Bài giảng môn học:

Kỹ nghệ tri thức và học máy (4080540)

CHƯƠNG 3:
HỌC CÓ GIÁM SÁT - 03
(Supervised Learning)

Giảng viên: Đặng Văn Nam

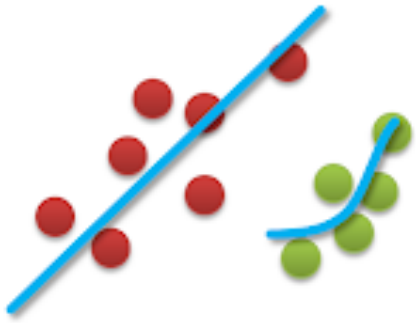
Email: dangvannam@humg.edu.vn

09/2020

Nội dung chương 3

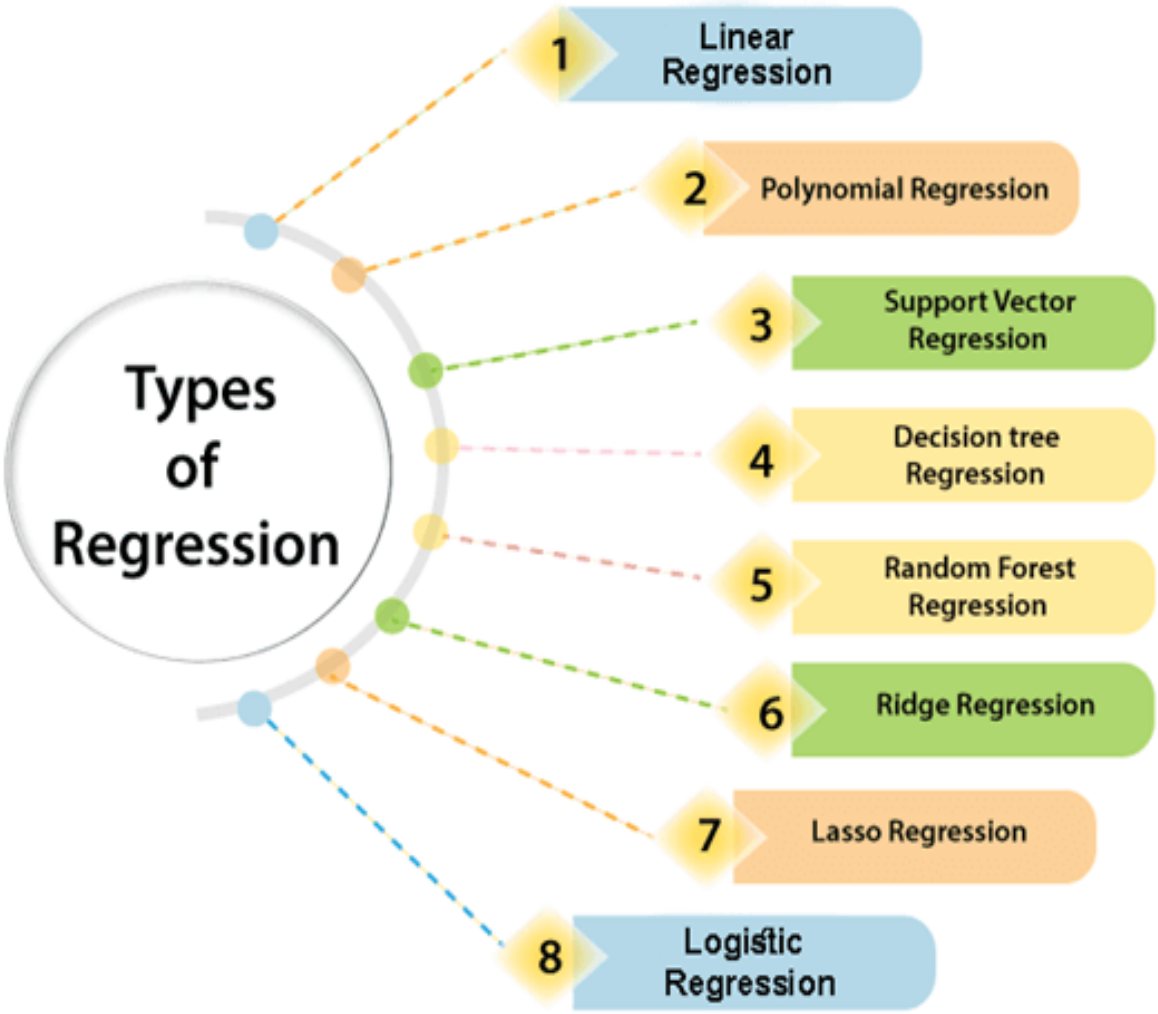
1. Các bước xây dựng một mô hình học máy
2. Datasets
3. Học có giám sát (Supervised Learning)
4. Phân loại học có giám sát (Classification – Regression)
5. Thuật toán phân loại (KNN, Decision Tree)
6. Thuật toán hồi quy (Linear, Polinomial, KNN regression)
7. Đánh giá độ chính xác của mô hình phân lớp,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

6. Mô hình hồi quy (Regression)

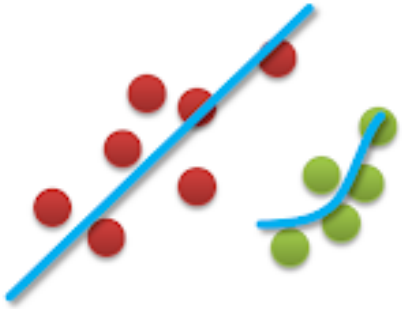


Thuật toán hồi quy

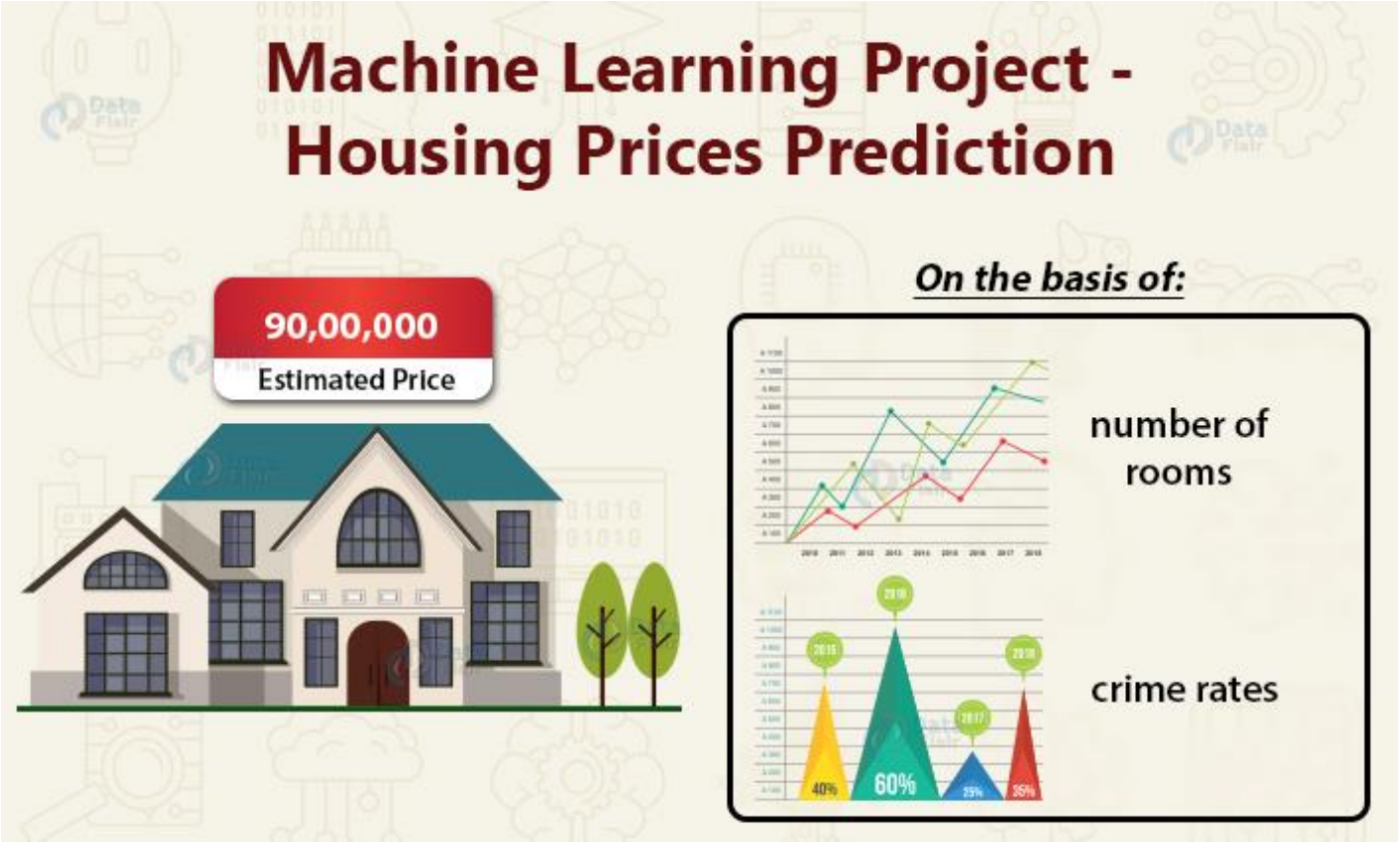
Advertisement	Sales
$90	$1000
$120	$1300
$150	$1800
$100	$1200
$130	$1380
$200	??



6.1 Hồi quy tuyến tính (Linear Regression)



Bài toán dự đoán giá nhà.



- Tập dữ liệu bao gồm 506 mẫu:
  - 13 thuộc tính đầu vào (features)
  - Thuộc tính target (MEDV)

Tham khảo tiến trình thực hiện trong file code trên Jupyter Notebook

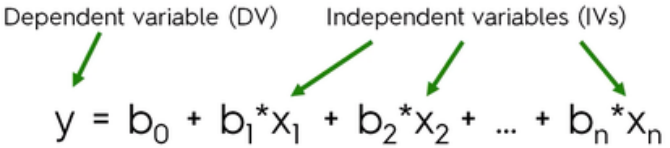
Hồi quy tuyến tính (Linear Regression).

- Hồi quy tuyến tính với 1 biến độc lập X là biến đầu vào (input) để xác định 1 biến đầu ra y (target) – **Simple Linear Regression.**
- Hồi quy tuyến tính với n biến độc lập X_1, …., X_n để xác định 1 biến đầu ra y (target) – **Multiple Linear Regression.**

Simple Linear Regression

$$y = b_0 + b_1 * x_1$$

Multiple Linear Regression



Linear hay *tuyến tính* hiểu một cách đơn giản là *thẳng, phẳng*.

- Không gian hai chiều, một hàm số được gọi là *tuyến tính* nếu đồ thị của nó có dạng một *đường thẳng*.
- Không gian ba chiều: một hàm số được goi là *tuyến tính* nếu đồ thị của nó có dạng một *mặt phẳng*.
- Không gian nhiều hơn 3 chiều, khái niệm *mặt phẳng* không còn phù hợp nữa, thay vào đó, một khái niệm khác ra đời được gọi là *siêu mặt phẳng* (*hyperplane*).

Simple Linear Regression.

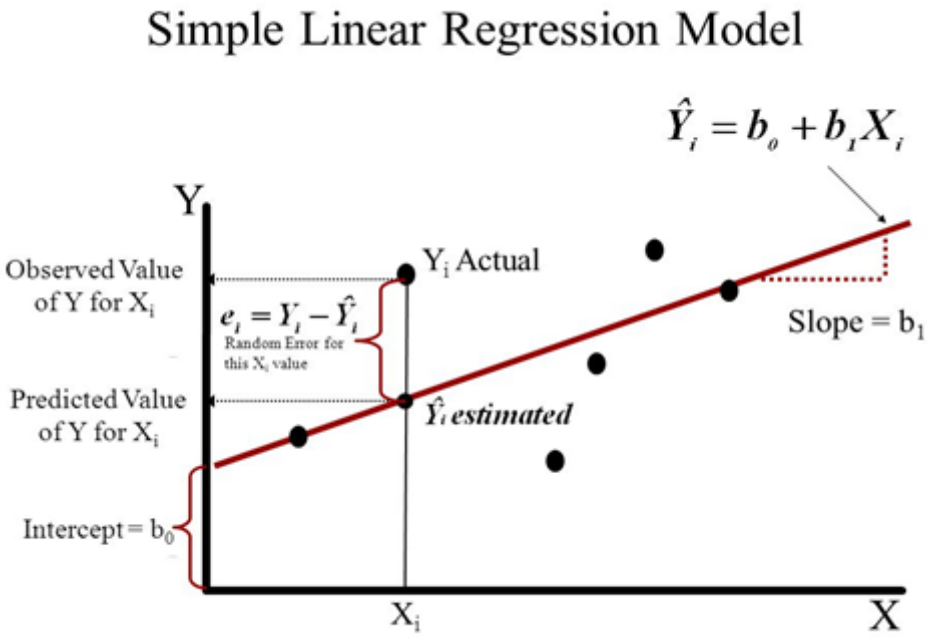
Hồi quy tuyến tính với 1 biến độc lập X là biến đầu vào (input) để xác định 1 biến đầu ra (target)

→ Xác định phương trình:

y = f(x)

$$\hat{y} = \beta_0 + \beta_1 X$$

target — coefficients — input



Mục tiêu ước lượng các tham số b_i sao cho sai số nhỏ nhất.

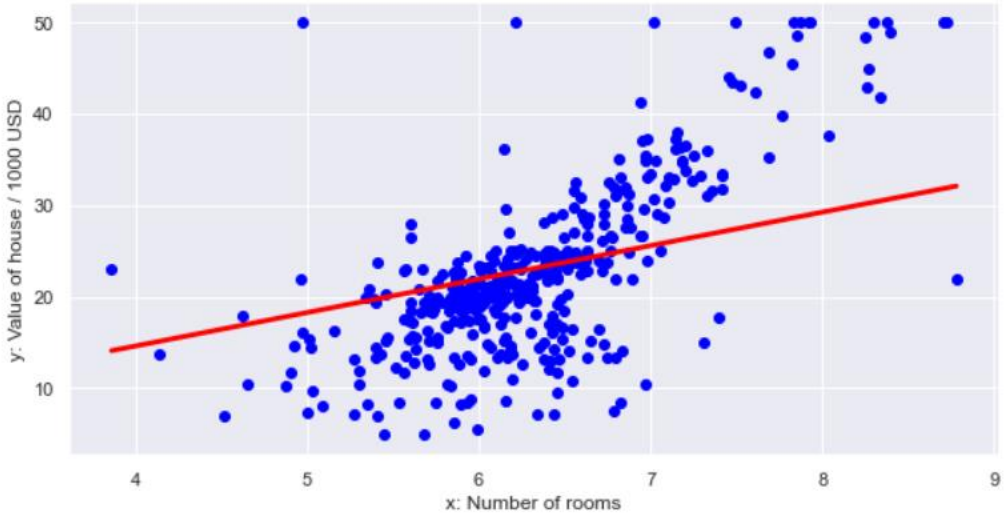
$$RSS = \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

- **y:** Giá trị thật trong tập train (Outcome).
- **ŷ:** Giá trị mà mô hình linear regression dự đoán được.

Simple Linear Regression.

Dự đoán giá nhà với 1 biến độc lập – RM (số phòng trung bình của căn nhà)

RM	MEDV
6.575	24.0
6.421	21.6
7.185	34.7
6.998	33.4
7.147	36.2
6.430	28.7
6.012	22.9
6.172	27.1
5.631	16.5
6.004	18.9
6.377	15.0



$\hat{y}_{MEDV} = f(x) = b_0 + b_1 * X_{RM} = 0 + 3.65279843 * X_{RM}$

Sai số RMSE: **7.67452585343132**

Multiple Linear Regression.

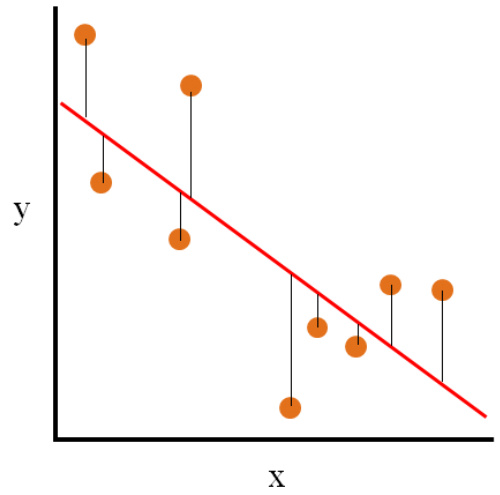
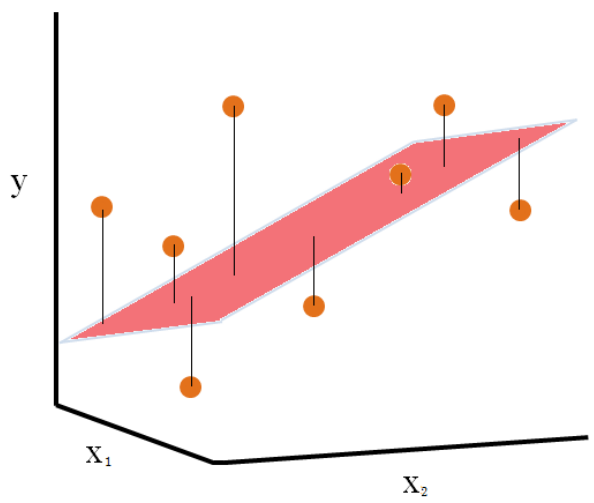
Hồi quy tuyến tính với n biến độc lập (X_1, X_2 …. X_n)

$$\hat{y} = \beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n$$

coefficients

target

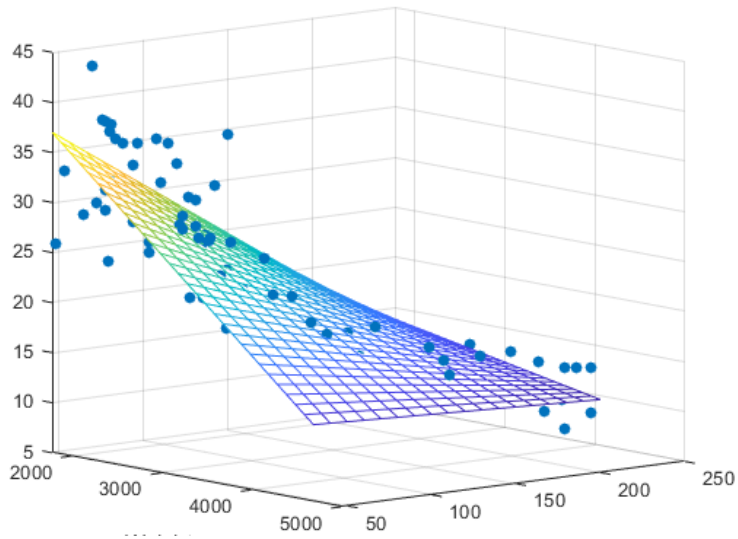
inputs

Simple Linear Regression	Multiple Linear Regression (2 Independent Variables (x_1, x_2))
	

Multiple Linear Regression.

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0.00632	18.0	2.31	0.0	0.538	6.575	65.2	4.0900	1.0	296.0	15.3	396.90	4.98	24.0
0.02731	0.0	7.07	0.0	0.469	6.421	78.9	4.9671	2.0	242.0	17.8	396.90	9.14	21.6
0.02729	0.0	7.07	0.0	0.469	7.185	61.1	4.9671	2.0	242.0	17.8	392.83	4.03	34.7
0.03237	0.0	2.18	0.0	0.458	6.998	45.8	6.0622	3.0	222.0	18.7	394.63	2.94	33.4
0.06905	0.0	2.18	0.0	0.458	7.147	54.2	6.0622	3.0	222.0	18.7	396.90	5.33	36.2
0.02985	0.0	2.18	0.0	0.458	6.430	58.7	6.0622	3.0	222.0	18.7	394.12	5.21	28.7
0.08829	12.5	7.87	0.0	0.524	6.012	66.6	5.5605	5.0	311.0	15.2	395.60	12.43	22.9
0.14455	12.5	7.87	0.0	0.524	6.172	96.1	5.9505	5.0	311.0	15.2	396.90	19.15	27.1
0.21124	12.5	7.87	0.0	0.524	5.631	100.0	6.0821	5.0	311.0	15.2	386.63	29.93	16.5
0.17004	12.5	7.87	0.0	0.524	6.004	85.9	6.5921	5.0	311.0	15.2	386.71	17.10	18.9

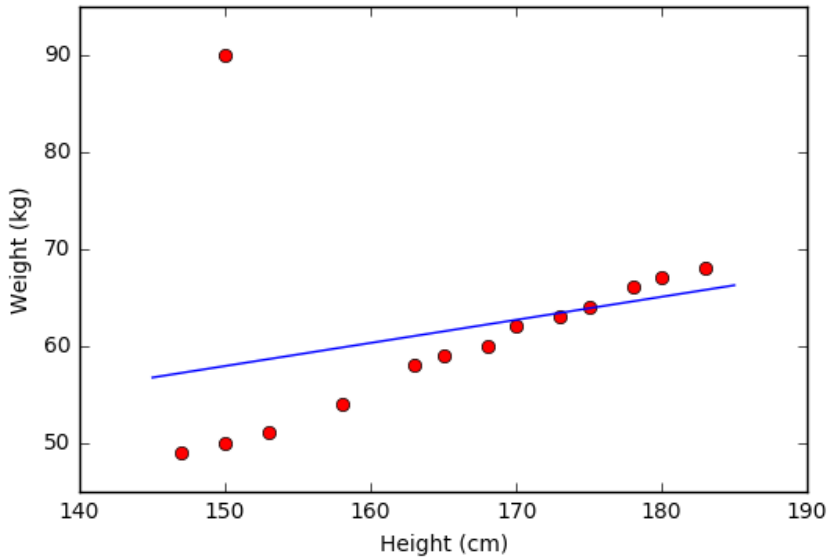
$$\hat{y}_{MEDV} = f(x) = b_0 + b_1 * X_{CRIM} + b_2 * X_{ZN} + b_3 * X_{INDUS} + b_4 * X_{CHAS} + b_5 * X_{NOX} + b_6 * X_{RM} + b_7 * X_{AGE} + b_8 * X_{DIS} + b_9 * X_{RAD} + b_{10} * X_{TAX} + b_{11} * X_{PTRATIO} + b_{12} * X_B + b_{13} * X_{LSTAT}$$



Nhược điểm của hồi quy tuyến tính

- Linear Regression **rất nhạy cảm với nhiễu** (sensitive to noise). Vi vậy trước khi thực hiện Linear Regressiong, các giá trị ngoại lai (outlier) cần phải được loại bỏ.

- Linear Regression **không biểu diễn được các mô hình phức tạp.**



THỰC HÀNH 3.3

Yêu cầu 1:

- Sinh viên tìm hiểu về tập dữ liệu mẫu Diabetes Dataset của Sklearn (xác định các features đầu vào (input) và label đầu ra (target))

Data Set Characteristics:

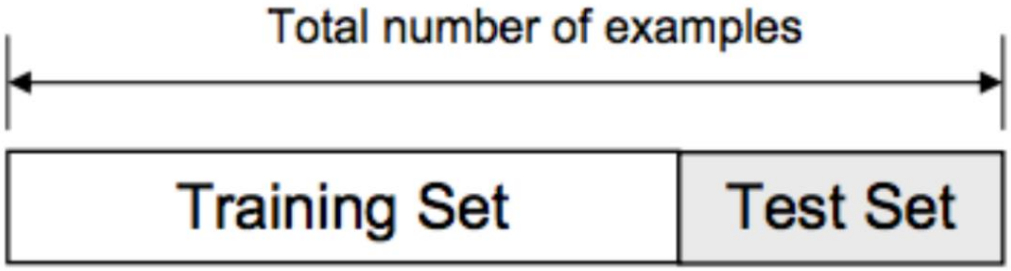
Number of Instances:	442
Number of Attributes:	First 10 columns are numeric predictive values
Target:	Column 11 is a quantitative measure of disease progression one year after baseline
Attribute Information:	• age age in years • sex • bmi body mass index • bp average blood pressure • s1 tc, T-Cells (a type of white blood cells) • s2 ldl, low-density lipoproteins • s3 hdl, high-density lipoproteins • s4 tch, thyroid stimulating hormone • s5 ltg, lamotrigine • s6 glu, blood sugar level

Samples total	442
Dimensionality	10
Features	real, -.2 < x < .2
Targets	integer 25 - 346

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html#sklearn.datasets.load_diabetes

Yêu cầu 2:

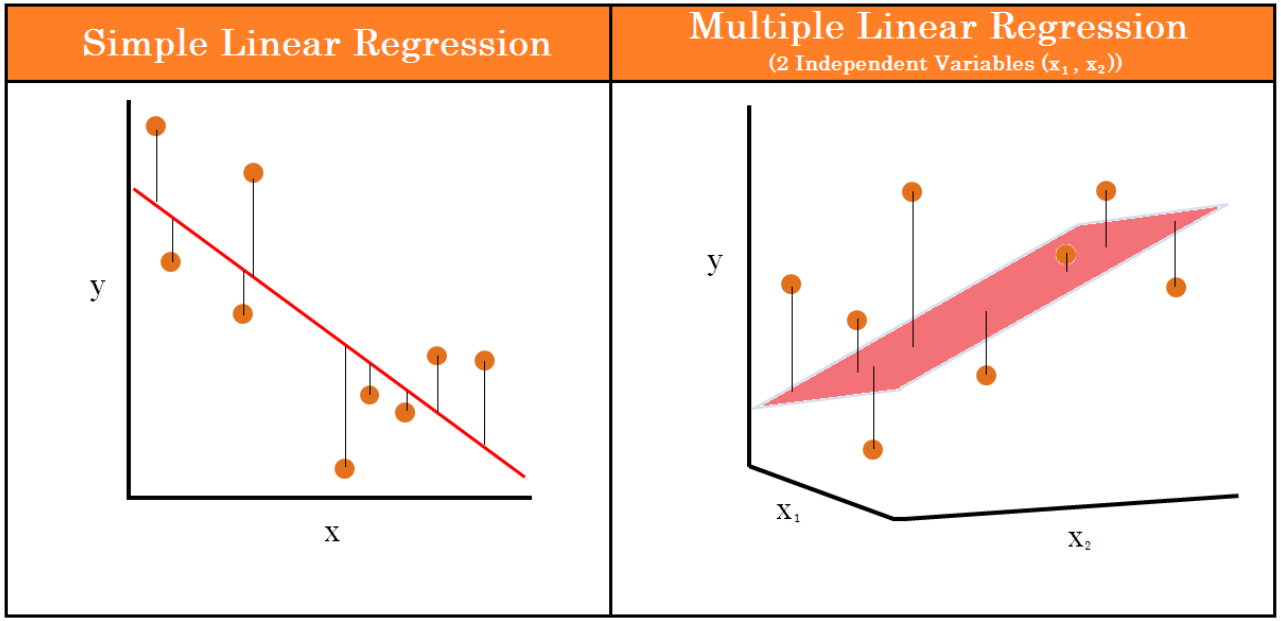
- Trong tập dữ liệu Diabetes xác định thuộc tính có ảnh hưởng lớn nhất (hệ số tương quan cao nhất) tới thuộc tính target.
- Tách tập dữ liệu thành 2 phần Train – Test với tỷ lệ 75%-25%



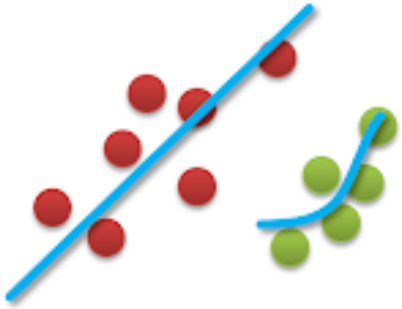
Yêu cầu 3:

Xây dựng mô hình hồi quy tuyến tính đơn giản (Simple Linear Regression) với thuộc tính có ảnh hưởng cao nhất tới thuộc tính Target. Xác định sai số RMSE và R^2 trên tập Train và Test.

Xây dựng mô hình hồi quy tuyến tính với tất các thuộc tính đầu vào (input). Xác định sai số RMSE và R^2 trên tập Train và Test.

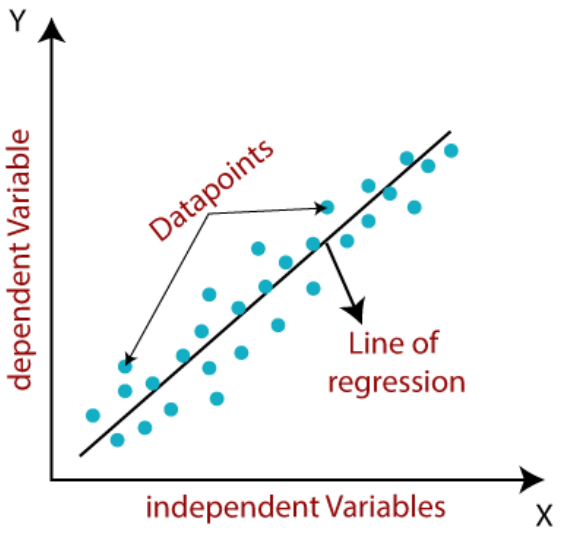


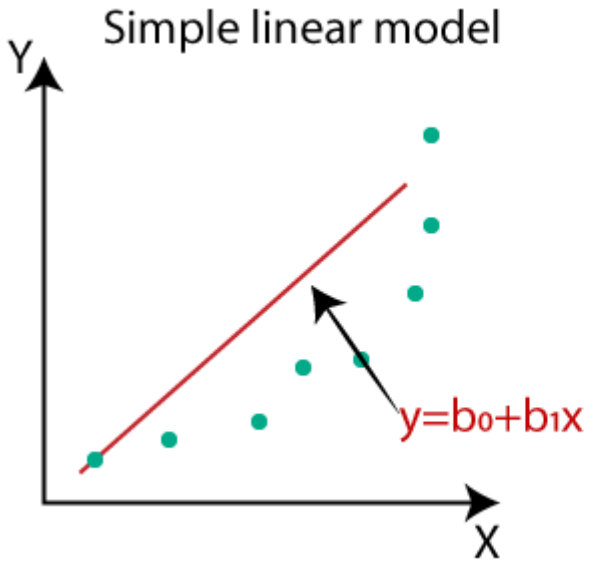
6.2 Hồi quy đa thức (Polynomial Regression)

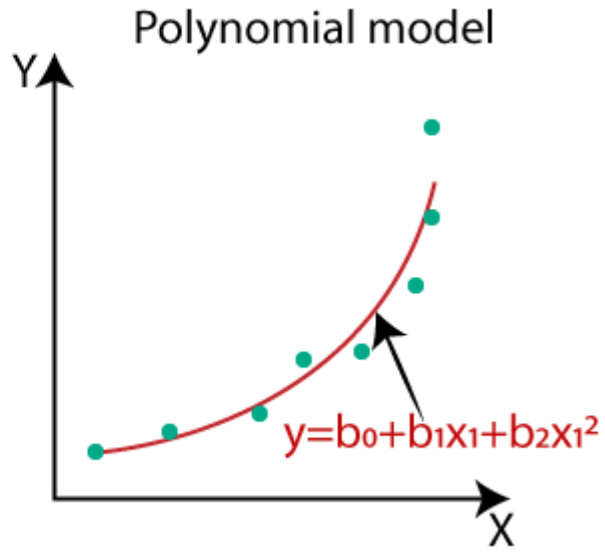


Hồi quy đa thức

Trong trường hợp dữ liệu không tuyến tính việc áp dụng mô hình tuyến tính sẽ không hiệu quả tỷ lệ lỗi cao, độ chính xác giảm.





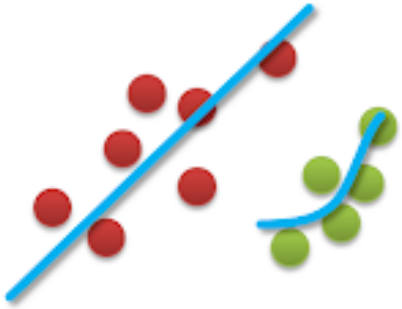


Hồi quy đa thức bậc n của biến độc lập x_1

Polynomial Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

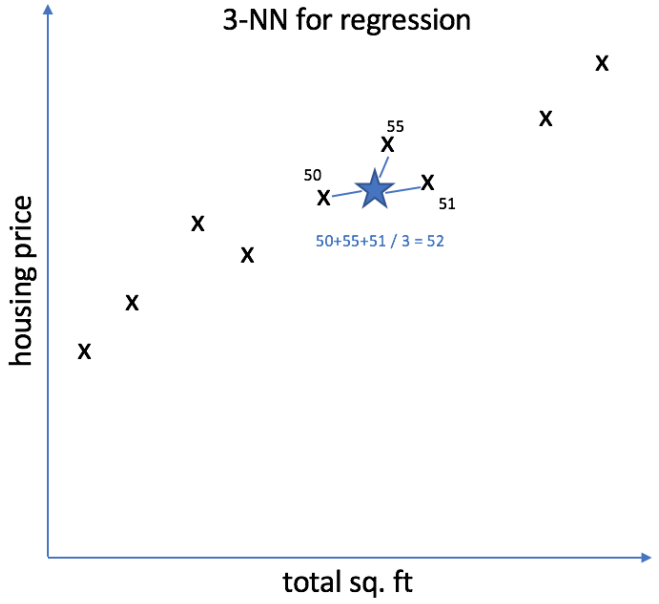
6.3 KNN cho bài toán Hồi quy

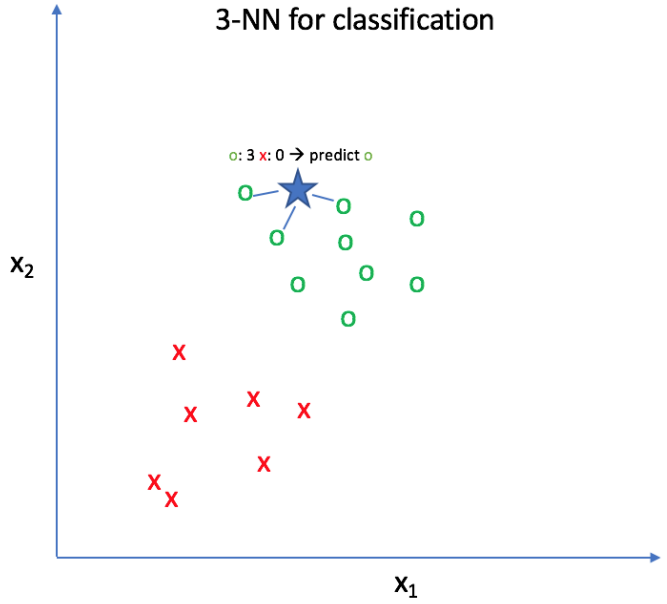


KNN

Tương tự như đối với bài toán phân lớp. Xác định những điểm dữ liệu gần nhất với điểm dữ liệu mới.

Nhãn của điểm dữ liệu mới được là nhãn của điểm dữ liệu đã biết gần nhất (K=1) hoặc trung bình có trọng số của những điểm gần nhất.

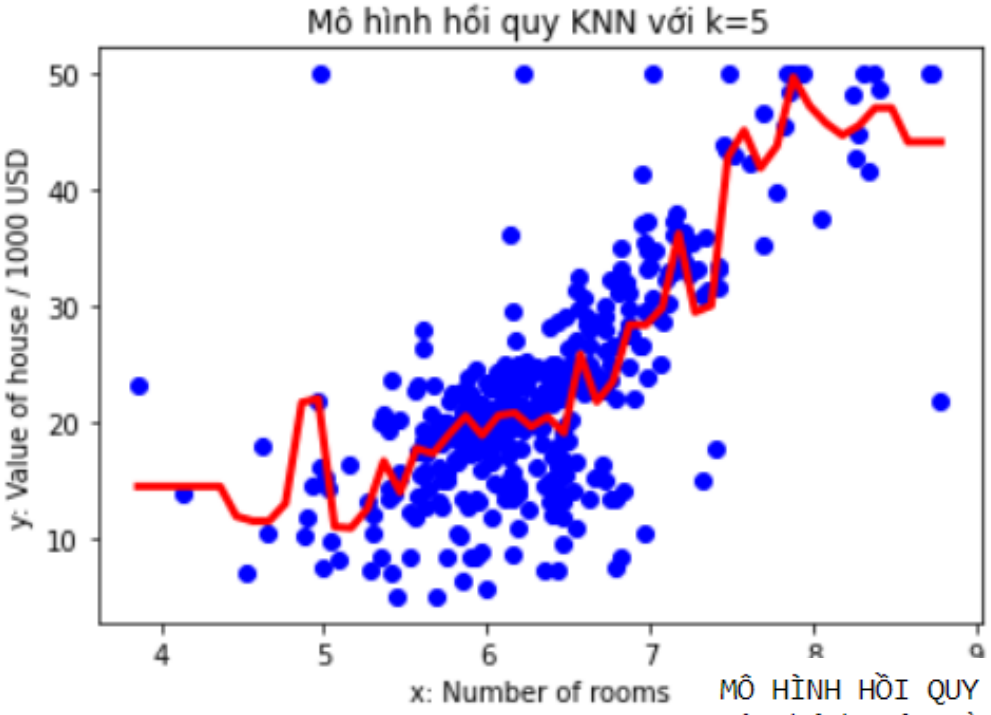




KNN

Dự đoán giá nhà với 1 biến độc lập – RM (số phòng trung bình của căn nhà)

RM	MEDV
6.575	24.0
6.421	21.6
7.185	34.7
6.998	33.4
7.147	36.2
6.430	28.7
6.012	22.9
6.172	27.1
5.631	16.5
6.004	18.9
6.377	15.0




```
MÔ HÌNH HỒI QUY KNN SỬ DỤNG 1 BIẾN ĐỘC LẬP-RM
Độ chính xác của mô hình trên tập huấn luyện:
--------------------------------------------
Sai số RMSE 5.166827358301712
Sai số R2   0.6906745137827078


Độ chính xác của mô hình trên tập kiểm thử:
--------------------------------------------
Sai số RMSE 6.498274280718765
Sai số R2   0.4508457090315743
```

THỰC HÀNH 3.4

Yêu cầu 1: (Đã làm trong TH 3.3)

- Sinh viên tìm hiểu về tập dữ liệu mẫu Diabetes Dataset của Sklearn (xác định các features đầu vào (input) và label đầu ra (target))

Data Set Characteristics:

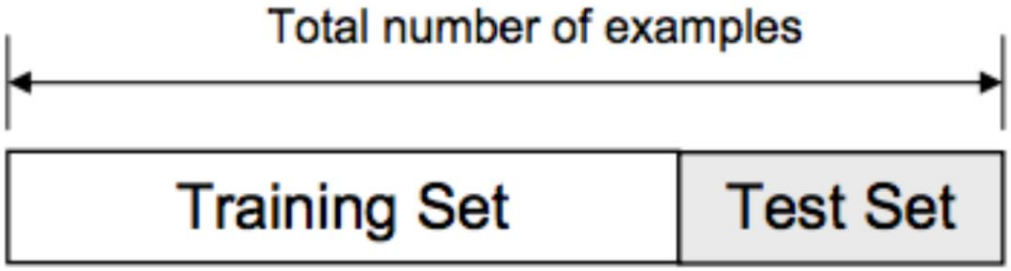
Number of Instances:	442
Number of Attributes:	First 10 columns are numeric predictive values
Target:	Column 11 is a quantitative measure of disease progression one year after baseline
Attribute Information:	• age age in years • sex • bmi body mass index • bp average blood pressure • s1 tc, T-Cells (a type of white blood cells) • s2 ldl, low-density lipoproteins • s3 hdl, high-density lipoproteins • s4 tch, thyroid stimulating hormone • s5 ltg, lamotrigine • s6 glu, blood sugar level

Samples total	442
Dimensionality	10
Features	real, -.2 < x < .2
Targets	integer 25 - 346

https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_diabetes.html#sklearn.datasets.load_diabetes

Yêu cầu 2: (Đã làm trong TH 3.3)

- Trong tập dữ liệu Diabetes xác định thuộc tính có ảnh hưởng lớn nhất (hệ số tương quan cao nhất) tới thuộc tính target.
- Tách tập dữ liệu thành 2 phần Train – Test với tỷ lệ 75%-25%



Yêu cầu 3:

1) Xây dựng mô hình KNN cho bài toán hồi quy (Simple Linear Regression) với thuộc tính có ảnh hưởng cao nhất tới thuộc tính Target. Xác định sai số RMSE và R^2 trên tập Train và Test.

2) Xây dựng mô hình KNN cho bài toán hồi quy với tất các thuộc tính đầu vào (input). Xác định sai số RMSE và R^2 trên tập Train và Test.


HANOI UNIVERSITY OF MINING AND GEOLOGY
KHOA
CÔNG NGHỆ THÔNG TIN
FACULTY OF INFORMATION TECHNOLOGY

Thank you!