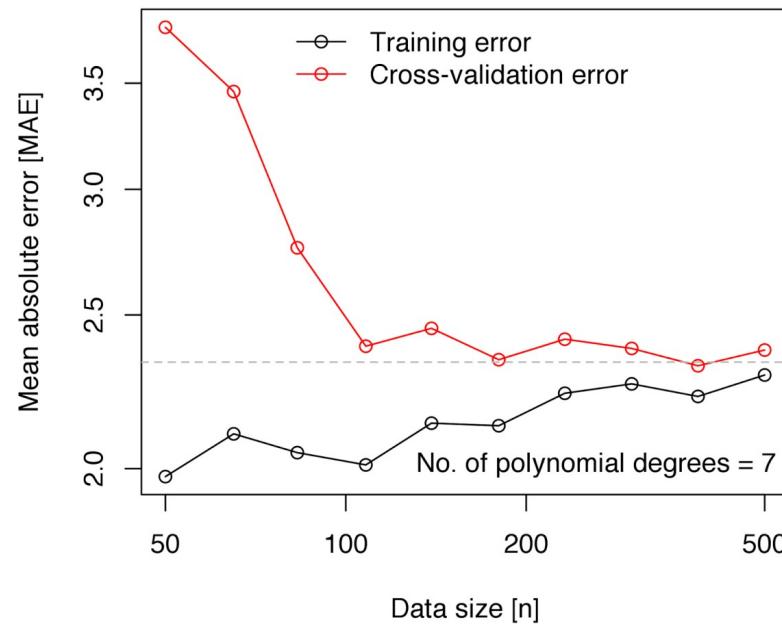
Hình phải: 10-fold CV được chạy nhiều lần, đồ thị biểu diễn sai khác nhau về lỗi CV

LOOCV là trường hợp đặc biệt của k-fold, khi k = N

Cả hai đều ổn định, tuy nhiên LOOCV mất nhiều thời gian tính toán hơn!



Kỹ thuật đánh giá chéo



Ta cần thêm biến (mô hình mới) hoặc thêm dữ liệu?

Kỹ thuật đánh giá chéo

- Nhược điểm lớn của *cross-validation* là số lượng *training runs* tỉ lệ thuận với k . Trong các bài toán Machine Learning, lượng tham số cần xác định thường lớn và khoảng giá trị của mỗi tham số cũng rộng.
- Vậy việc chỉ xây dựng một mô hình thôi đã rất phức tạp.

Câu hỏi?

Nội dung

1. Giới thiệu mô hình hồi quy
2. Overfitting, kỹ thuật đánh giá chéo
3. Phân tích dữ liệu với R
4. Hồi quy tuyến tính
5. Hồi quy phi tuyến
6. Real-life problem

Phân tích dữ liệu bằng R

R

- R và R-studio
- Gói caret

Nội dung

1. Giới thiệu mô hình hồi quy
2. Overfitting, kỹ thuật đánh giá chéo
3. Phân tích dữ liệu với R
4. Hồi quy tuyến tính
5. Hồi quy phi tuyến
6. Real-life problem

Hồi quy tuyến tính

- *Hồi quy tuyến tính*: là phương pháp học máy có giám sát đơn giản, được sử dụng để dự đoán giá trị biến đầu ra dạng số (định lượng)
 - Nhiều phương pháp học máy là dạng tổng quát hóa của hồi quy tuyến tính
 - Là ví dụ để minh họa các khái niệm quan trọng trong bài toán học máy có giám sát

Hồi quy tuyến tính

- Tại sao dùng hồi quy tuyến tính?
 - Mỗi quan hệ tuyến tính: là sự biến đổi tuân theo quy luật hàm bậc nhất
 - Tìm một mô hình (phương trình) để mô tả một mối liên quan giữa X và Y
 - Ta có thể biến đổi các biến đầu vào để tạo ra mối quan hệ tuyến tính
 - Diễn giải các mối quan hệ giữa biến đầu vào và đầu ra - sử dụng cho bài toán suy diễn



Hồi quy tuyến tính đơn giản

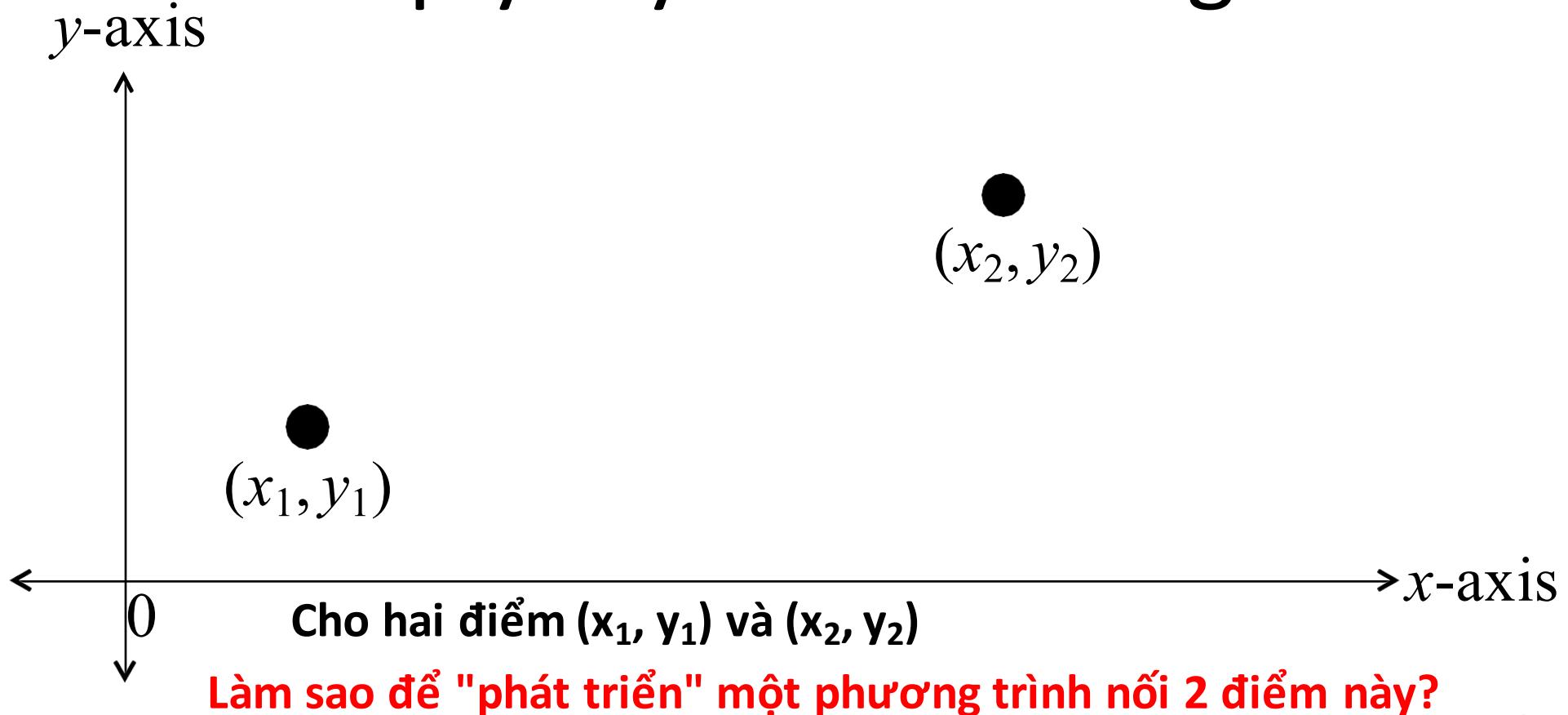
- Biến đầu ra Y và biến đầu vào X có mối quan hệ tuyến tính giữa X và Y như sau:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

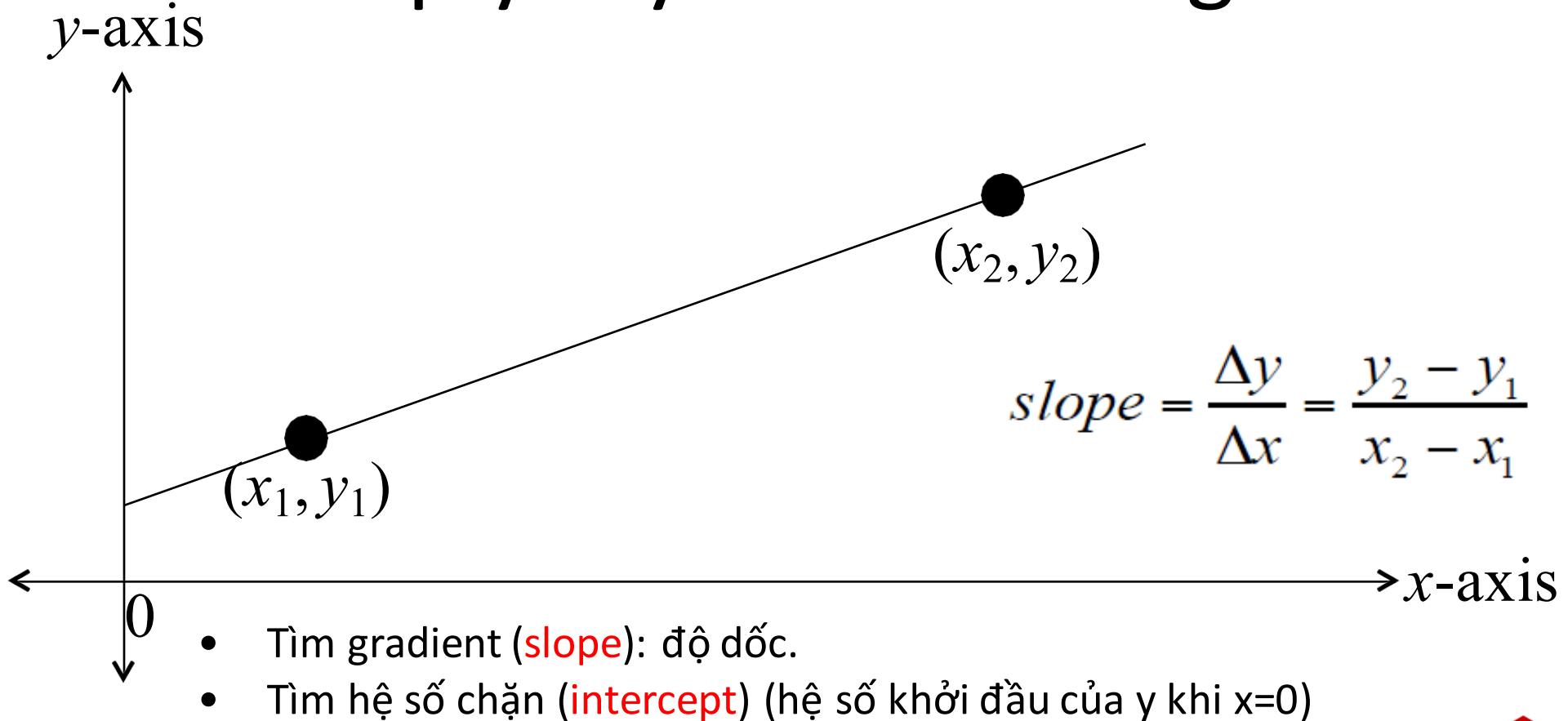
- Các tham số của mô hình:

β_0 intercept hệ số chẵn (khi các $x_i=0$)
 β_1 slope độ dốc

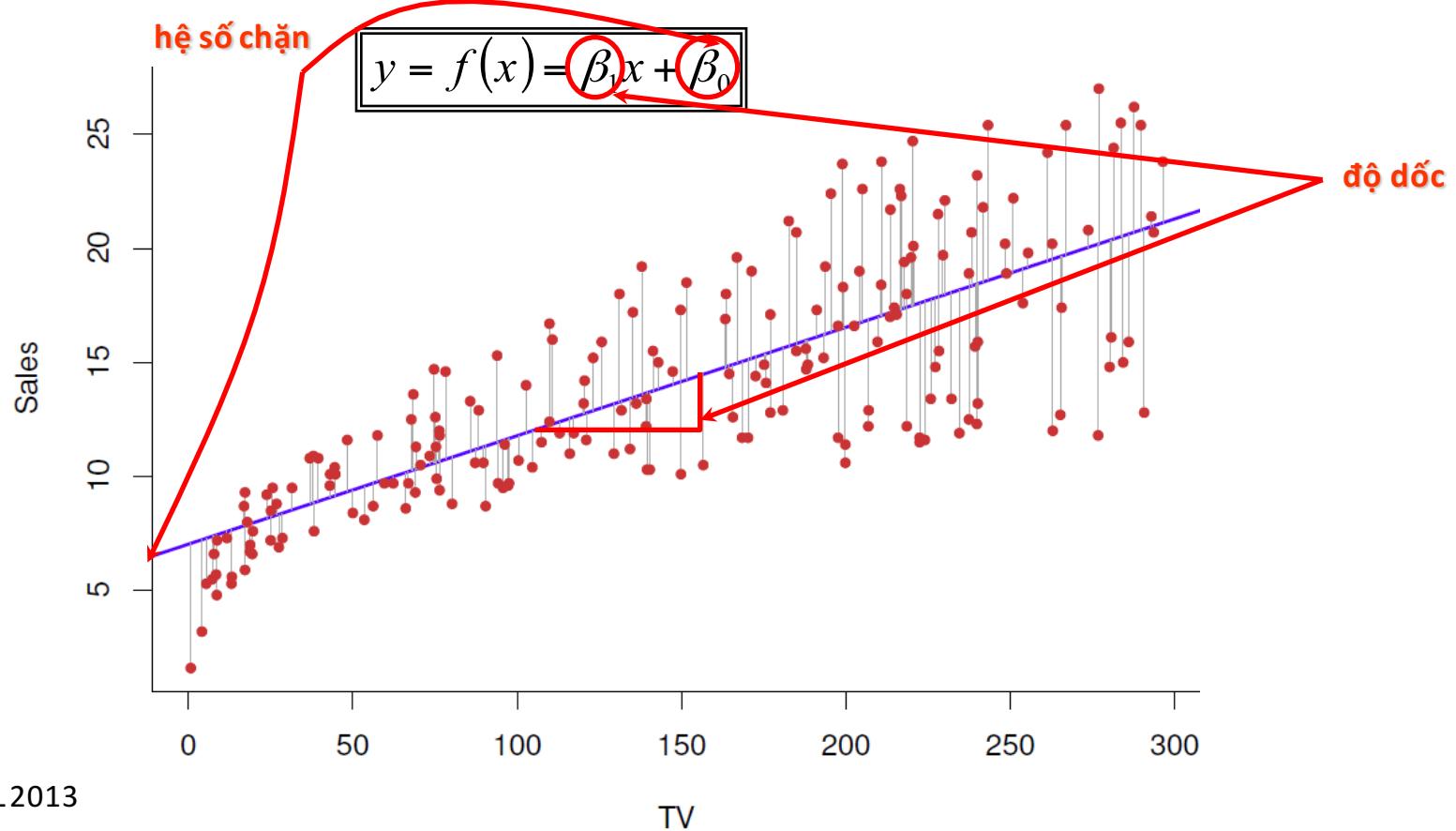
Hồi quy tuyến tính đơn giản



Hồi quy tuyến tính đơn giản



Hồi quy tuyến tính đơn giản



Hồi quy tuyến tính đơn giản

- β_0 và β_1 chưa biết \rightarrow Ta ước tính giá trị của chúng từ dữ liệu đầu vào

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- Lấy $\hat{\beta}_0, \hat{\beta}_1$ sao cho mô hình đạt “xấp xỉ tốt nhất” (“good fit”) đối với tập huấn luyện

$$Y^{(i)} \approx \hat{\beta}_0 + \hat{\beta}_1 X^{(i)}, \quad i = 1, \dots, n$$

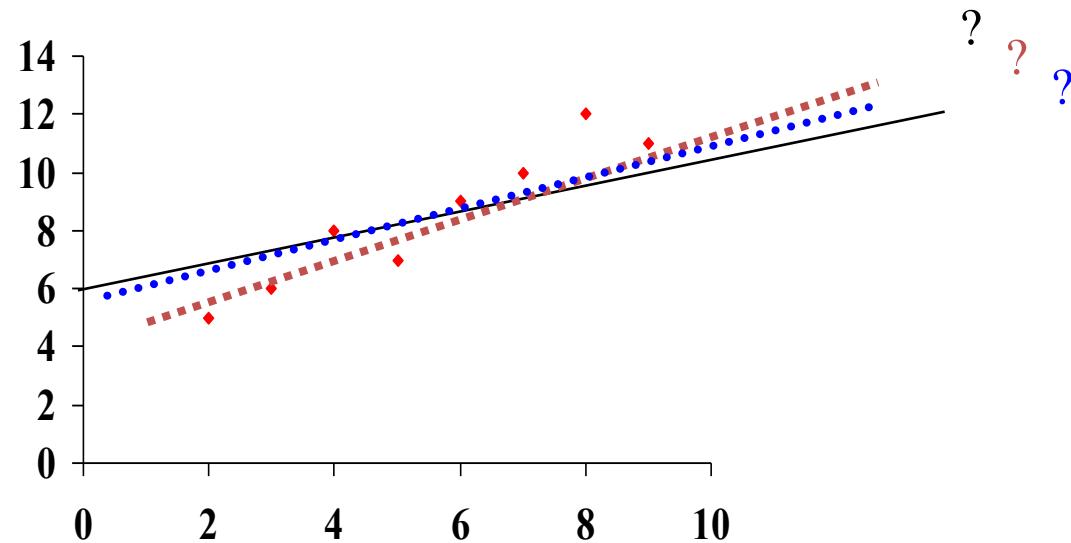
Các giả định

- Mỗi liên quan giữa X và Y là tuyến tính (linear) về *tham số*
- X không có sai số ngẫu nhiên
- Giá trị của Y độc lập với nhau (vd, Y_1 không liên quan với Y_2) ;
- Sai số ngẫu nhiên (ε): phân bố chuẩn, trung bình 0, phương sai bất biến

$$\varepsilon \sim N(0, \sigma^2)$$

Đường thẳng phù hợp nhất

Cho tập dữ liệu đầu vào, ta cần tìm cách tính toán các tham số của phương trình đường thẳng



Bình phương nhỏ nhất

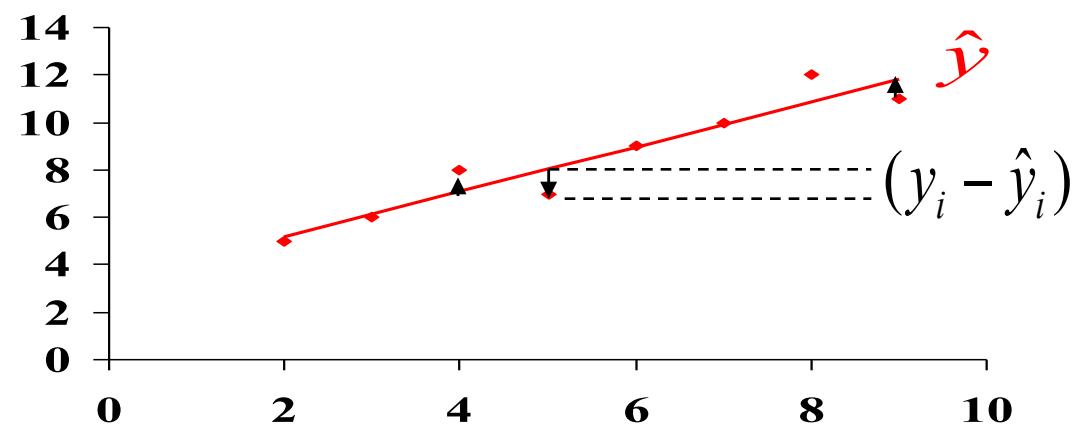
- Thông thường, để đánh giá độ phù hợp của mô hình từ dữ liệu quan sát ta sử dụng phương pháp *bình phương nhỏ nhất* (*least squares*)
- Lỗi bình phương trung bình (Mean squared error):

$$MSE = \frac{1}{n} \sum_{i=1}^n \left(Y^{(i)} - \hat{Y}^{(i)} \right)^2$$

Đường thẳng phù hợp nhất

Rất hiếm để có 1 đường thẳng khớp chính xác với dữ liệu, do vậy luôn tồn tại lỗi gắn liền với đường thẳng

Đường thẳng phù hợp nhất là đường giảm thiểu độ dao động của các lỗi này



Phần dư (lỗi)

Biểu thức $(y_i - \hat{y})$ được gọi là lỗi hoặc *phần dư*

$$\varepsilon_i = (y_i - \hat{y})$$

Đường thẳng phù hợp nhất tìm thấy khi tổng bình phương lỗi là nhỏ nhất

$$SSE = \sum_{i=1}^n (y_i - \hat{y})^2$$

Ước lượng tham số

- Các ước số $\hat{\beta}_0, \hat{\beta}_1$ tính được bằng cách cực tiểu hóa MSE

$$\min_{(\hat{\beta}_0, \hat{\beta}_1)} \left[\frac{1}{n} \sum_{i=1}^n \left(Y^{(i)} - (\hat{\beta}_0 + \hat{\beta}_1 X^{(i)}) \right)^2 \right]$$

- Hệ số chẵn của đường thẳng $\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$

trong đó: $SS_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ và $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$

Ước lượng tham số

Hệ số chặn của đường thẳng

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

trong đó

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Hồi quy tuyến tính đơn giản

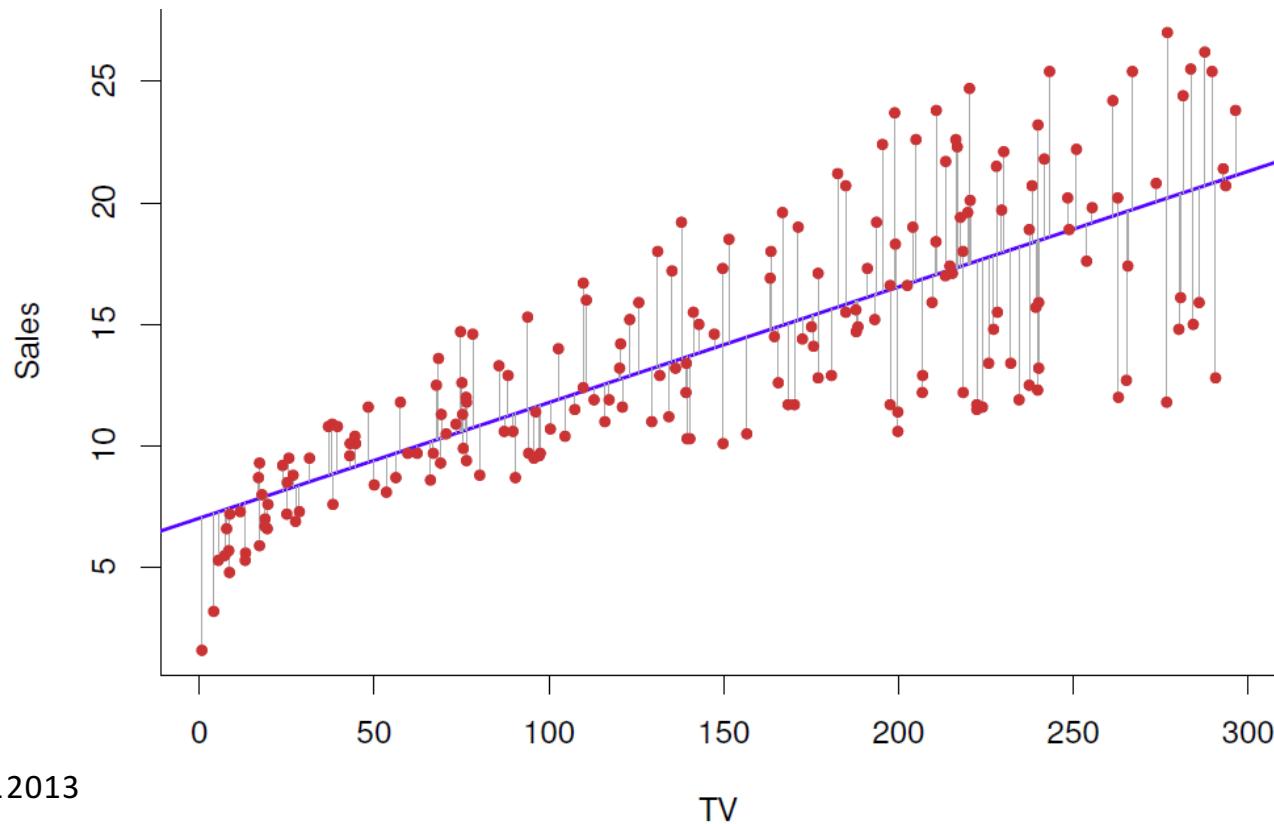
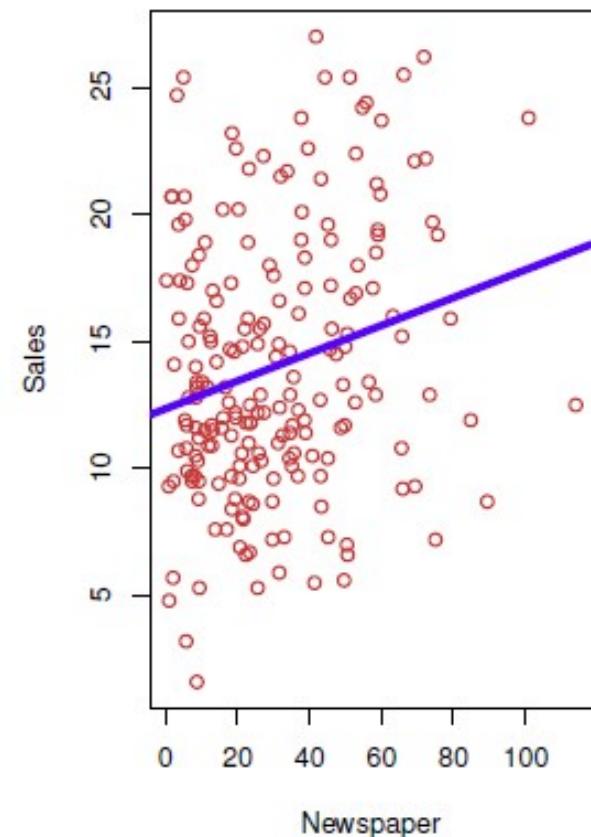
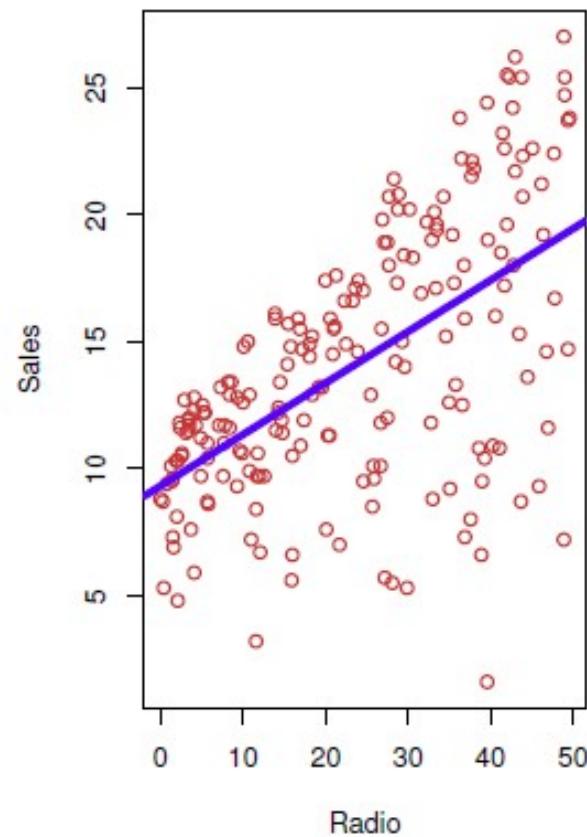
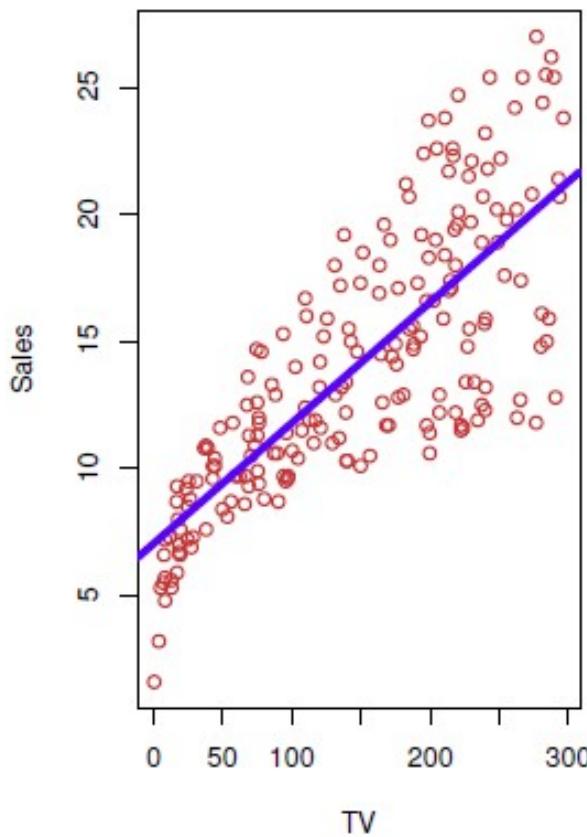


Figure 3.1 , ISL2013

Hồi quy tuyến tính đơn giản



Phương pháp đánh giá

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2}; MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - \hat{Y}_i|$$

và $R^2 = 1 - \sum_{i=1}^N (Y_i - \hat{Y}_i)^2 / \sum_{i=1}^N (Y_i - \bar{Y}_i)^2$.

Ví dụ

X kilograms	Y cost \$
----------------	--------------

17	132
21	150
35	160
39	162
50	149
65	170

$$\bar{x} = 37.83$$

$$\bar{y} = 153.83$$

$$SS_{xy} = 891.83$$

$$SS_x = 1612.83$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{891.83}{1612.83} = 0.533$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 153.83 - 0.533 \times 37.83 = 132.91$$

phương trình tìm được là

$$Y = 132.91 + 0.533 * X$$

R

Residuals:

	1	2	3	4	5	6
	-10.313	5.475	7.733	7.522	-11.561	1.145

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	132.9130	10.1079	13.149	0.000193 ***
X	0.5530	0.2451	2.256	0.087095 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

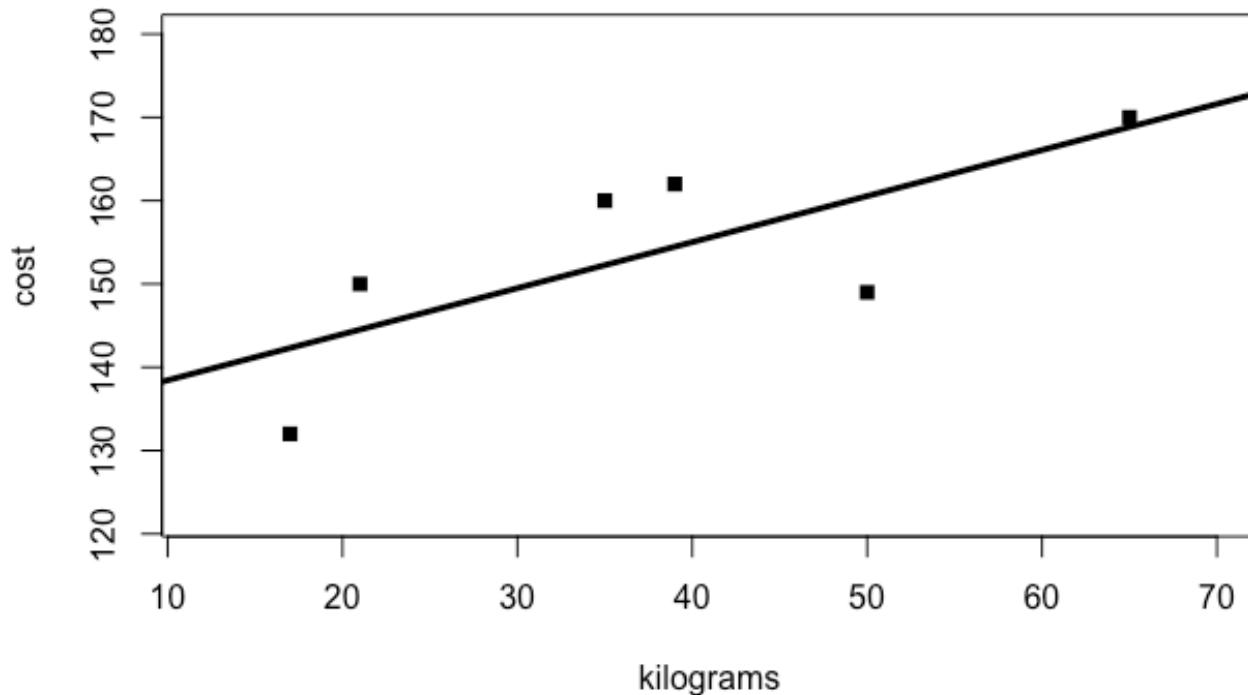
Residual standard error: 9.845 on 4 degrees of freedom
Multiple R-squared: 0.5599, Adjusted R-squared: 0.4498
F-statistic: 5.088 on 1 and 4 DF, p-value: 0.08709

```
X<-c(17, 21, 35, 39, 50, 65)
Y<-c(132, 150, 160, 162, 149, 170)
model=lm(Y ~ X)
plot(X, Y, xlim=c(min(X)-5, max(X)+5), ylim=c(min(Y)-10,
max(Y)+10), xlab="kilograms", ylab="cost",
pch=15)abline(model, lwd=3)
Summary(model)
```



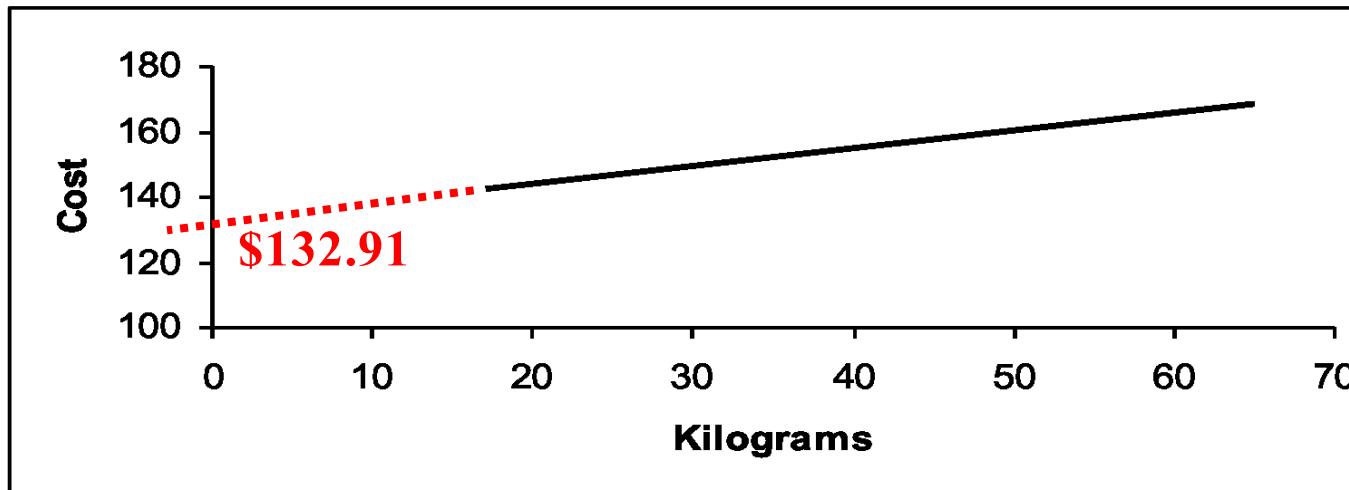
Diễn giải tham số

Trong ví dụ trước, tham số ước lượng $\hat{\beta}_1$ của độ dốc là 0.553. Điều này có nghĩa là khi thay đổi 1 kg của X, giá của Y thay đổi 0.553 \$



Diễn giải tham số

$\hat{\beta}_0$ là hệ số chặn của Y. Nghĩa là, điểm mà đường thẳng cắt trục tung Y. Trong ví dụ này là \$132.91



Đây là giá trị của Y khi X = 0

Dữ liệu phân tích: Boston

- **Boston data:** liên quan đến giá nhà đất
- Các biến số
 - **crim:** tỉ lệ tội phạm của thị trấn
 - **zn:** tỉ lệ khu đất có diện tích trên 25,000 feet vuông
 - **indus:** tỉ lệ doanh nghiệp tương đối lớn
 - **chas:** gần sông Charles (1=yes, 0=no)
 - **nos:** nồng độ nitric oxides (parts/10 triệu)
 - **rm:** số phòng trung bình mỗi nhà
 - **age:** tỉ lệ căn hộ (unit) xây trước 1940
 - **dis:** khoảng cách đến các trung tâm kĩ nghệ (tìm việc làm)

Dữ liệu phân tích: Boston

- **Boston data:** liên quan đến giá nhà đất
- Các biến số
 - **rad:** chỉ số gần xa lộ radial
 - **tax:** tỉ suất thuế tinh trên \$10,000
 - **ptratio:** tỉ số học trò trên giáo viên của thị trấn
 - **black:** chỉ số về số người da đen trong thị trấn ($Bk - 0.63)^2$
 - **Istat:** tỉ lệ dân số thành phần kinh tế thấp
 - **medv:** trị giá nhà (\$1000)

▲	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	black	lstat	medv
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2
16	0.62739	0.0	8.14	0	0.5380	5.834	56.5	4.4986	4	307	21.0	395.62	8.47	19.9
17	1.05393	0.0	8.14	0	0.5380	5.935	29.3	4.4986	4	307	21.0	386.85	6.58	23.1
18	0.78420	0.0	8.14	0	0.5380	5.990	81.7	4.2579	4	307	21.0	386.75	14.67	17.5
19	0.80271	0.0	8.14	0	0.5380	5.456	36.6	3.7965	4	307	21.0	288.99	11.69	20.2
20	0.72580	0.0	8.14	0	0.5380	5.727	69.5	3.7965	4	307	21.0	390.95	11.28	18.2

Ước tính bằng R

- Chúng ta muốn ước tính mối liên quan giữa số phòng (rm) và giá căn nhà
- Mô hình hồi qui tuyến tính:

$$\text{medv} = \beta_0 + \beta_1 * \text{rm} + \varepsilon$$

- R

`lm(medv ~ rm, data=Boston)`

Phân tích bằng R

```
attach(Boston)
```

```
# Phân tích hồi qui tuyến tính
```

```
m1 = lm(medv ~ rm, data= Boston)
```

```
summary(m1)
```

```
# vẽ biểu đồ
```

```
plot(medv ~ rm, pch=16)
```

```
abline(m1, col="red")
```

Phân tích bằng R

Residuals:

Min	1Q	Median	3Q	Max
-23.346	-2.547	0.090	2.986	39.433

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-34.671	2.650	-13.08	<2e-16	***
rm	9.102	0.419	21.72	<2e-16	***

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'	0.1 '
' 1					

Residual standard error: 6.616 on 504 degrees of freedom

Multiple R-squared: 0.4835, Adjusted R-squared: 0.4825

F-statistic: 471.8 on 1 and 504 DF, p-value: < 2.2e-16

Diễn giải kết quả

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-34.671	2.650	-13.08	<2e-16 ***	
rm	9.102	0.419	21.72	<2e-16 ***	

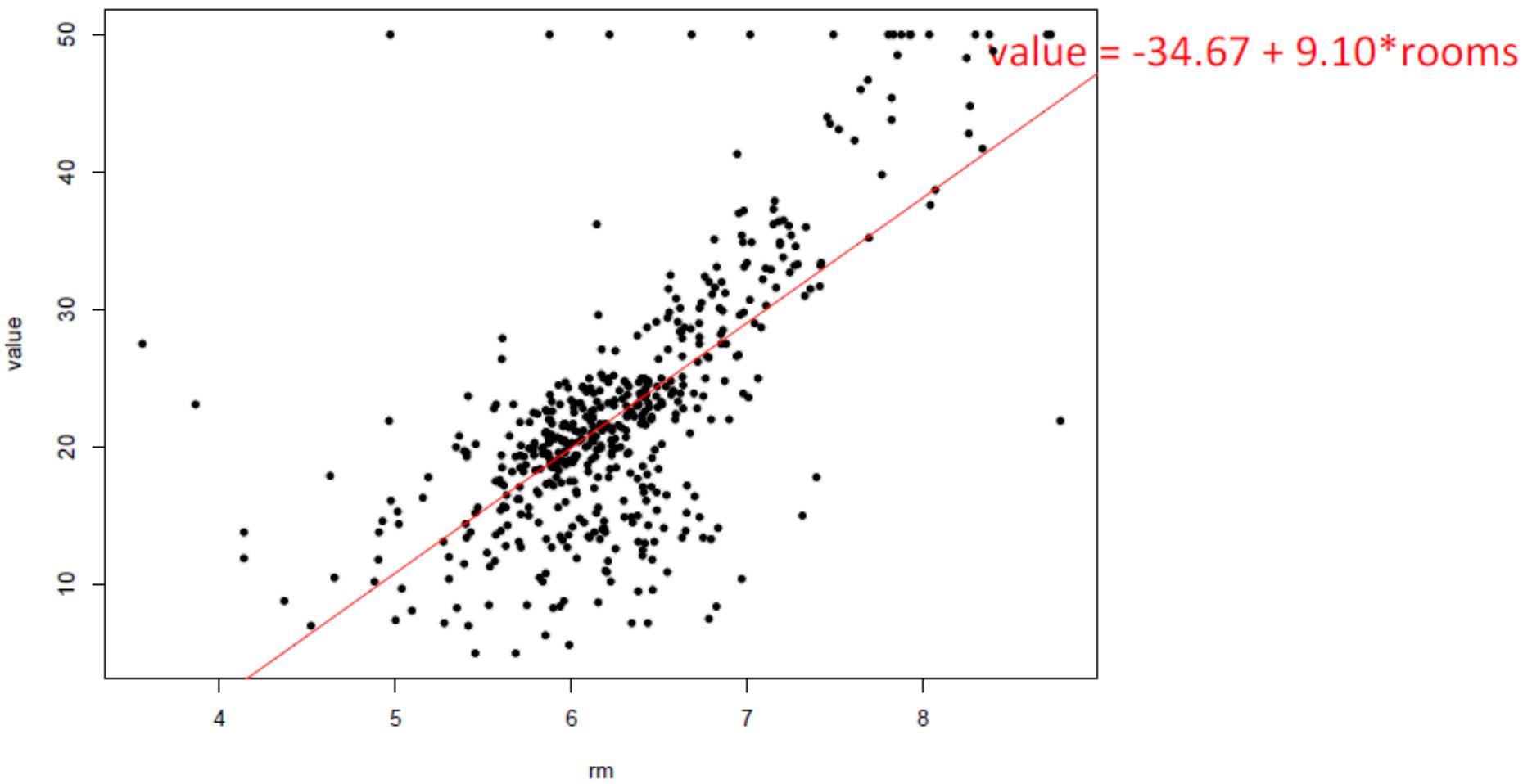
- Nhớ rằng mô hình là:

$$\text{medv} = \beta_0 + \beta_1 * \text{rm}$$

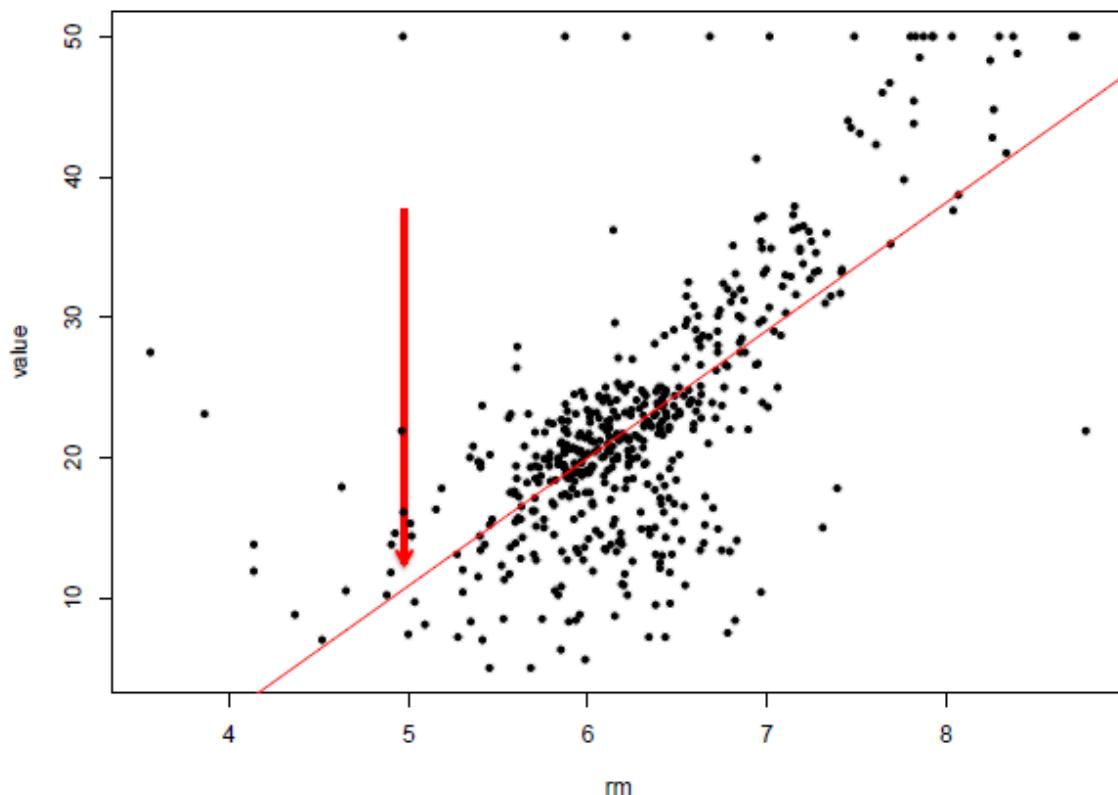
- Phương trình:

$$\text{medv} = -34.67 + 9.10 * \text{rooms}$$

- Ý nghĩa: nhà có thêm 1 phòng tăng 9100 USD cho giá trị căn nhà. Mỗi tương quan này có **ý nghĩa thống kê ($P < 0.0001$)**



Ý nghĩa của đường biểu diễn



Giá trị trung bình (kì vọng)

$$\text{medv} = -34.67 + 9.10 \cdot \text{rooms}$$

Khi room = 5,

$$\text{medv} = -34.67 + 9.10 \cdot 5 = 10.83$$

Khi room = 6

$$\text{medv} = -34.67 + 9.10 \cdot 6 = 19.93$$

Khi room = 8

$$\text{medv} = -34.67 + 9.10 \cdot 8 = 38.13$$

Hồi quy tuyến tính đa biến

- **Hồi quy tuyến tính đa biến:** mô hình có nhiều hơn 1 biến dùng để dự đoán biến đích

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_d X_d + \epsilon$$

Hồi quy tuyến tính đa biến

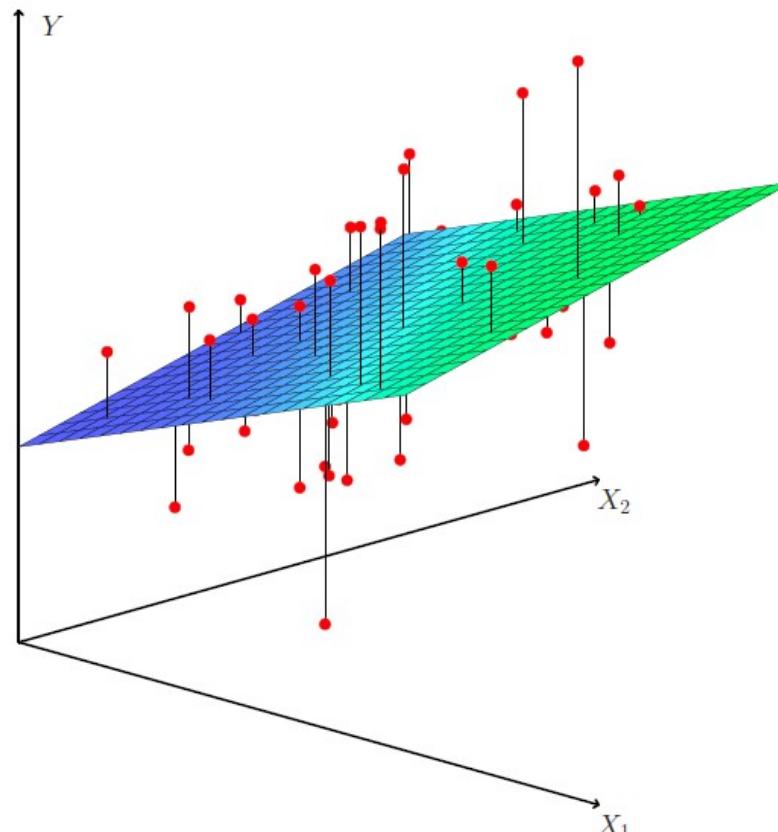


Figure 3.4 , ISL2013

Hồi quy tuyến tính đa biến

- Diễn giải hệ số β_j :

khi tăng X_j lên một đơn vị $\rightarrow Y$ sẽ tăng trung bình một lượng là β_j

	Coefficient
Intercept	2.939
TV	0.046
radio	0.189
newspaper	-0.001

Bình phương nhỏ nhất

- Tìm các ước số bằng phương pháp bình phương nhỏ nhất

$$\hat{\beta} = \arg \min_{\beta} \|Y - X^T \beta\|^2 \quad X = \begin{bmatrix} 1 & X^{(1)T} \\ & \cdots \\ 1 & X^{(n)T} \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_d \end{bmatrix}$$
$$Y = \begin{bmatrix} Y^{(1)} \\ \vdots \\ Y^{(n)} \end{bmatrix}$$

- Giải phương trình để tìm $\hat{\beta}$:

$$X^T X \hat{\beta} = X^T Y \quad \rightarrow \quad \hat{\beta} = (X^T X)^{-1} X^T Y$$

Hồi quy tuyến tính đa biến

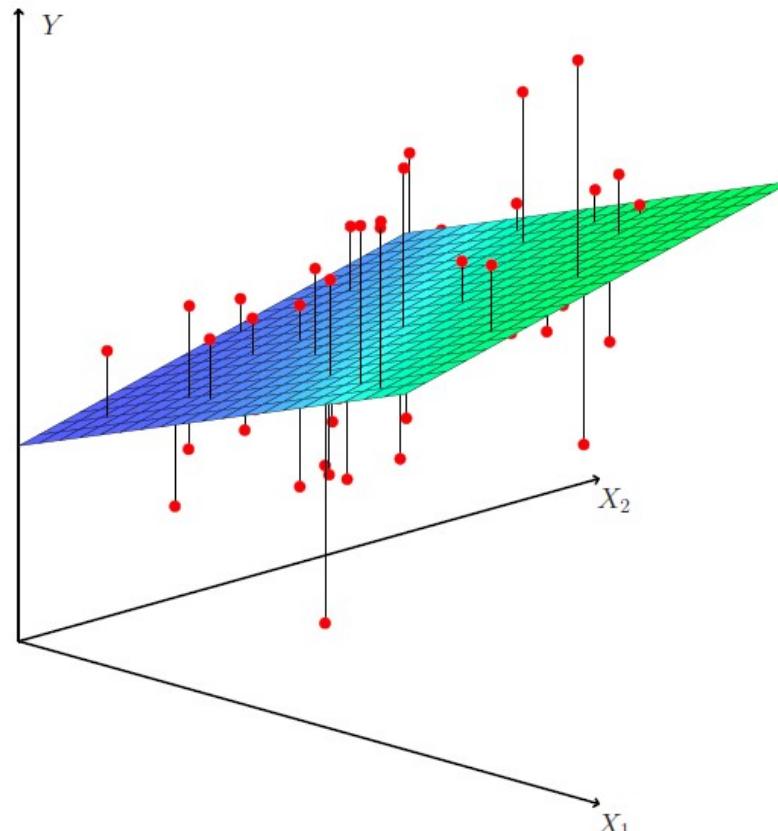


Figure 3.4 , ISL2013

Ví dụ

Cho

$$\mathbf{y} = \begin{bmatrix} 6 \\ 9 \\ 12 \\ 5 \\ 13 \\ 2 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 3 & 9 & 16 \\ 1 & 6 & 13 & 13 \\ 1 & 4 & 3 & 17 \\ 1 & 8 & 2 & 10 \\ 1 & 3 & 4 & 9 \\ 1 & 2 & 4 & 7 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

Ví dụ

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 8 & 3 & 2 \\ 9 & 13 & 3 & 2 & 4 & 4 \\ 16 & 13 & 17 & 10 & 9 & 7 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 6 & 26 & 35 & 72 \\ 26 & 138 & 153 & 315 \\ 35 & 153 & 295 & 448 \\ 72 & 315 & 448 & 944 \end{bmatrix} \quad X^T y = \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

Ví dụ

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{bmatrix} 2.59578 & -0.15375 & -0.01962 & -0.13737 \\ -0.15375 & 0.03965 & -0.00014 & -0.00144 \\ -0.01962 & -0.00014 & 0.01234 & -0.00431 \\ -0.13737 & -0.00144 & -0.00431 & 0.01406 \end{bmatrix} \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$
$$= \begin{bmatrix} 3.20975 \\ -0.07573 \\ -0.11162 \\ 0.46691 \end{bmatrix}$$

$$\hat{\beta}_0 = 3.20975 \quad \hat{\beta}_1 = -0.07573 \quad \hat{\beta}_2 = -0.11162 \quad \hat{\beta}_3 = 0.46691$$

$$\hat{y} = 3.20975 - 0.07573x_1 - 0.11162x_2 + 0.46691x_3$$

Hồi quy tuyến tính

- Ưu điểm:
 - Mô hình đơn giản, dễ hiểu
 - Dễ diễn giải hệ số hồi quy
 - Nhận được kết quả tốt khi dữ liệu quan sát nhỏ
 - Nhiều cải tiến/mở rộng
- Nhược điểm:
 - Mô hình hơi đơn giản nên khó dự đoán chính xác với dữ liệu có miền giá trị rộng
 - Khả năng ngoại suy (extrapolation) kém
 - Nhạy cảm với dữ liệu ngoại lai (outliers) – do dung phương pháp bình phương nhỏ nhất

Bài tập tại lớp

Cho bảng dữ liệu về chiều cao và cân nặng của 15 người như sau:

Chiều cao (cm)	Cân nặng (kg)	Chiều cao (cm)	Cân nặng (kg)
147	49	168	60
150	50	170	72
153	51	173	63
155	52	175	64
158	54	178	66
160	56	180	67
163	58	183	68
165	59		

Bài toán đặt ra là: liệu có thể dự đoán cân nặng của một người dựa vào chiều cao của họ không?

Câu hỏi?

Nội dung

1. Giới thiệu mô hình hồi quy
2. Overfitting, kỹ thuật đánh giá chéo
3. Phân tích dữ liệu với R
4. Hồi quy tuyến tính
5. Hồi quy phi tuyến
6. Real-life problem

Phương pháp kết hợp các mô hình (ensemble models)

Cây phân loại và hồi quy

Classification and Regression Trees

(CART)

Xây dựng cây CART thế nào?

Có 2 dạng:

1. Hồi quy

2. Phân loại (lớp)

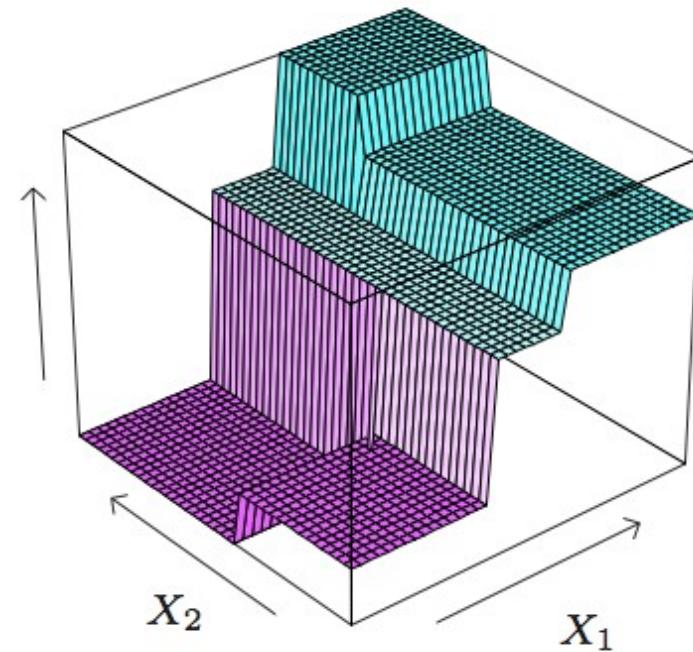
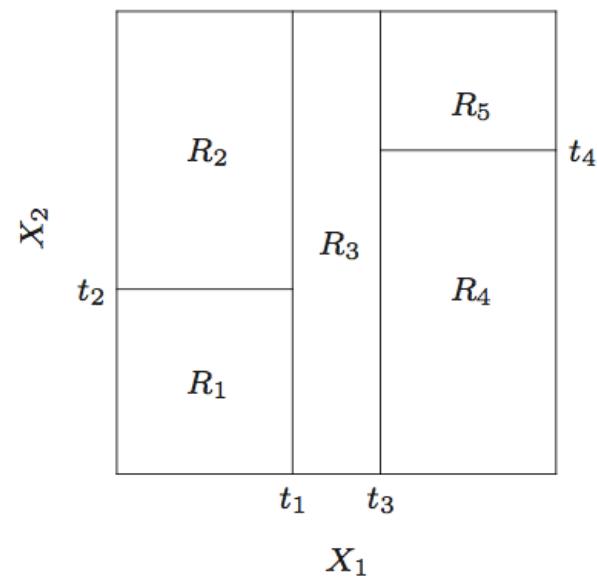
Mô hình liên tục từng đoạn(piecewise)

- Dự đoán liên tục trong mỗi vùng

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m)$$

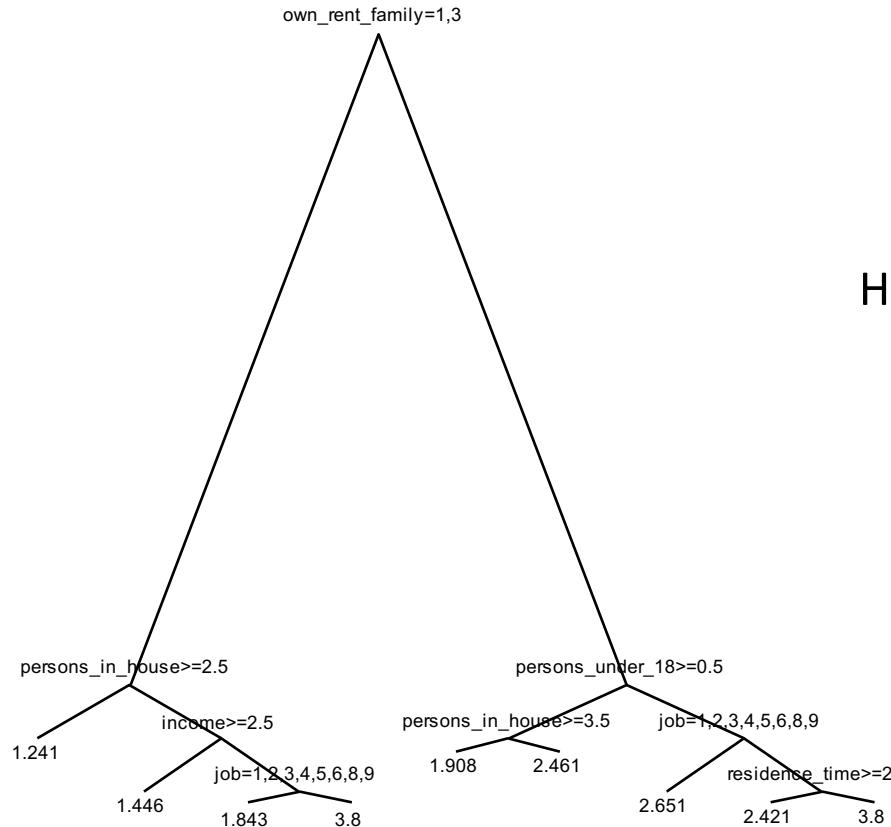
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Mô hình liên tục từng đoạn



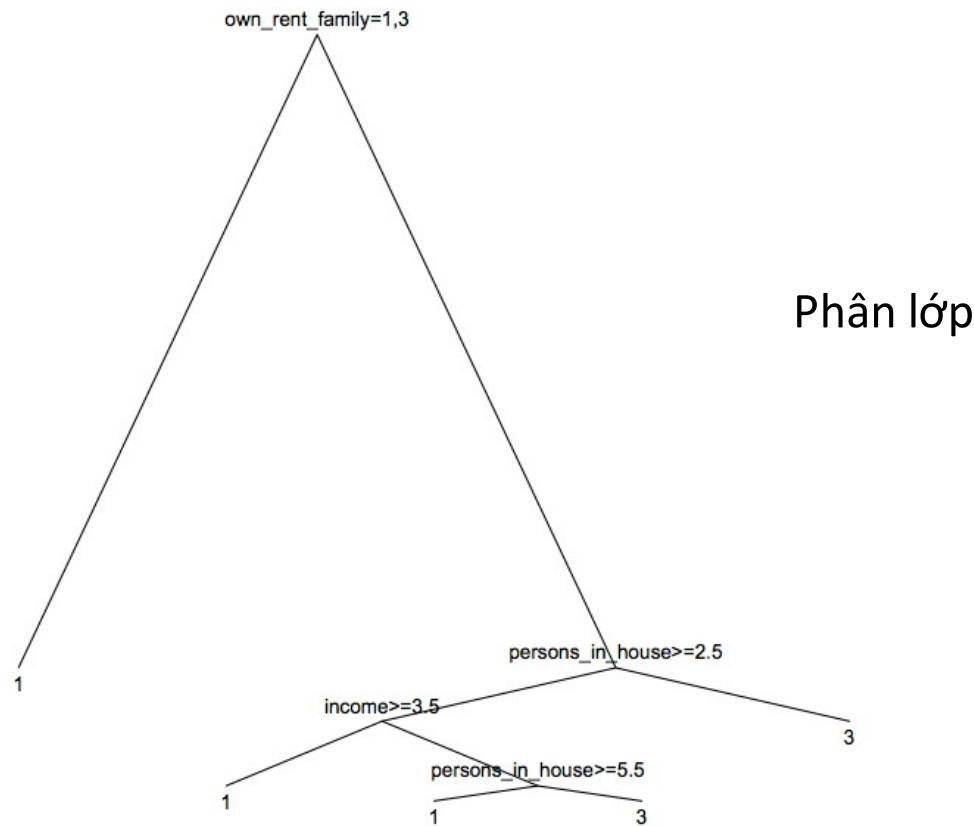
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Minh họa cây CART



Hồi quy

Minh họa cây CART



Cây hồi quy

$$f(x) = \sum_{m=1}^M c_m I(x \in R_m)$$

$$\hat{c}_m = \text{ave}(y_i | x_i \in R_m)$$

Giá trị dự đoán lưu tại lá của cây hồi quy. Nó được tính bằng giá trị trung bình của tất cả các mẫu (bản ghi) tại lá đó.

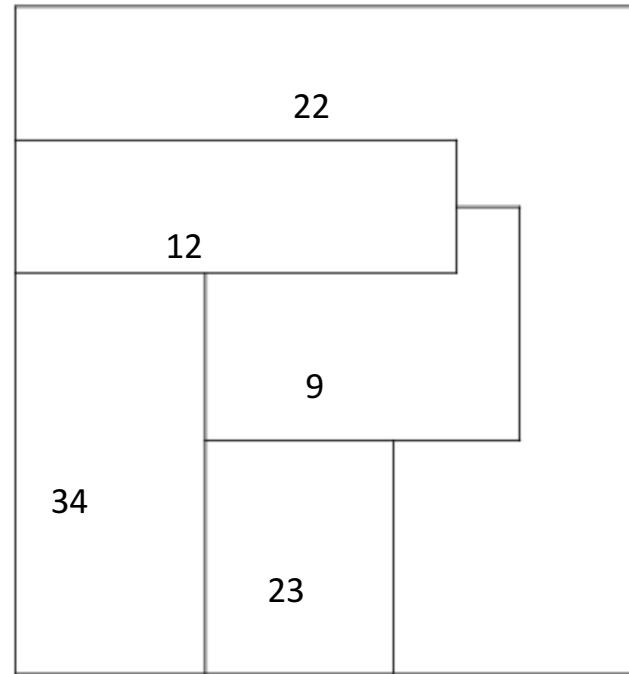
Cây hồi quy

- Giả sử ta có 2 vùng R_1 và R_2 với $\hat{Y}_1 = 10, \hat{Y}_2 = 20$
- Với các giá trị của X mà $X \in R_1$ ta sẽ có giá trị dự đoán là 10, ngược lại $X \in R_2$ ta có kết quả dự đoán là 20.

Cây hồi quy

- Cho 2 biến đầu vào và 5 vùng
- Tùy theo từng vùng của giá trị mới X ta sẽ có dự đoán 1 trong 5 giá trị cho Y.

X_2



X_1

Tách các biến X

Ta tạo ra các phân
vùng bằng cách
tách lặp đi lặp lại
một trong các biến
X thành hai vùng



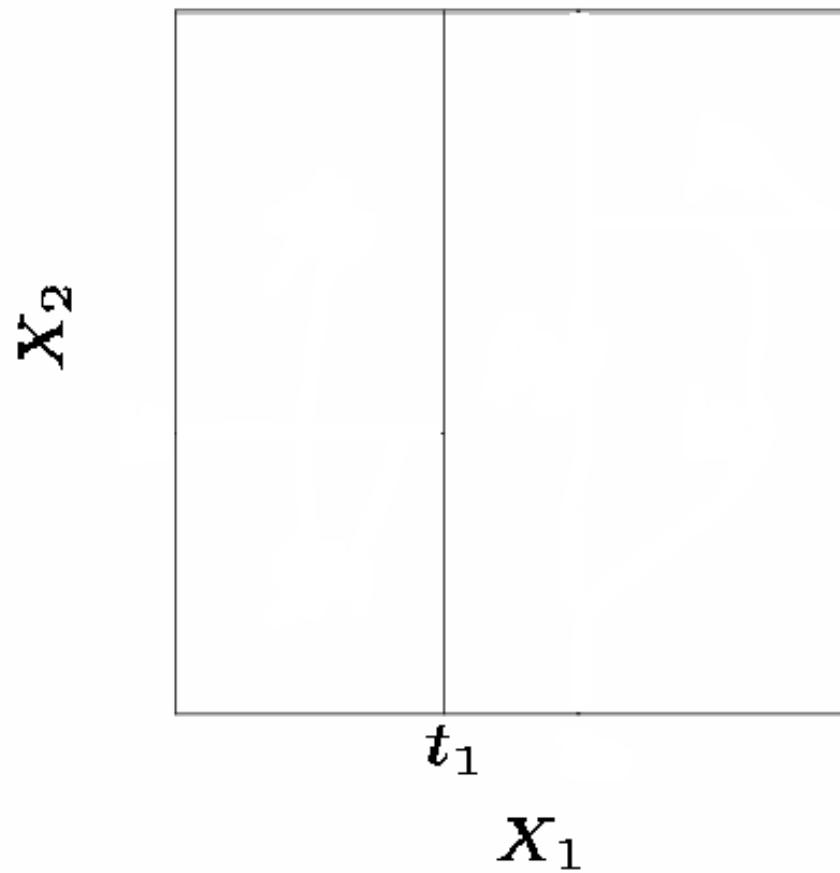
X_2

X_1



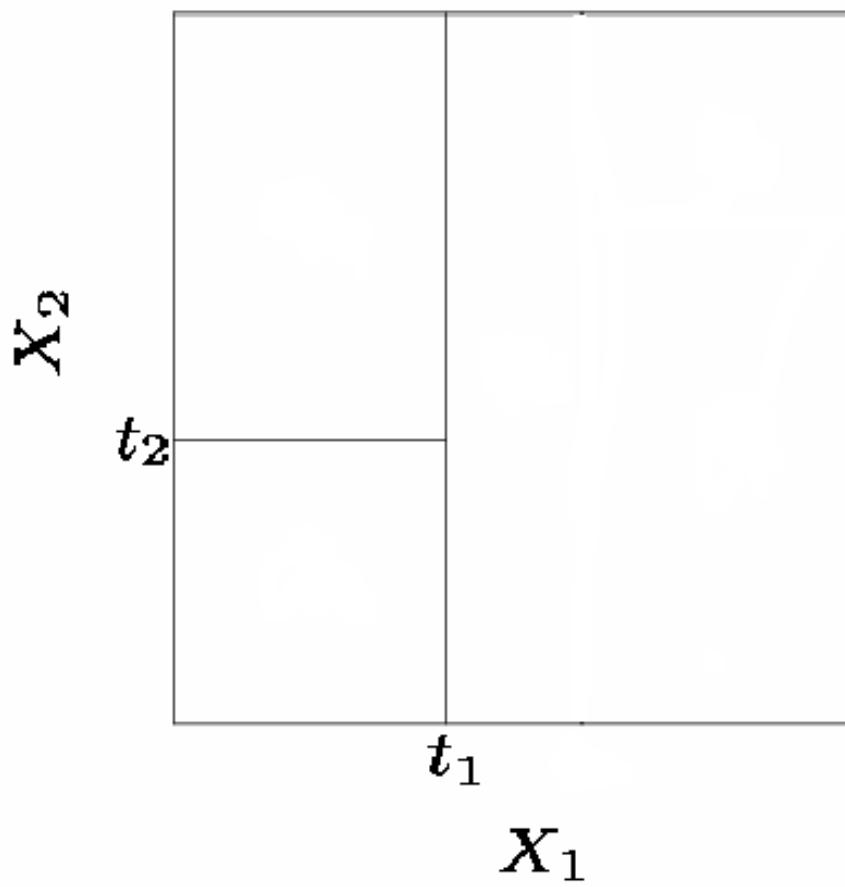
Tách các biến X

1. Đầu tiên tách
trên $X_1=t_1$



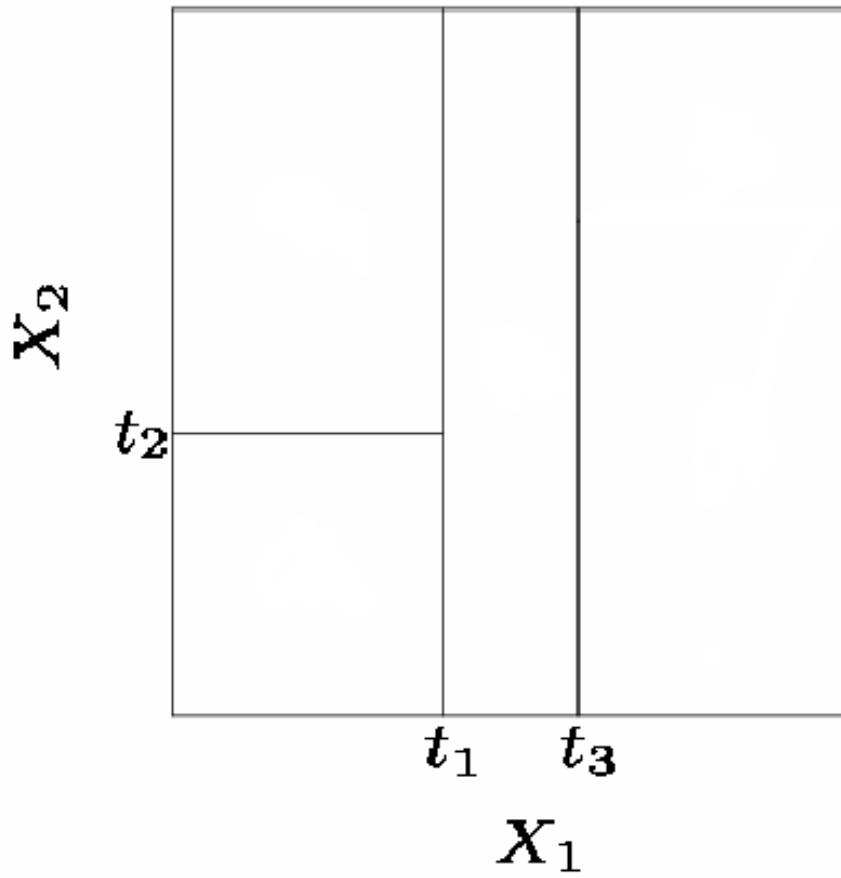
Tách các biến X

1. Đầu tiên tách
trên $X_1=t_1$
2. Nếu $X_1 < t_1$,
tách trên $X_2=t_2$



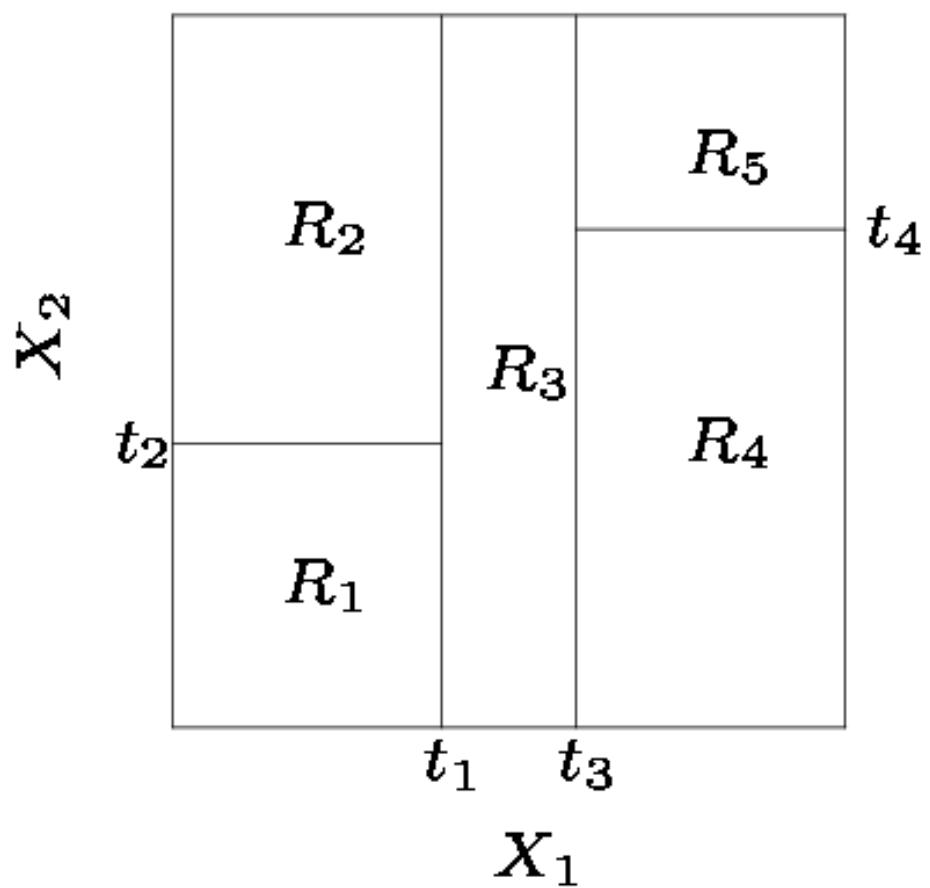
Tách các biến X

1. Đầu tiên tách
trên $X_1=t_1$
2. Nếu $X_1 < t_1$,
tách trên $X_2=t_2$
3. Nếu $X_1 > t_1$,
tách trên $X_1=t_3$

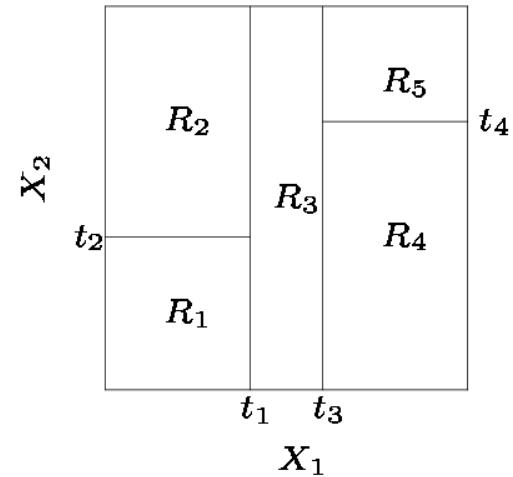
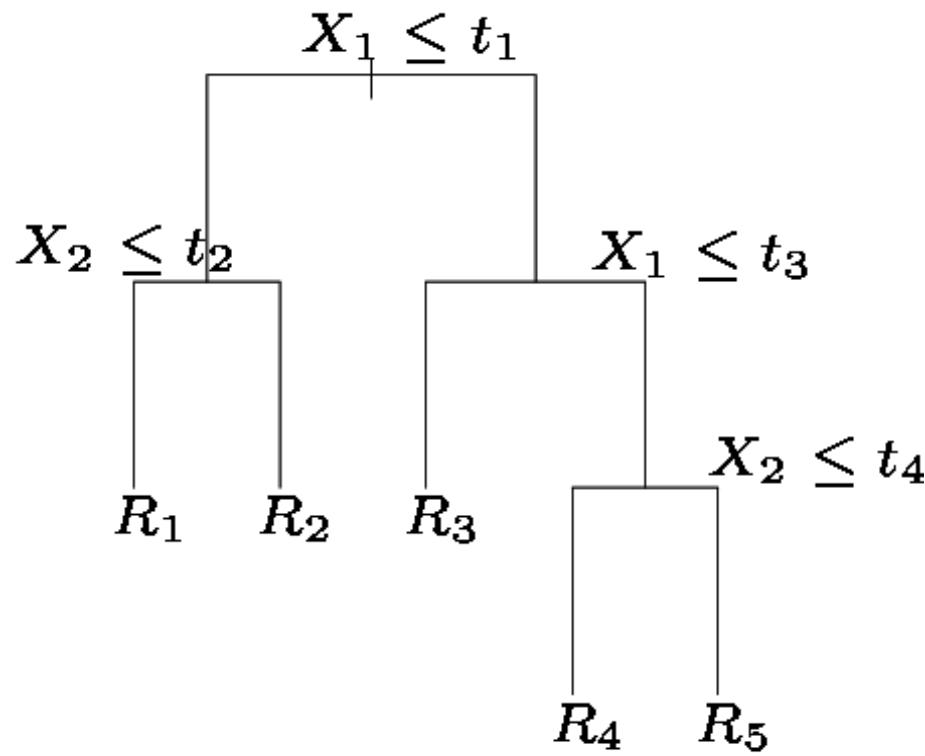


Tách các biến X

1. Đầu tiên tách
trên $X_1=t_1$
2. Nếu $X_1 < t_1$,
tách trên $X_2=t_2$
3. Nếu $X_1 > t_1$,
tách trên $X_1=t_3$
4. Nếu $X_1 > t_3$,
tách $X_2=t_4$



Tách các biến X



- Khi ta tạo các vùng theo phương pháp này, ta có thể biểu diễn chúng dùng cấu trúc cây.
- Phương pháp này dễ diễn giải mô hình dự đoán, dễ diễn giải kết quả

Giải thuật tham lam: hồi quy

- Tìm thuộc tính tách j và điểm tách s mà nó cực tiểu lỗi dự đoán

$$\min_{j, s} \left[\min_{c_1} \sum_{x_i \in R_1(j, s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j, s)} (y_i - c_2)^2 \right]$$

Ưu điểm của CART

- Dễ xử lý dữ liệu thiếu (surrogate splits)
- Mạnh trong xử lý dữ liệu chứa thông tin rác (non-informative data)
- Cho phép tự động lựa chọn thuộc tính (variable selection)
- Dễ giải thích, lý tưởng để giải thích “tại sao” đối với người ra quyết định
- Xử lý được tính tương tác cao giữa các thuộc tính

Nhược điểm của CART

- Cây không ổn định (Instability of trees)
- Thiếu tính trơn (Lack of smoothness)
- Khó nắm bắt độ cộng tính (Hard to capture additivity)

Ensemble Models

*Some characteristics of different learning methods. Key: ▲ = good,
◆ = fair, and ▼ = poor.*

Characteristic	Neural Nets	SVM	Trees	Random forest	k-NN, Kernels
Natural handling of data of “mixed” type	▼	▼	▲	▲	▼
Handling of missing values	▼	▼	▲	▲	▲
Robustness to outliers in input space	▼	▼	▲	▲	▲
Insensitive to monotone transformations of inputs	▼	▼	▲	▲	▼
Computational scalability (large N)	▼	▼	▲	▲	▼
Ability to deal with irrelevant inputs	▼	▼	▲	▲	▼
Ability to extract linear combinations of features	▲	▲	▼	▲	◆
Interpretability	▼	▼	◆	▼	▼
Predictive power	▲	▲	▼	▲	▲

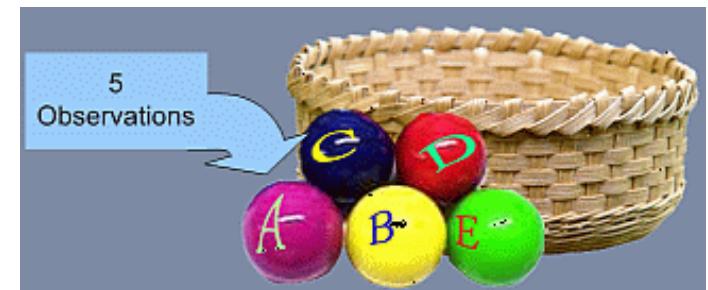
[Fernández-Delgado, Manuel](#), et al. "Do we need hundreds of classifiers to solve real world classification problems?." *The Journal of Machine Learning Research* 15.1 (2014): 3133-3181.

Kết luận của nghiên cứu trên của nhóm Manuel là phương pháp Random Forests hầu hết cho kết quả tốt nhất.

Bootstrap là gì?

- Giả sử ta có 5 quả bóng gắn nhãn A,B,C,D, E và bỏ tất cả chúng vào trong 1 cái giỏ.
- Lấy ra ngẫu nhiên 1 quả từ giỏ và ghi lại nhãn, sau đó bỏ lại quả bóng vừa bốc được vào giỏ.
- Tiếp tục lấy ra ngẫu nhiên một quả bóng và lặp lại quá trình trên cho đến khi việc lấy mẫu kết thúc. Việc lấy mẫu này gọi là lấy mẫu có hoàn lại.
- Kết quả của việc lấy mẫu như trên có thể như sau (giả sử kích thước mẫu là 10):

C, D, E, E, A, B, C, B, A, E



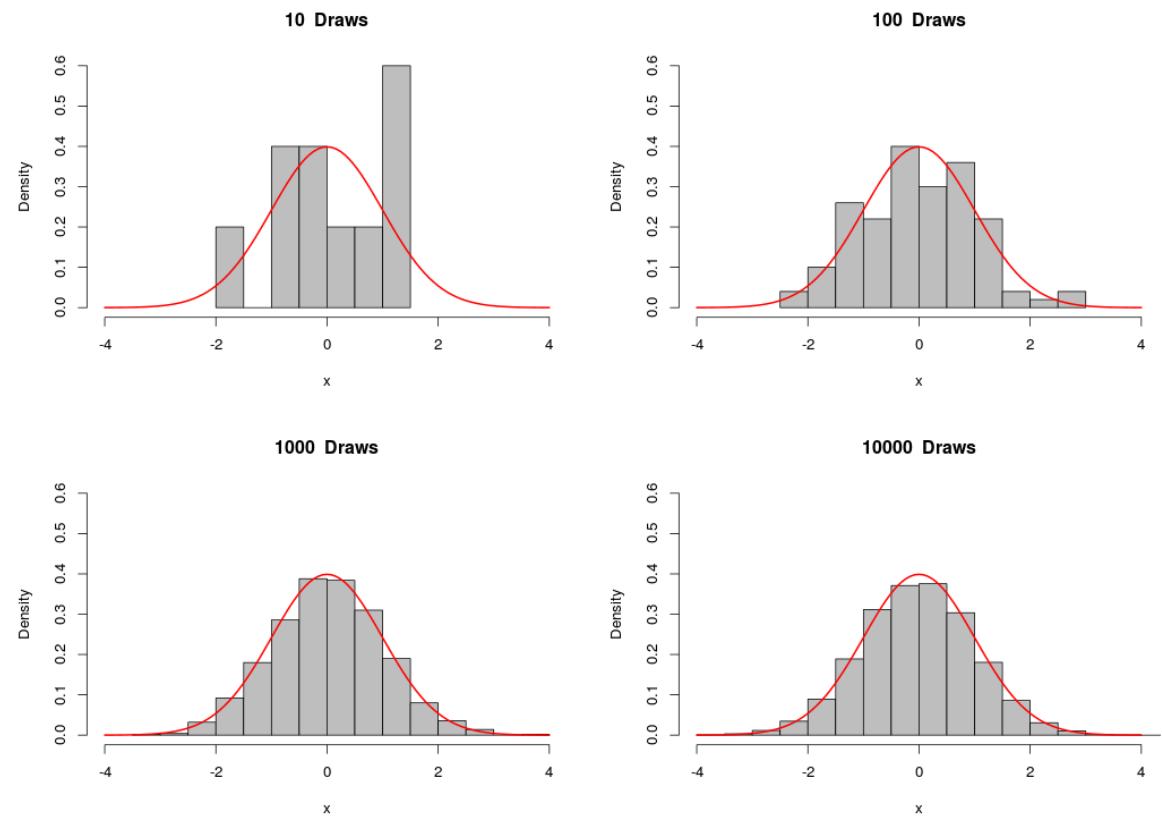
Nguồn: bis.net.vn/forums

Vietnam Institute for
Advanced Study in Mathematics



Bootstrap là gì?

- Bootstrap là phương pháp lấy mẫu có hoàn lại (sampling with replacement) -> một mẫu có thể xuất hiện nhiều lần trong một lần lấy mẫu

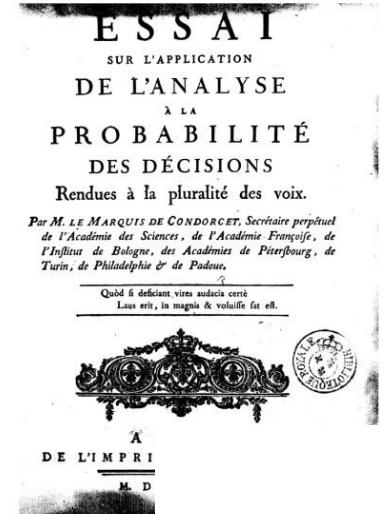


Bootstrap là gì?

- Là kỹ thuật rất quan trọng trong thống kê
- Lấy mẫu có hoàn lại từ tập dữ liệu ban đầu để tạo ra các tập dữ liệu mới

Sức mạnh của các bộ phân lớp yếu

Condorcet's Jury Theorem – Nếu p lớn hơn $1/2$ (mỗi cử tri bỏ phiếu đúng mong muốn của họ), càng thêm nhiều cử tri sẽ tăng xác suất theo quyết định số đông sẽ chính xác. Trong giới hạn, xác suất bầu chọn theo số đông tiến đến 1 khi số cử tri tăng lên.



Source gallica.bnf.fr / Bibliothèque nationale de France

Sức mạnh của các bộ phân lớp yếu

Condorcet's Jury Theorem – Nếu p lớn hơn $1/2$ (mỗi cử tri bỏ phiếu đúng mong muốn của họ), càng thêm nhiều cử tri sẽ tăng xác suất theo quyết định số đông sẽ chính xác. Trong giới hạn, xác suất bầu chọn theo số đông tiến đến 1 khi số cử tri tăng lên.



Sức mạnh của các bộ phân lớp yếu

- Việc lấy trung bình làm giảm phương sai và không làm tăng bias (bias vẫn được giữ nguyên) $\text{Var}[\bar{Y}] = \sigma^2/n$

Sức mạnh của các bộ phân lớp yếu

- Việc lấy trung bình làm giảm phương sai và không làm tăng bias (bias vẫn được giữ nguyên) $\text{Var}[\bar{Y}] = \sigma^2/n$
- Các phiếu bầu của các bộ phân lớp tương quan không trợ giúp được nhiều

THE CHOICE OF A CANDIDATE

THE NEW YORK TIMES supported Franklin D. Roosevelt for the Presidency in 1932 and again in 1936. In 1940 it will support Wendell Willkie.



Sức mạnh của các bộ phân lớp yếu

- Việc lấy trung bình làm giảm phương sai và không làm tăng bias (bias vẫn được giữ nguyên) $\text{Var}[\bar{Y}] = \sigma^2/n$
- Các phiếu bầu của các bộ phân lớp tương quan không trợ giúp được nhiều $\text{Var}[\bar{Y}] = \sigma^2/n + (\rho\sigma^2)(n-1)/n$

THE CHOICE OF A CANDIDATE

THE NEW YORK TIMES supported Franklin D. Roosevelt for the Presidency in 1932 and again in 1936. In 1940 it will support Wendell Willkie.

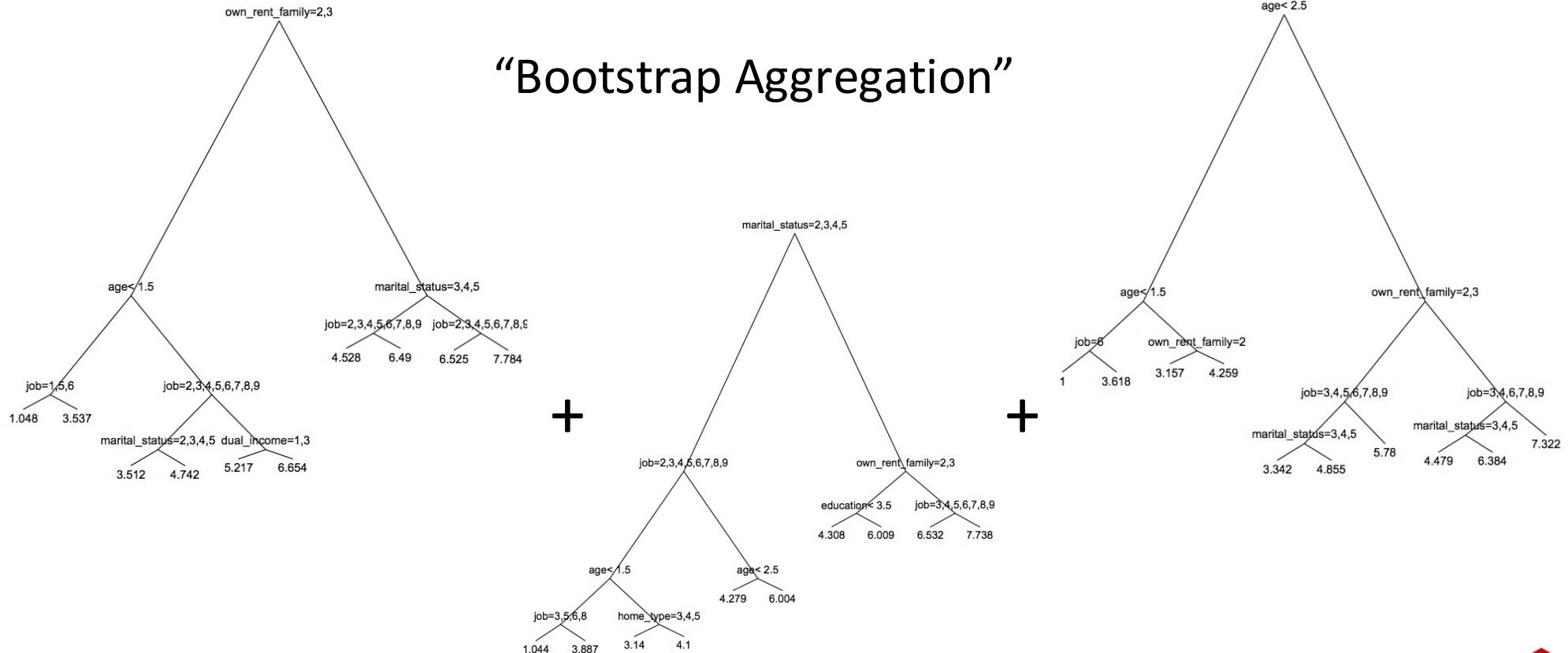
Kết hợp các bộ phân lớp

$$\alpha \times \{CART\} + (1-\alpha) \times \{LinearModel\}$$

Các phương pháp kết hợp: Bagging

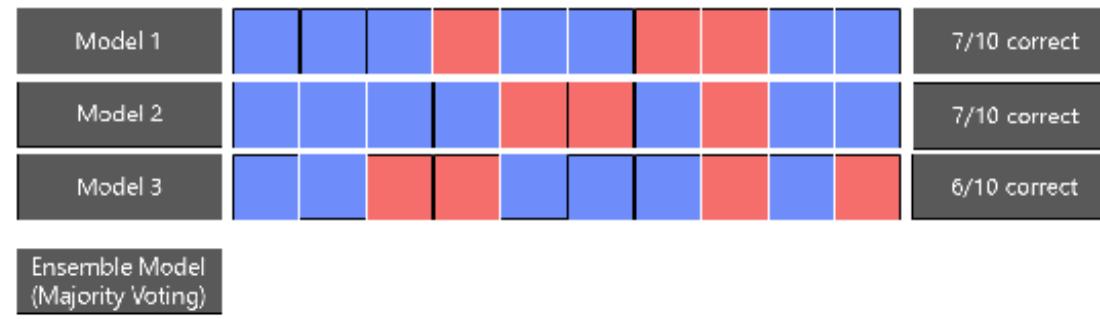
Bagging là gì?

“Bootstrap Aggregation”



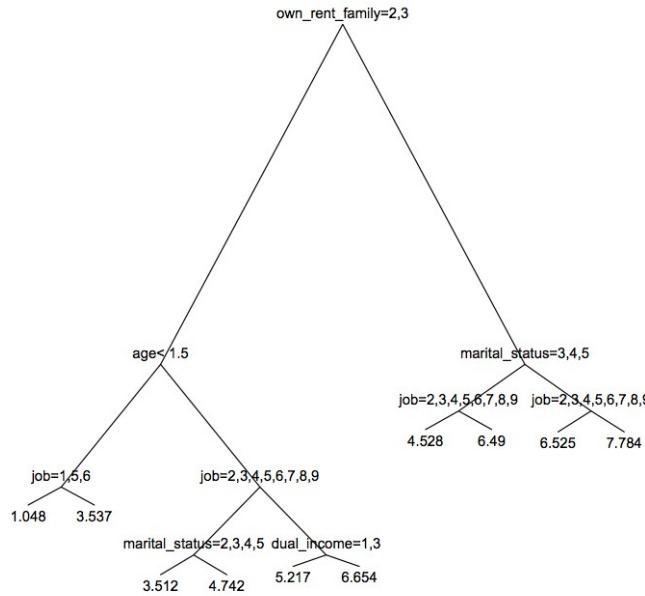
Bagging là gì?

“Bootstrap Aggregation” $\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x)$

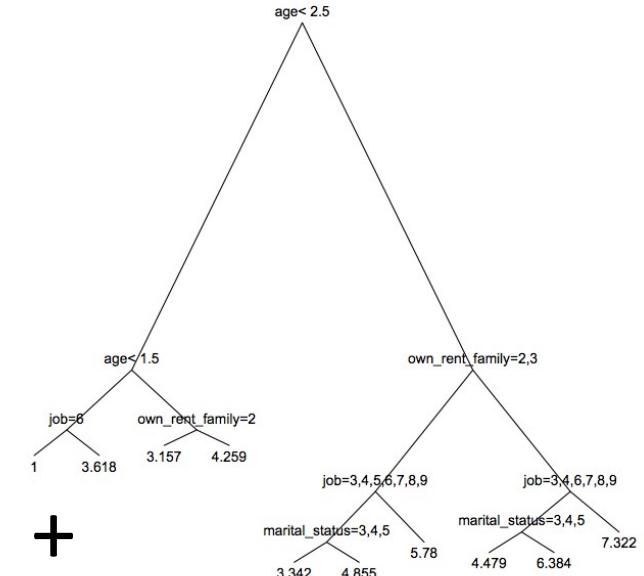
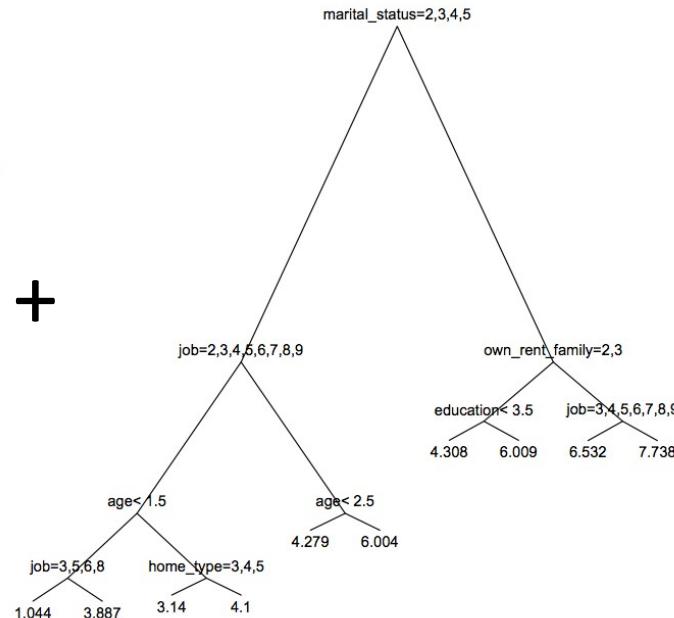


Bagging

Giai quyết được tính thiểu ổn định của CART

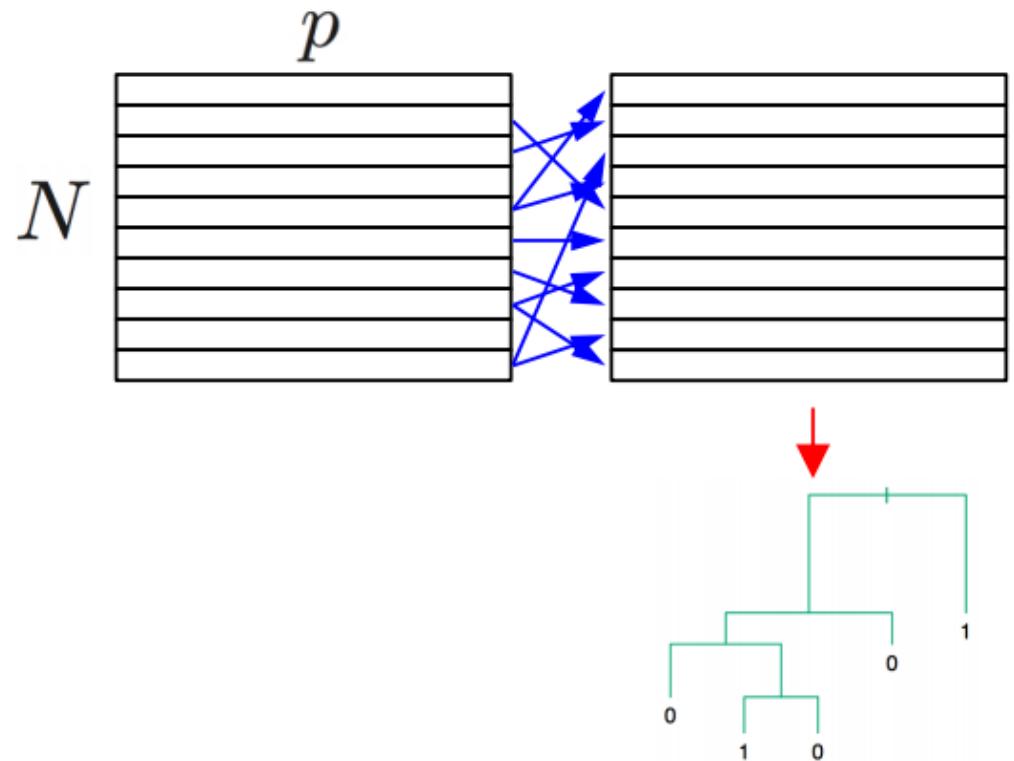


+



Bagging

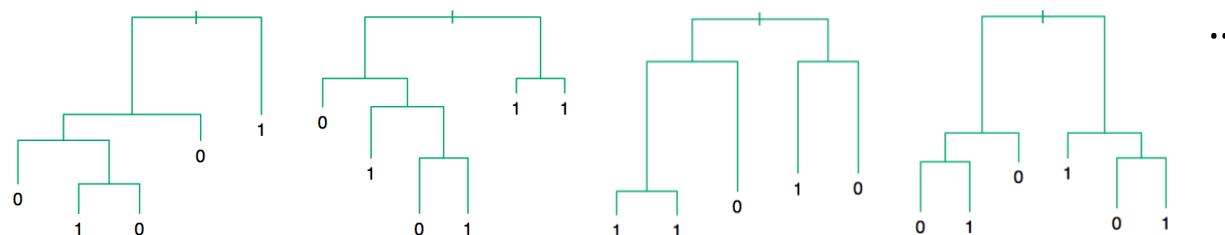
- Lấy mẫu tập dữ liệu huấn luyện theo Bootstrap để tạo ra tập hợp các dự đoán.



Bagging

- Lấy mẫu tập dữ liệu huấn luyện theo Bootstrap để tạo ra tập hợp các dự đoán.

Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.



- Lấy trung bình (hoặc bình chọn theo số đông- majority vote) các bộ dự đoán độc lập.
- Bagging giảm phương sai (variance) và giữ bias.

Bagging

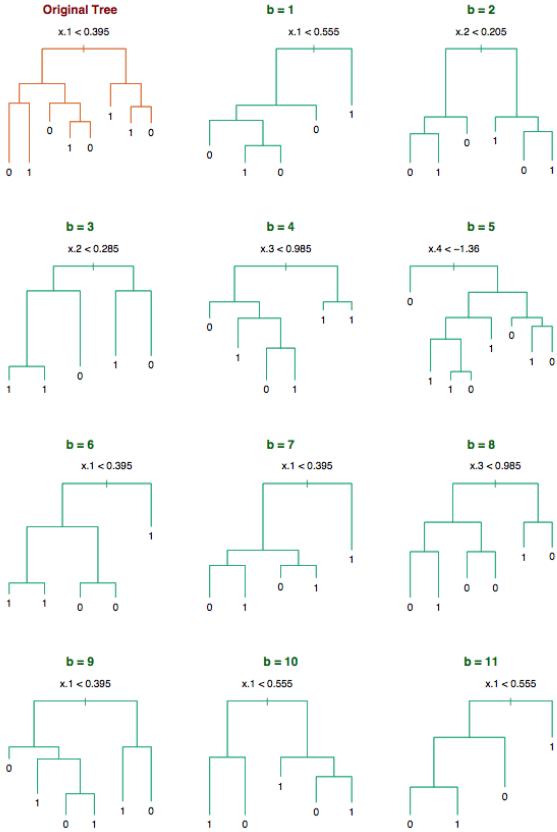
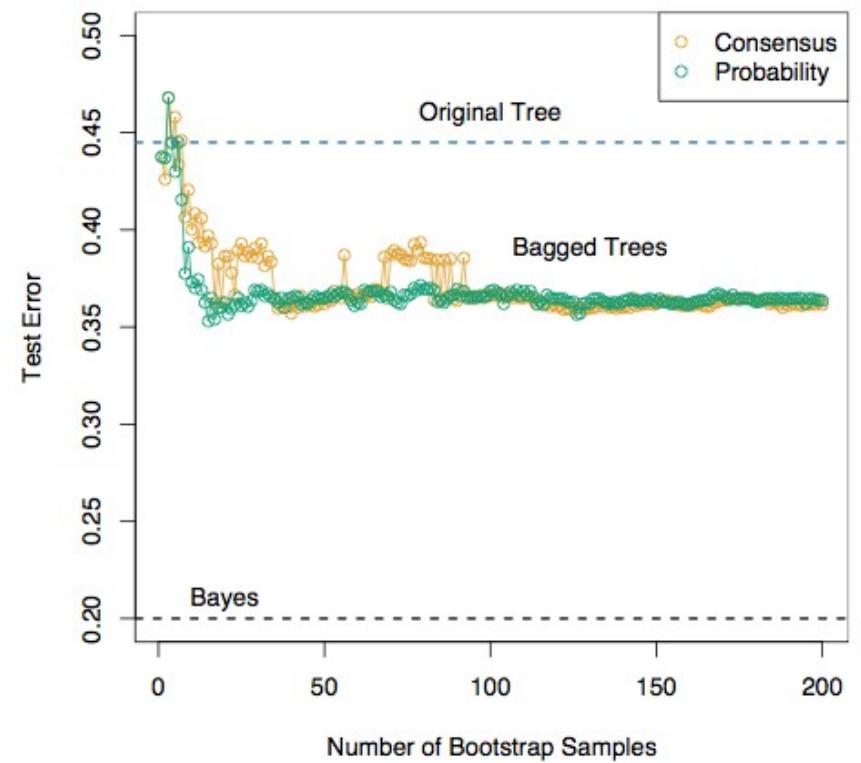


FIGURE 8.9. Bagging trees on simulated dataset. The top left panel shows the original tree. Eleven trees grown on bootstrap samples are shown. For each tree, the top split is annotated.

Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.



Bagging

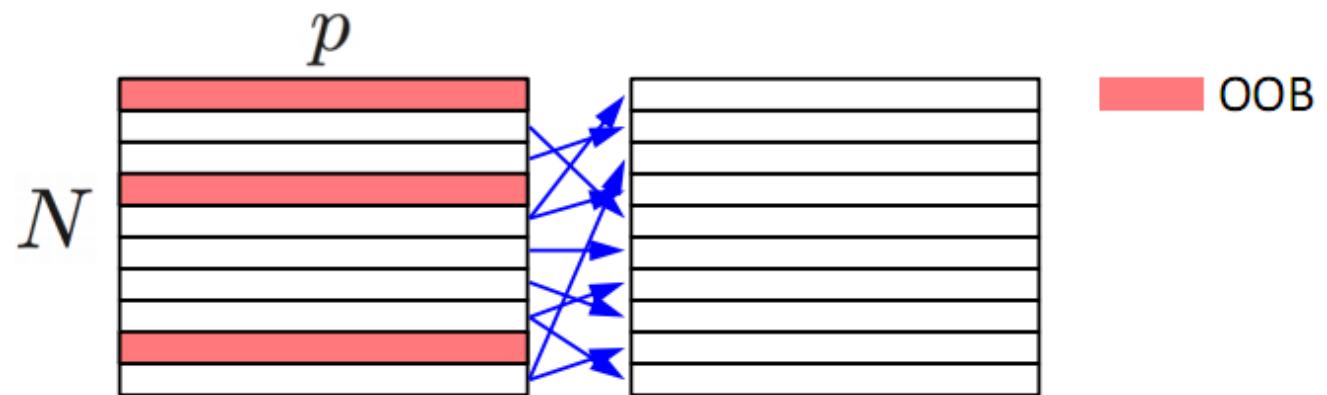
Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Lấy mẫu có hoàn lại
- Xây dựng bộ phân lớp trên mỗi mẫu bootstrap
- Mỗi mẫu bootstrap chứa xấp xỉ 63.2% số lượng mẫu trong tập dữ liệu ban đầu
- Số lượng mẫu còn lại (36.8%) được dùng để kiểm thử

Bonus! Out-of-bag cross-validation

Các mẫu Out-of-bag (OOB)

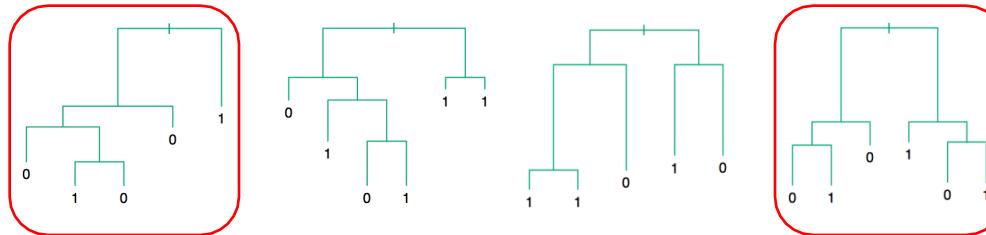
- Quá trình Bootstrapping:



- Mỗi cây chỉ sử dụng một tập con các mẫu huấn luyện (trung bình số mẫu $\sim 2/3$).
- Số mẫu cho OOB khoảng $\sim 1/3$ của cây quyết định.

Dự đoán mẫu OOB

- Với mỗi mẫu, tìm các cây mà nó là OOB.



- Dự đoán giá trị của chúng từ các cây này.
- Ước lượng lỗi dự đoán của cây (bagged trees) dùng tất cả các dự đoán OOB.
- Tương tự như kỹ thuật kiểm tra chéo (cross-validation).

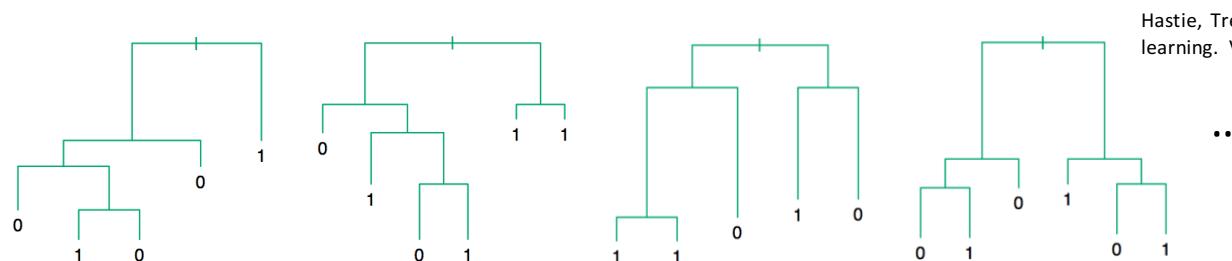
Phương pháp Rừng ngẫu nhiên Random Forests (RF)

Động lực để có Random forest

- Mô hình dựa trên cây phân loại và hồi quy (CART).
- Các mô hình cây có lỗi bias thấp, tuy nhiên phương sai lại cao (high variance).
- Phương pháp Bagging dùng để giảm phương sai.

Nhắc lại: Bagging

- Lấy mẫu tập dữ liệu huấn luyện theo Bootstrap để tạo ra tập hợp các dự đoán.



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

- Lấy trung bình (hoặc bình chọn theo số đông-majority vote) các bộ dự đoán độc lập.
- Bagging giảm phương sai (variance) và giữ bias.

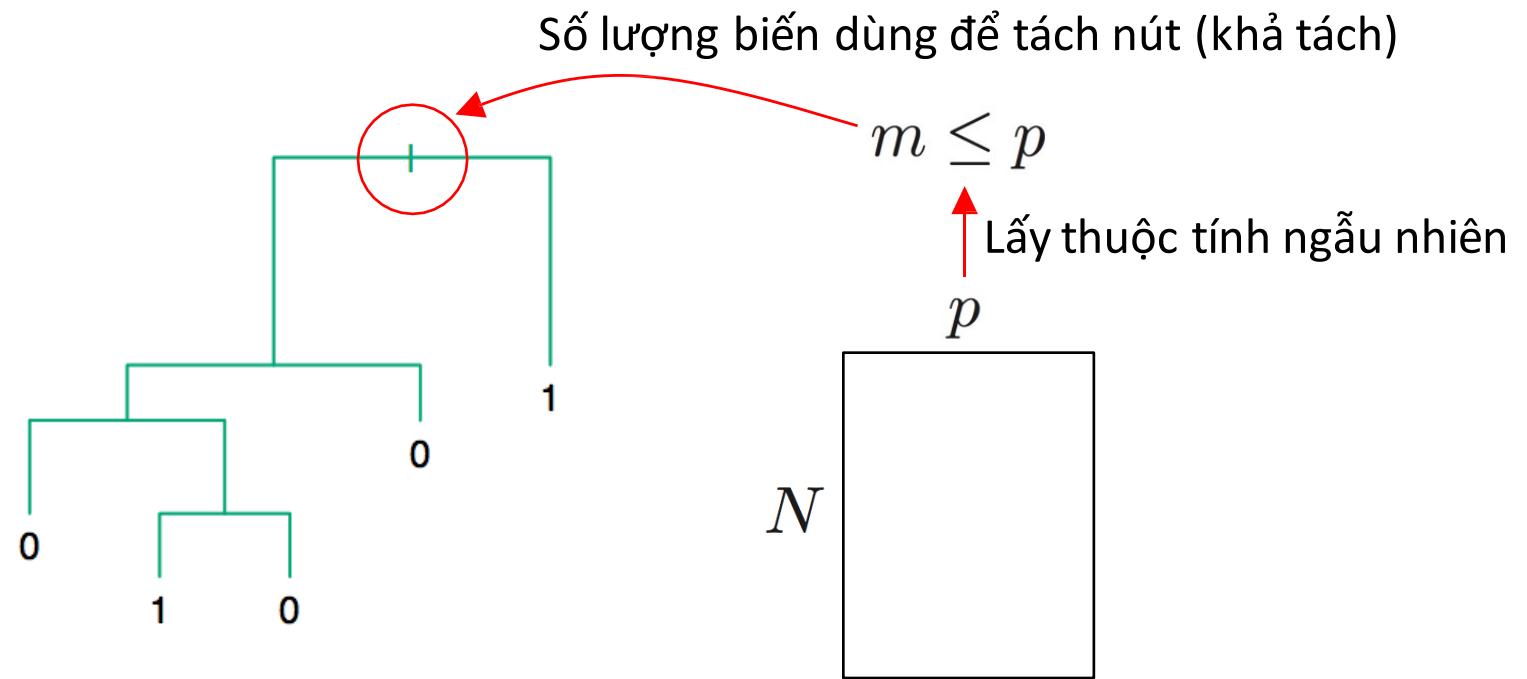
Bagged trees vs. random forests

- Phương pháp Bagging biểu thị sự biến thiên (variability) giữa các cây bởi việc chọn mẫu ngẫu nhiên từ dữ liệu huấn luyện.
- Cây được sinh ra từ phương pháp Bagging vẫn có tương quan lẫn nhau, do đó hạn chế trong việc giảm phương sai.

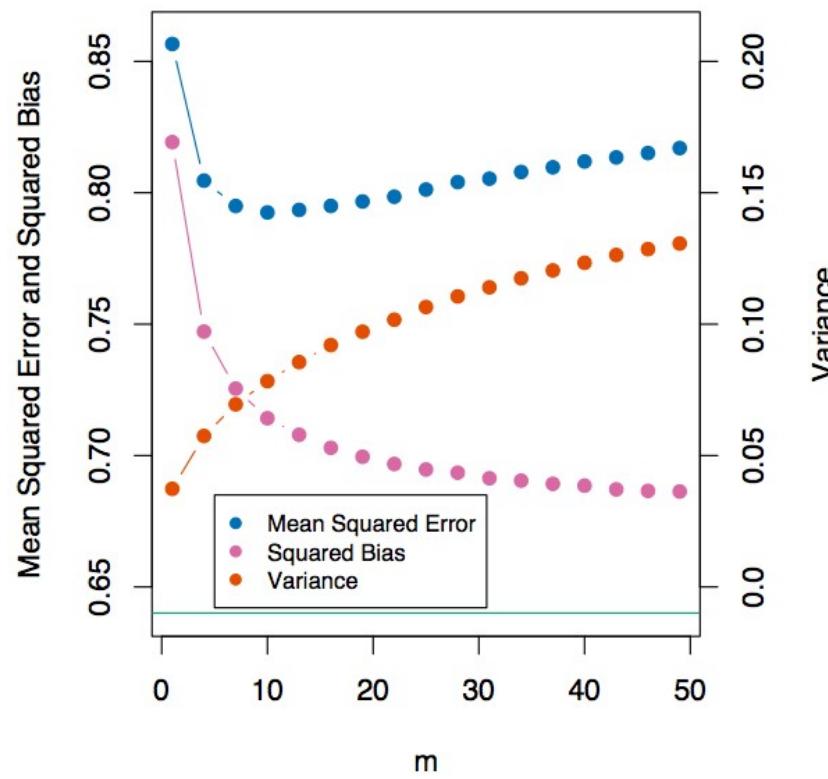
Random forests đưa ra thêm tính ngẫu nhiên (randomness):

- Làm giảm mối tương quan giữa các cây bằng cách lấy ngẫu nhiên các biến khi tách nút của cây.

Các biến dùng cho tách nút

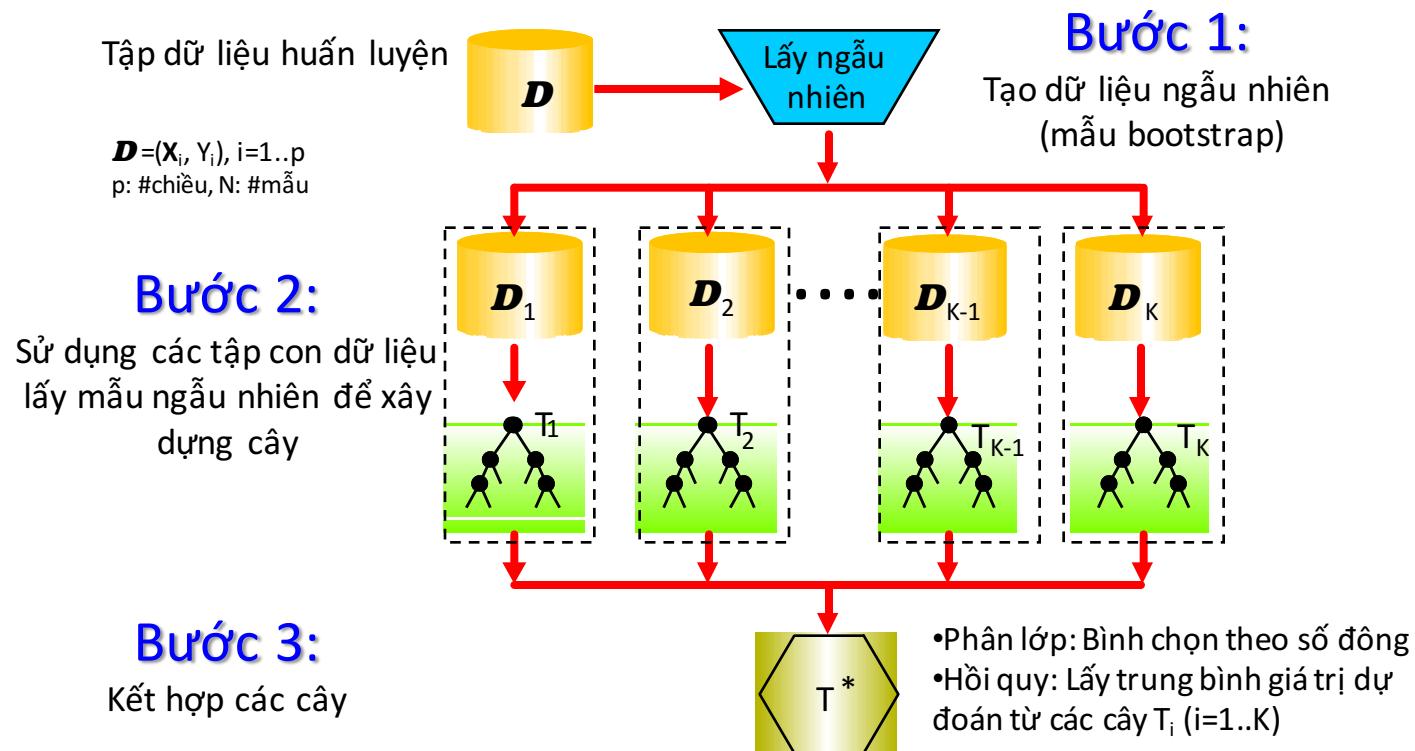


Các biến dùng cho tách nút



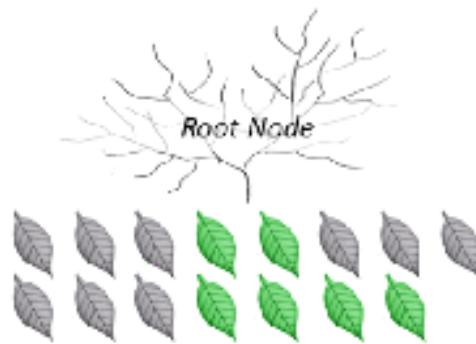
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Rừng ngẫu nhiên



Introduction to Data Mining – Tan, Steinbach, Kumar

Rừng ngẫu nhiên



For more tutorials: algobeans.com

Các tham số chính

Các tham số quan trọng của Rừng ngẫu nhiên:

- Số lượng biến khả tách tại mỗi nút (m)
- Độ sâu của từng cây trong rừng (số lượng mẫu tối thiểu tại mỗi nút của cây-minimum node size)
- Số lượng cây trong rừng

Số lượng biến khả tách

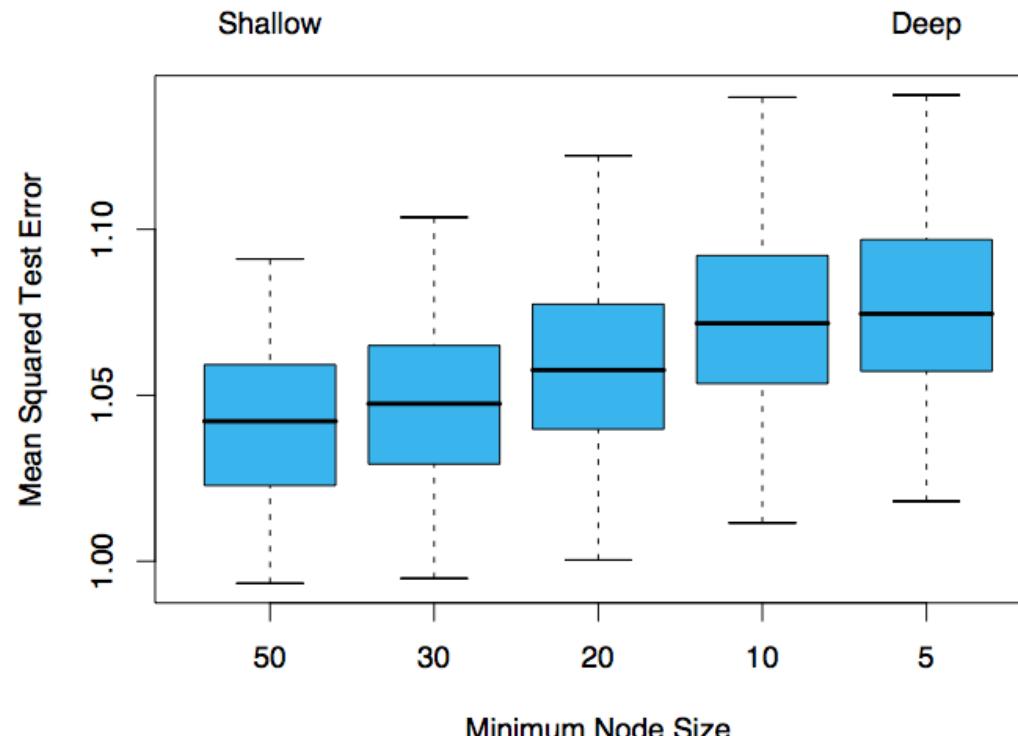
Giá trị mặc định

Bài toán phân lớp $m = \lfloor \sqrt{p} \rfloor$

Bài toán hồi quy $m = \lfloor p/3 \rfloor$

gói randomForest trong R dùng *mtry*

Độ sâu của từng cây (số lượng mẫu tối thiểu tại mỗi nút của cây)



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

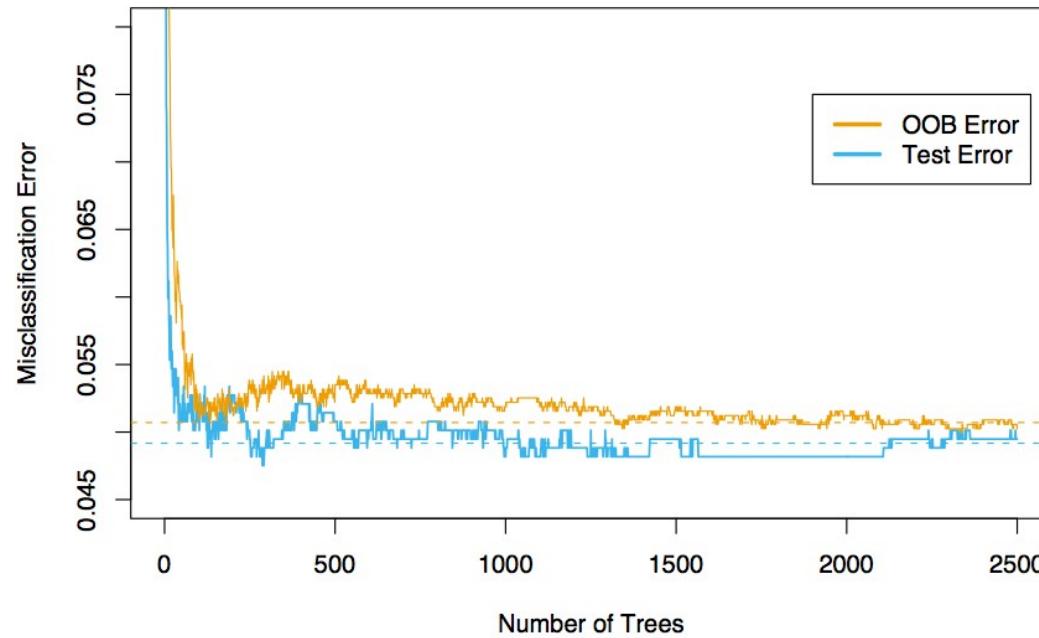
Độ sâu của cây

Giá trị mặc định

Bài toán phân lớp 1

Bài toán hồi quy 5

Số lượng cây trong rừng



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

- Thêm nhiều cây không gây ra overfitting.

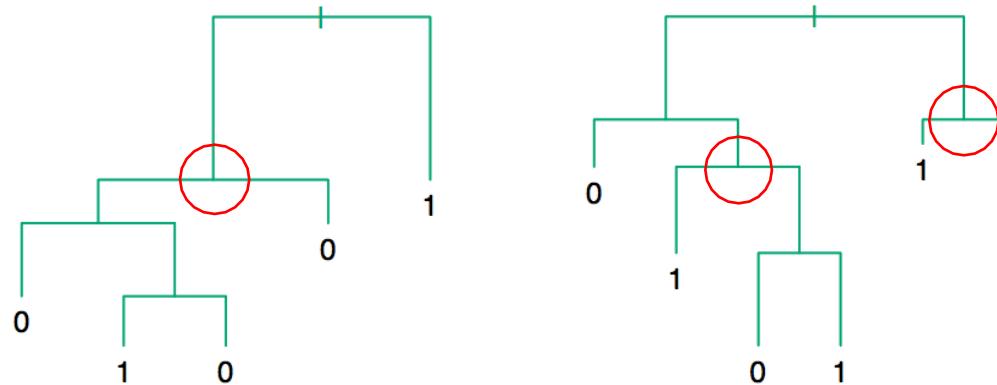
Các tính năng khác của RF

- Các mẫu Out-of-bag (OOB)
- Độ quan trọng của biến (Variable importance measurements)

Độ quan trọng của biến

Dạng 1:

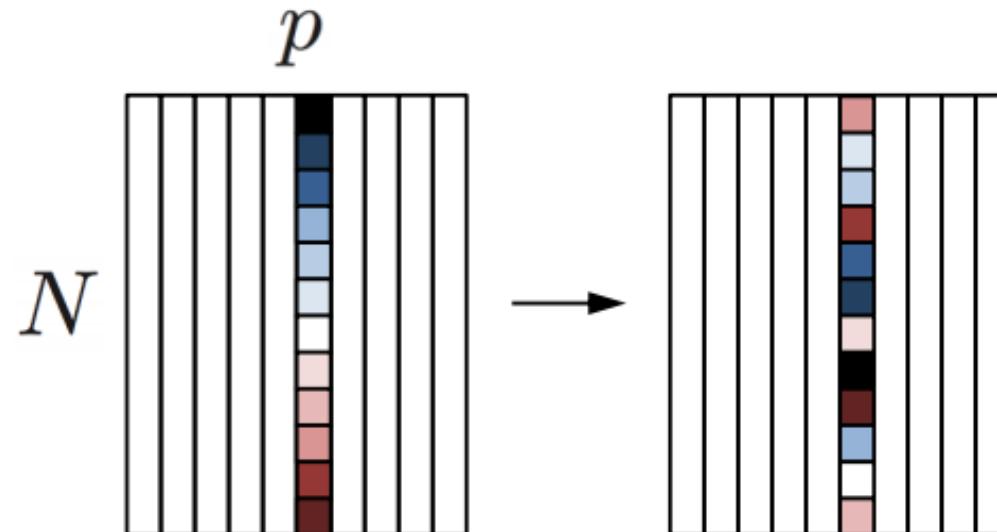
Độ giảm của lỗi dự đoán hoặc impurity từ các điểm tách nút liên quan đến các biến đó, cuối cùng lấy trung bình trên các cây trong rừng.



Độ quan trọng của biến

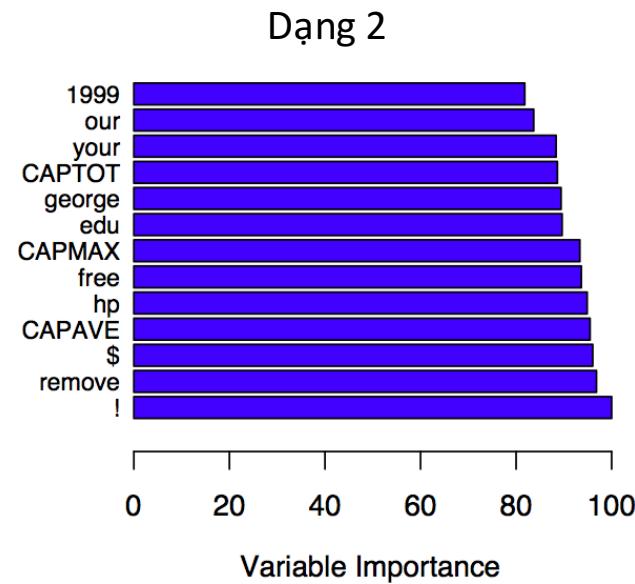
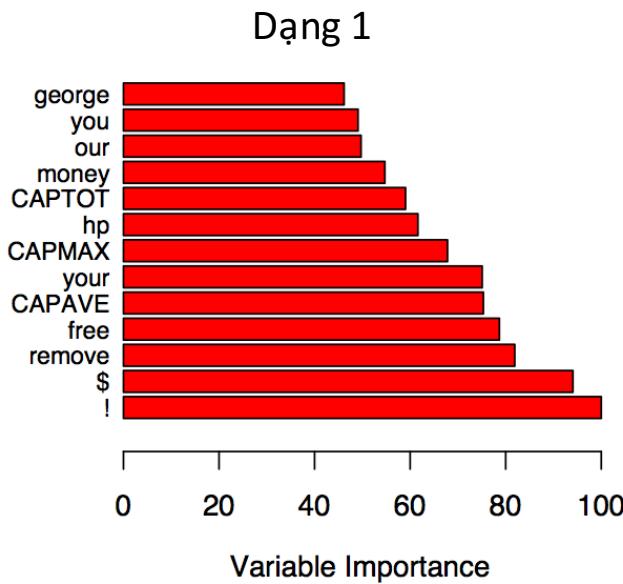
Dạng 2:

Độ tăng lỗi dự đoán tổng thể khi các giá trị của biến được hoán vị ngẫu nhiên giữa các mẫu.



Ví dụ về độ quan trọng của biến

- Cả 2 dạng biểu thị gần giống nhau, tuy nhiên có sự khác biệt về xếp hạng các biến:



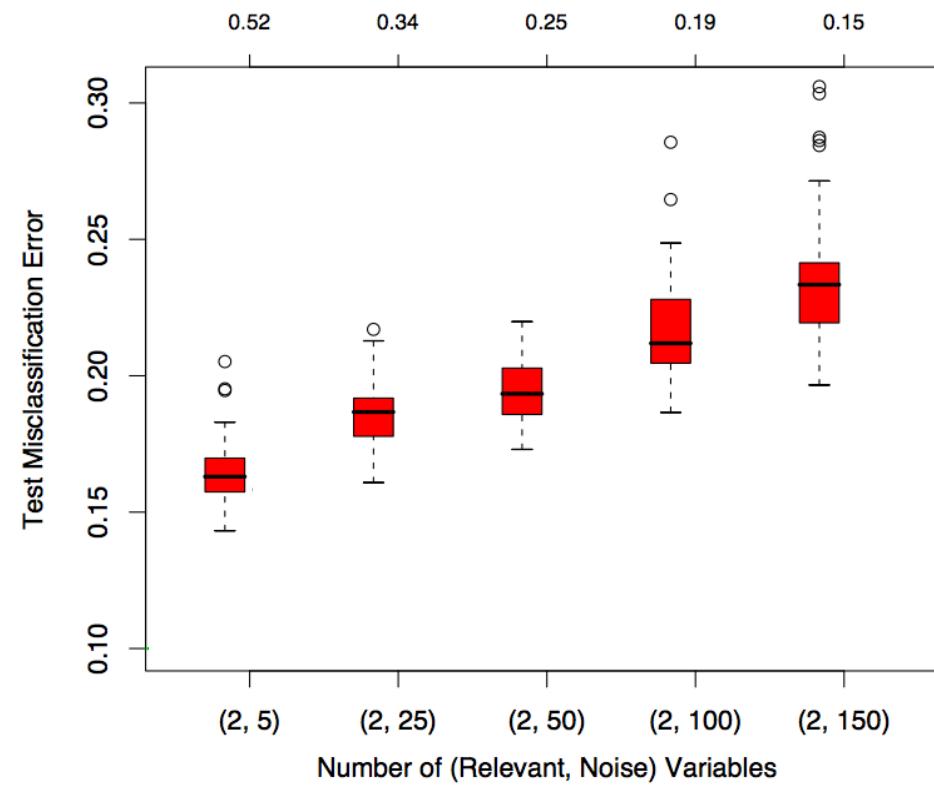
Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Ưu điểm của RF

Tương tự như CART:

- Tương đối mạnh trong việc xử lý biến rác (non-informative variable)
(Việc lựa chọn biến tích hợp sẵn khi xây dựng mô hình, built-in variable selection)

Ảnh hưởng của biến rác



Hastie, Trevor, et al. The elements of statistical learning. Vol. 2. No. 1. New York: Springer, 2009.

Ưu điểm của RF

Tương tự như CART:

- Tương đối mạnh trong việc xử lý biến rác (non-informative variable)
- Xử lý (nắm bắt) được độ tương tác bậc cao giữa các biến (Capture high-order interactions between variables)
- Có lỗi bias thấp
- Dễ xử lý các biến hỗn hợp (biến rời rạc, phân loại)

Ưu điểm của RF

Ưu điểm vượt trội CART:

- Lỗi phương sai thấp hơn (mạnh hơn vì sử dụng phương pháp bootstrapping lấy mẫu từ tập huấn luyện)
- Ít bị overfitting hơn
- Không cần tỉa cây (No need for pruning)
- Kiểm tra chéo được tích hợp sẵn trong mô hình (dùng các mẫu OOB)

Nhược điểm của RF

Tương tự như CART:

- Khó nắm bắt độ cộng tính

Nhược điểm so với CART:

- Khó diễn giải/giải thích mô hình dự đoán

Câu hỏi?

Nội dung

1. Giới thiệu mô hình hồi quy
2. Overfitting, kỹ thuật đánh giá chéo
3. Phân tích dữ liệu với R
4. Hồi quy tuyến tính
5. Hồi quy phi tuyến
6. Real-life problem

Giới thiệu bài toán dự đoán

- Cho tập dữ liệu đầu vào $\mathcal{L} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$, trong đó N là số lượng mẫu.
 - Đầu vào là tập biến ngẫu nhiên $X \in \mathbb{R}^M$, M số thuộc tính.
 - Đầu ra là biến ngẫu nhiên $Y \in \mathbb{R}^1$.
- $x_i \in X$ và $y_i \in Y$ nhận các giá trị ngẫu nhiên từ phân bố xác suất $P_{x,y}$ ($1 \leq i \leq N$).
- Mục tiêu của bài toán dự đoán là tìm mô hình $f_L : X \rightarrow Y$ cực tiểu hóa

$$\text{Err}(f_L) = E_{X,Y} \{ L(Y, f_L(X)) \},$$

Trong đó hàm lỗi là

$$L(Y, f_L(X)) \} = (Y - f_L(X))^2.$$

Dự đoán sự hài lòng của các hộ dùng nước tưới tiêu tại đồng bằng sông Hồng

5. Đáp ứng (RES)

- Nhân viên thủy lợi cho ông bà biết khi nào thực hiện dịch vụ tưới tiêu
- Nhân viên thủy lợi nhanh chóng thực hiện dịch vụ cho ông bà.
- Tổ chức cung cấp nước thực hiện đúng lịch cấp nước
- Tổ chức cung cấp nước cung cấp tối đa khả năng cấp nước.
- Khối lượng nước cấp đáp ứng tốt nhu cầu theo từng giai đoạn sinh trưởng, phát triển của cây trồng.
- Nhân viên thủy lợi cung cấp luôn sẵn sàng đáp ứng yêu cầu của ông bà.
- Chất lượng nước tưới được đảm bảo
- Thời gian khắc phục hư hỏng nhanh chóng
- Ông bà không bao giờ phải lặp lại các khiếu nại trước (9 biến quan sát)

3. Đảm bảo (ASS)

- Cách cư xử của nhân viên gây niềm tin cho ông bà
- Ông bà cảm thấy rất an toàn khi giao dịch với tổ chức cung cấp nước
- Nhân viên thủy lợi có đủ hiểu biết để trả lời tất cả các câu hỏi của ông bà liên quan đến hệ thống tưới, tiêu.
- Nhân viên thủy lợi của tổ chức cung cấp nước luôn luôn niềm nở với ông bà
- Thời gian phân phối nước tới các thửa ruộng luôn luôn đủ nước trong mỗi đợt tưới.
- Từ năm 2008 đến nay nhân viên thủy lợi trả lời được tất cả các thắc mắc của ông bà liên quan đến số tiền ông bà trả trong tháng
- Nhân viên thủy lợi rất nhanh khắc phục khi hệ thống tưới, tiêu có sự cố (7 biến quan sát)

1. Phương tiện hữu hình (TAN)

- Các hệ thống tưới, tiêu có chất lượng tốt, đảm bảo chuyển nước và phân phối nước đến các diện tích cần tưới, tiêu
 - Các đơn vị cung cấp dành đủ kinh phí cho công tác quản lý, vận hành và bảo dưỡng hệ thống tưới, tiêu.
 - Nhân viên thủy lợi mặc đồng phục đơn vị
 - Tổ chức cung cấp nước có tài liệu hướng dẫn quản lý vận hành công trình thủy lợi.
 - Hợp đồng cung cấp dịch vụ được trình bày rất dễ hiểu
 - Các thiết bị của tổ chức cung cấp nước có chất lượng tốt
 - Việc duy tu, bảo dưỡng hệ thống tưới được thực hiện đều đặn và khi cần.
- (7 biến quan sát)

2. Tin cậy (REL)

- Đơn vị cung cấp dịch vụ tưới, tiêu giới thiệu đầy đủ nội dung hợp đồng với tổ chức cung cấp nước cũng như các kỹ thuật và cách sử dụng khi ông bà muốn đăng ký sử dụng
- Tổ chức cung cấp nước thực hiện đúng dịch vụ tưới tiêu như hợp đồng
- Tổ chức cung cấp nước xử lý sự cố ngay khi công trình hư hỏng, xuống cấp.
- Từ năm 2008 đến nay tổ chức cung cấp nước không để xảy ra bất kỳ sai sót nào khi tính chi phí hàng tháng (4 biến quan sát)

Sự hài lòng (SAT)

Ông bà hoàn toàn hài lòng về chất lượng dịch vụ tưới tiêu hiện đang sử dụng.
(Giá trị từ 0.5, kiểu thập phân).

4. Sự đồng cảm (EMP)

- Nhân viên kỹ thuật thủy lợi luôn làm việc vào những giờ thuận tiện cho ông bà.
 - Không có bất cứ ai ở Tổ chức cung cấp nước quan tâm đến những bức xúc của ông bà về dịch vụ tưới, tiêu.
 - Lịch phân phối nước rất thuận tiện theo giờ sản xuất của gia đình ông bà.
 - Ông bà được quan tâm và chú ý mỗi khi thắc mắc về dịch vụ tưới, tiêu.
 - Tổ chức cung cấp nước điều chỉnh lịch tưới phù hợp với sự thay đổi của thời tiết.
 - Nhân viên của tổ chức cung cấp nước luôn hiểu rõ nhu cầu của ông bà.
 - Đơn vị cung cấp lấy lợi ích của ông bà là mục tiêu phát triển bền vững của họ
- (7 biến quan sát)

Một số mô hình học máy

- Linear Regression
- LASSO
- K-NN
- Support Vector Regression
- Artificial neural network
- Decision trees
- Random Forests
- Boosting
- Deep Learning

Kết quả thực nghiệm

- Phương pháp đánh giá:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \hat{Y}_i)^2}; MAE = \frac{1}{N} \sum_{i=1}^N |Y_i - \hat{Y}_i|$$

và $R^2 = 1 - \sum_{i=1}^N (Y_i - \hat{Y}_i) / \sum_{i=1}^N (Y_i - \bar{Y})$.

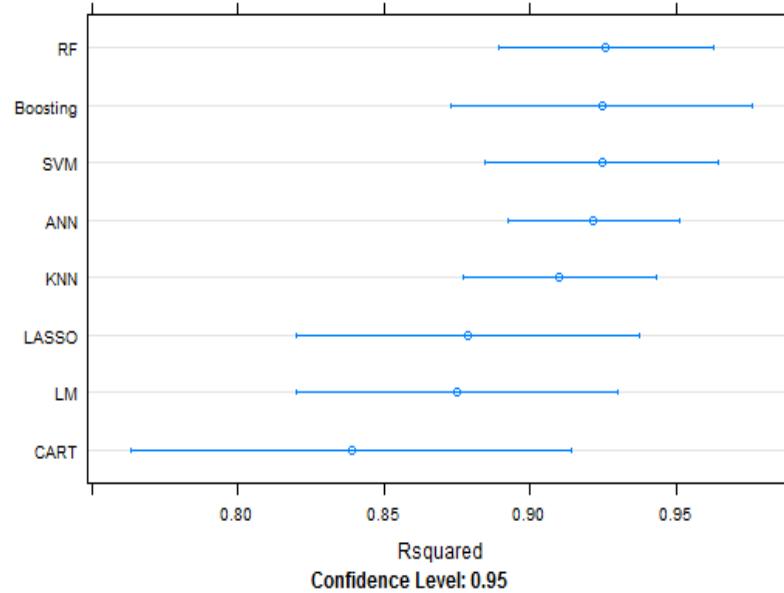
- Dữ liệu: Tập huấn luyện gồm 336 mẫu (70%) và tập dữ liệu kiểm thử gồm 144 mẫu (30%).
- Khi xây dựng mô hình hồi quy, kỹ thuật kiểm tra chéo 5-folds với 2 lần lặp và dựa trên hàm lỗi RMSE được dùng để tìm tham số tối ưu của từng mô hình, sau đó lựa chọn mô hình có RMSE nhỏ nhất với tham số tìm được để dự đoán dữ liệu kiểm thử.

Kết quả thực nghiệm

TT	Mô hình hồi quy	Tham số tối ưu	R ²	RMSE	MAE
1	Hồi quy tuyến tính (LM)	Mặc định	0.839	0.267	0.167
2	Hồi quy LASSO	$\lambda = 0.01$	0.844	0.263	0.163
3	K láng giềng (KNN)	$k = 1$	**0.894	**0.216	0.085
4	Cây hồi quy (CART)	Complexity parameter (cp)=0	0.835	0.272	0.156
5	Mạng nơ ron nhân tạo (ANN)	Trọng số phân rã=0.1 và số nơ-ron=9	***0.892	***0.218	**0.106
6	Máy véc-tơ hỗ trợ (SVR)	RBF, $\sigma = 0.032$, $\varepsilon = 0.1$ và $C = 32$	0.852	0.255	0.143
7	Rừng ngẫu nhiên (RF)	mtry = 9 và K=1000	0.902	0.208	***0.107
8	Boosting	K = 500, interaction.depth = 7 và shrinkage = 0.1	0.873	0.237	0.119

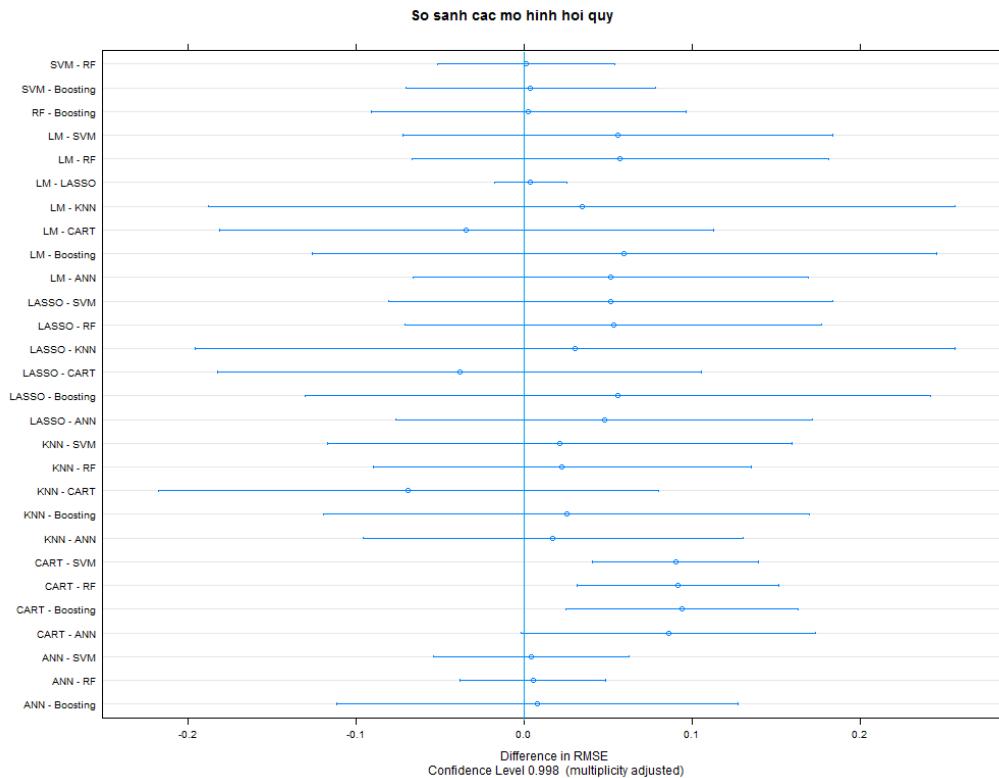
Kết quả thực nghiệm

- Mô hình rừng ngẫu nhiên cho kết quả tốt nhất, giải thích khoảng 93% các khác biệt về độ hài lòng giữa các hộ dùng nước tưới tiêu, theo sát là mô hình boosting có $R^2=92.445\%$ và SVR đạt $R^2=92.444\%$.
- Xếp cuối là phương pháp cây hồi quy có R^2 thấp nhất, khả năng giải thích của mô hình cây hồi quy khoảng 85% kém hơn mô hình hồi quy tuyến tính nhiều biến có $R^2=87.481\%$.

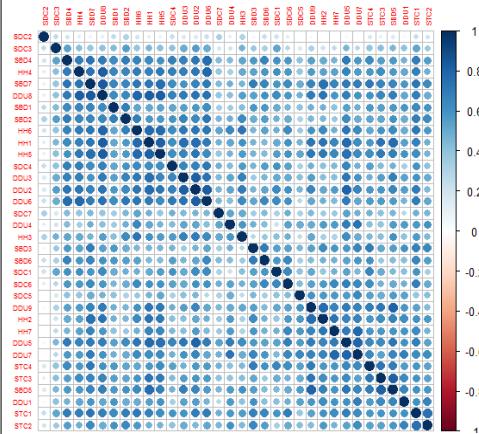


So sánh các mô hình hồi quy dựa trên kết quả huấn luyện theo hệ số xác định bội R^2

Kết quả thực nghiệm



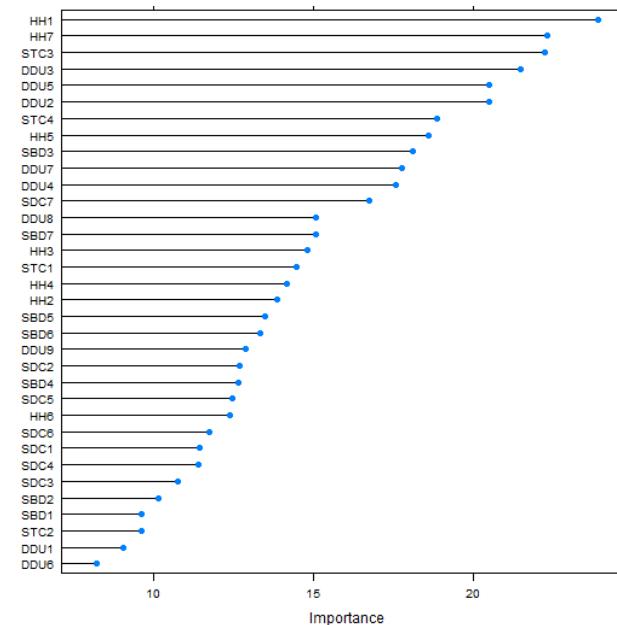
So sánh lỗi huấn luyện RMSE của các mô hình hồi quy theo từng cặp.



Kết quả thực nghiệm

- Độ đo sự quan trọng của 34 tiêu chí được sắp xếp theo chiều giảm dần, các độ đo này được tính từ rừng ngẫu nhiên.
- HH1, HH7, STC3 có độ quan trọng cao, trong đó HH1= "Các hệ thống tưới, tiêu có chất lượng tốt, đảm bảo chuyển nước và phân phôi nước đến các diện tích cần tưới, tiêu" có độ quan trọng cao nhất. Tiêu chí DDU6= "Nhân viên thủy lợi cung cấp luôn sẵn sàng đáp ứng yêu cầu của ông bà" có độ quan trọng thấp nhất.
- Như vậy, trong dịch vụ cung cấp nước tưới tiêu, hộ dùng nước quan tâm nhất đến các hệ thống tưới tiêu có chất lượng tốt, độ đáp ứng của đơn vị cung cấp nước, nó bao gồm những yếu tố như duy tu, bảo dưỡng được thực hiện đầy đủ và đều đặn, sửa chữa sự cố ngay khi công trình hư hỏng hoặc xuống cấp, thực hiện đúng lịch cấp nước, cung cấp tối đa khả năng cấp nước, đáp ứng tốt nhu cầu theo từng giai đoạn sinh trưởng và phát triển của cây trồng, chất lượng nước được đảm bảo.
- Nhân viên thủy lợi có hoặc không đáp ứng những yêu cầu cá nhân của các hộ dùng nước cũng không ảnh hưởng nhiều đến sự hài lòng chung về chất lượng dịch vụ tưới tiêu

Xem thêm bài báo [ở đây](#)

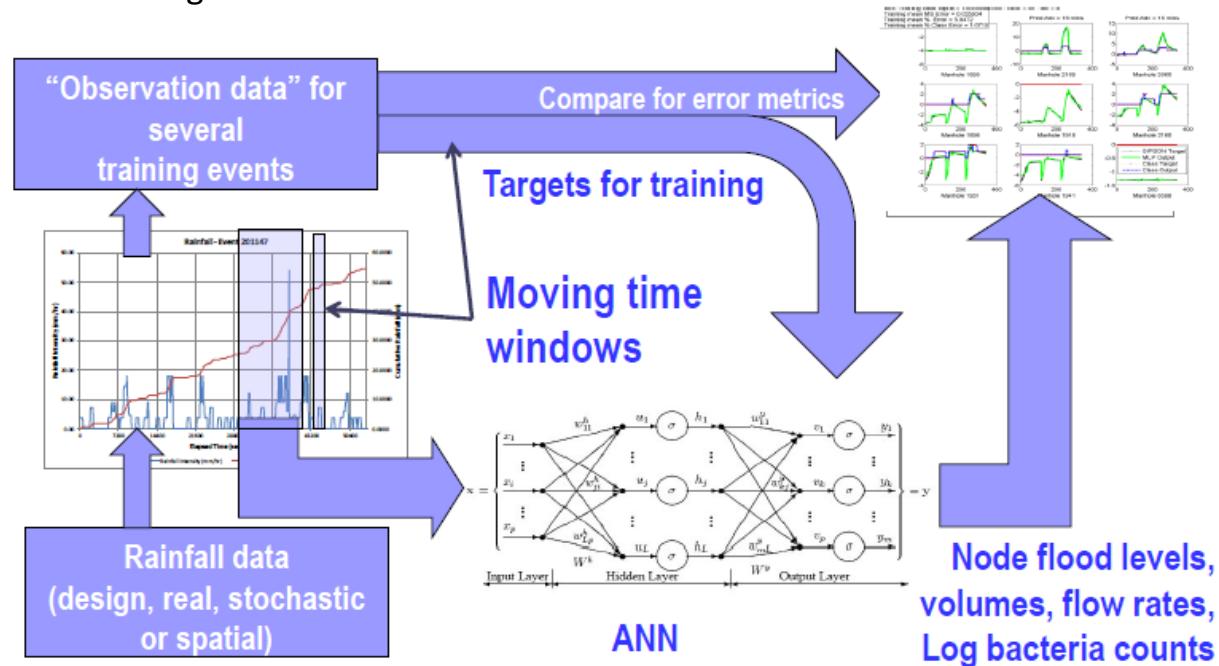


Độ đo sự quan trọng của các tiêu chí

Dự báo mực nước trên sông Mekong

Applications of Machine Learning

Flood forecasting

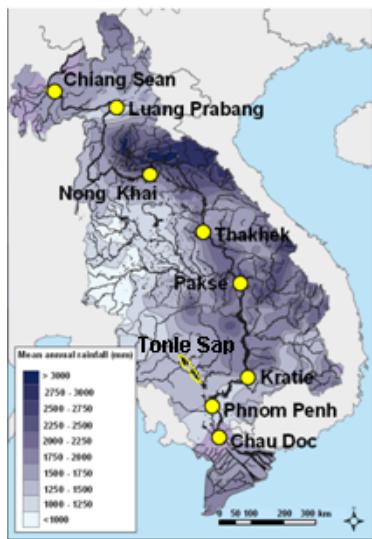


Motivation

- Two approaches to build the flood forecasting model:
 - physically based and
 - data-driven (machine learning) approaches.
- Physically based models are fully distributed models in increasing levels of complexity. The physically based modelling aims to reproduce the hydrological process in a physically realistic.
- Our solution: We use machine learning model, they are quickly developed and easily implemented for building the forecasting model.

Motivation

- Case study: Lower Mekong river.
- Inputs:
 - Rainfall intensity
 - Cumulative rainfall
- Outputs: the 5-lead-day water levels at Thakhek gauging station



Experiments

- Forecasting model: the 5-lead-day water levels at Thakhek station on the Mekong River, where it shows the major contribution to the flows in the Lower Mekong River.
- The relationship between the input-output features:

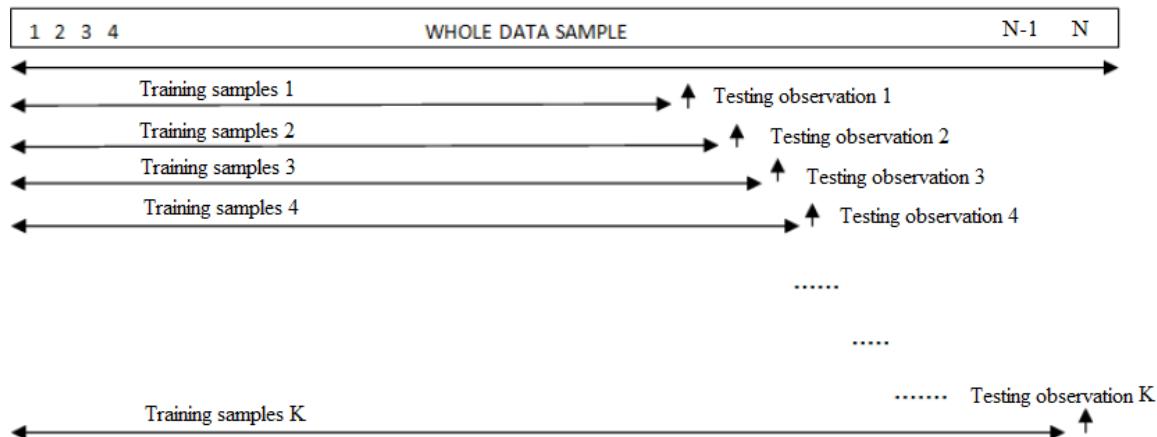
$$H_{Thakhek}(t+5) = f(H_{Thakhek}(t), H_{Thakhek}(t-1), H_{Thakhek}(t-2), H_{up}(t), H_{up}(t-1), H_{up}(t-2)).$$



where the output feature $H_{Thakhek}(t+5)$ is the water level forecasted for the next 5 days at Thakhek gauging station. $H_{Thakhek}(t)$, $H_{Thakhek}(t-1)$ and $H_{Thakhek}(t-2)$ are water levels measured in the current day and previous two days, respectively. $H_{up}(t)$, $H_{up}(t-1)$ and $H_{up}(t-2)$ are water levels measured in the current day and previous two days at NongKhai gauging station, respectively.

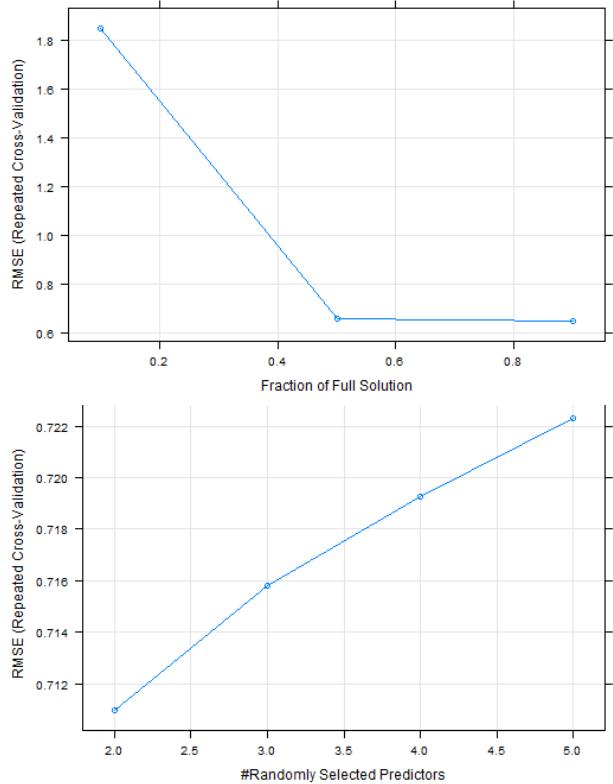
Experiments

- Design of the Forecast Evaluations

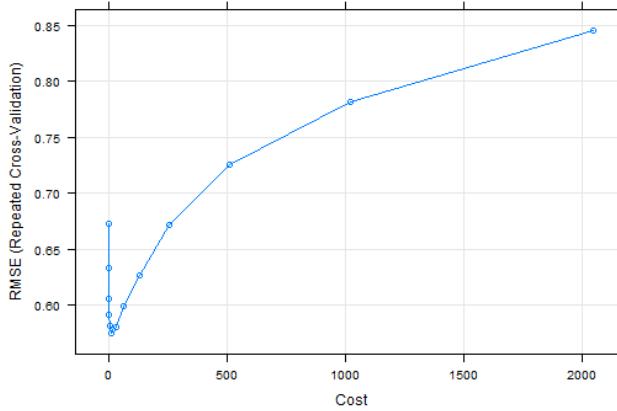


For each iteration, 1 sample from the testing data is added into the training data to build the forecasting model.

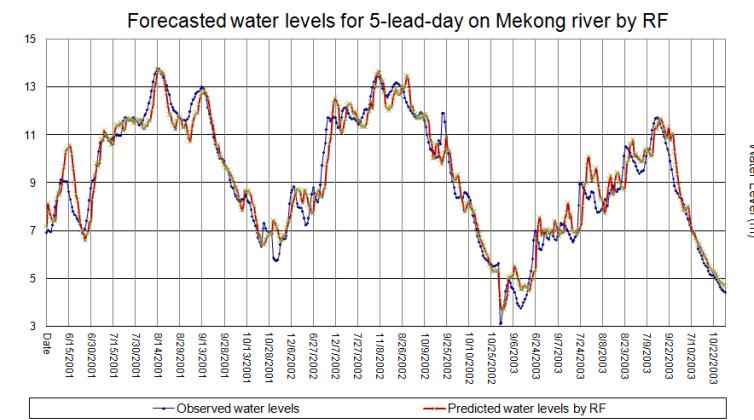
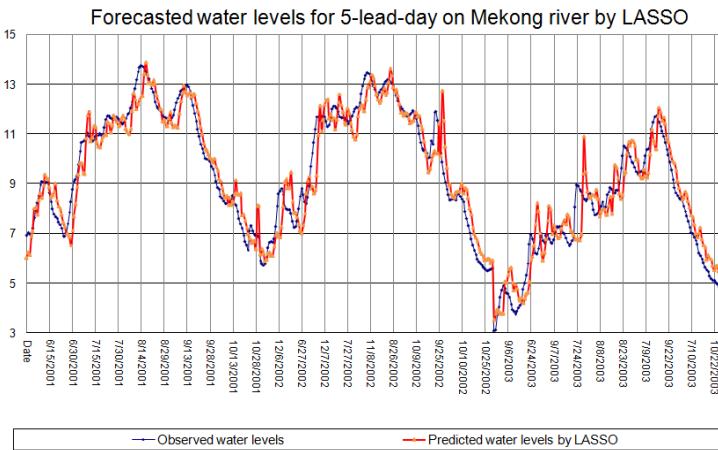
Experimental results



- Optimal parameters: k-folds cross-validation.



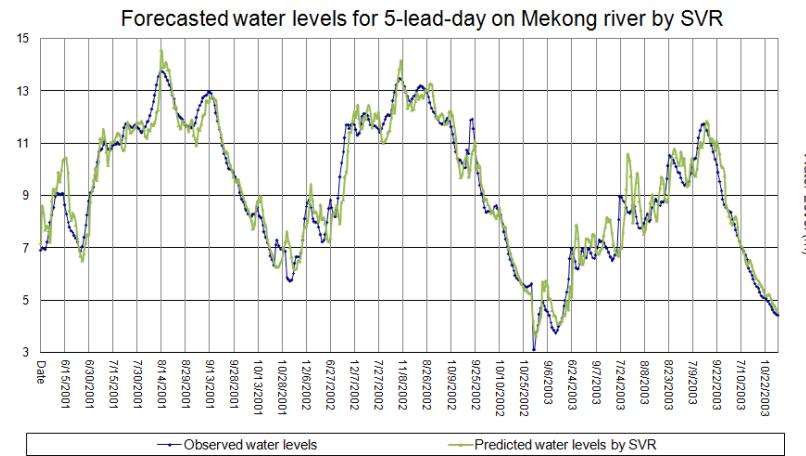
Experimental results



Experimental results

TABLE I. PREDICTIVE PERFORMANCE FOR FORECASTING 5-LEAD-DAY OF WATER LEVELS ON THE MEKONG RIVER.

Model used	Parameter	CE	RMSE	MAE
LASSO	default	0.911	0.761	0.604
RF	$mtry = 2, K = 1000$	0.936	0.649	0.491
SVR	$\varepsilon = 0.1, C = 8, \sigma = 0.235$	0.935	0.646	0.486



Câu hỏi?