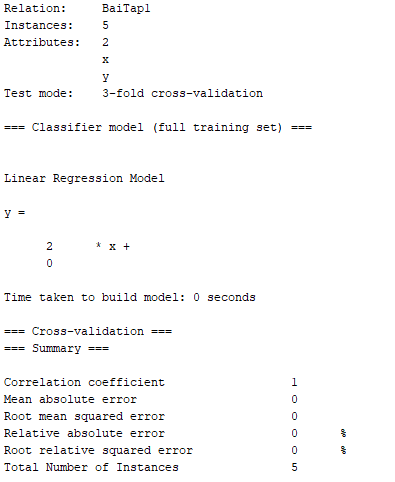
# **BÁO CÁO THỰC HÀNH SỐ 6**

# Hồi quy tuyến tính

### Bài tập 1



Phương trình hồi quy tìm được là: y = 2\*x

Hệ số tương quan: 0

Sai số tuyệt đối trung bình: 0

Lỗi bình phương gốc: 0

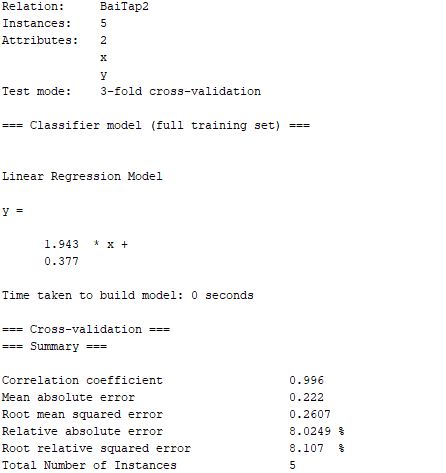
Sai số tuyệt đối tương đối: 0%

Lỗi bình phương tương đối gốc: 0%

Tổng số mẫu: 5

***KL: Hệ số tương quan bằng 0 nên không có mối quan hệ tuyến tính chuyển động của hai biến y và x***

### Bài tập 2



Phương trình hồi quy tìm được là: y = 1,943\*x + 0,377

Hệ số tương quan: 0,996

Sai số tuyệt đối trung bình: 0,222

Lỗi bình phương gốc: 0,2607

Sai số tuyệt đối tương đối: 8.0249%

Lỗi bình phương tương đối gốc: 8.107%

Tổng số mẫu: 5

***KL: Hệ số tương quan bằng 0,996 cho biết một sự tương quan thuận giữa hai biến, nghĩa là nếu giá trị của biến này tăng thì sẽ làm tăng giá trị của biến kia.***

### Computer Hardware

* 1. Mô tả dữ liệu

Đặc điểm tập dữ liệu: Đa biến

Đặc điểm thuộc tính: Số nguyên

Bài toán: Hồi quy

Số mẫu: 209

Số lượng thuộc tính: 9

Giá trị thiếu: 0

Nguồn:

Tác giả: Phillip Ein-Dor and Jacob Feldmesser Ein-Dor: Faculty of Management Tel Aviv University; Ramat-Aviv; Tel Aviv, 69978; Israel

Tài trợ: David W. Aha (aha '@' ics.uci.edu) (714) 856-8779

Tên nhà cung cấp: 30

adviser, amdahl,apollo, basf, bti, burroughs, c.r.d, cambex, cdc, dec,

dg, formation, four-phase, gould, honeywell, hp, ibm, ipl, magnuson,

microdata, nas, ncr, nixdorf, perkin-elmer, prime, siemens, sperry,

sratus, wang

Tên model: many unique symbols

MYCT: thời gian chu kỳ máy tính bằng nano giây (số nguyên)

MMIN: bộ nhớ chính tối thiểu tính bằng kilobyte (số nguyên)

MMAX: bộ nhớ chính tối đa tính bằng kilobyte (số nguyên)

CACH: bộ nhớ đệm tính bằng kilobyte (số nguyên)

CHMIN: kênh tối thiểu theo đơn vị (số nguyên)

CHMAX: kênh tối đa theo đơn vị (số nguyên)

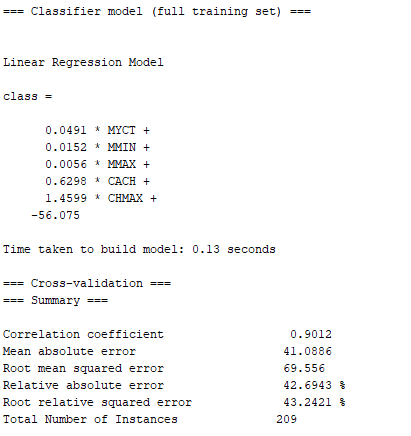
PRP: hiệu suất tương đối được công bố (số nguyên)

ERP: ước tính hiệu suất tương đối từ bài báo gốc (số nguyên)

* 1. Mô tả bài toán

Bài toán hồi quy tuyến tính tìm mối liên quan giữa PRP( class) với MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX

* 1. Kết quả và đánh giá



Phương trình hồi quy tuyến tính:

class = 0.0491 \* MYCT + 0.0152 \* MMIN + 0.0056 \* MMAX + 0.6298 \* CACH + 1.4599 \* CHMA - 56.075

Hệ số tương quan: 0,9012

Sai số tuyệt đối trung bình: 41,0886

Lỗi bình phương gốc: 69,556

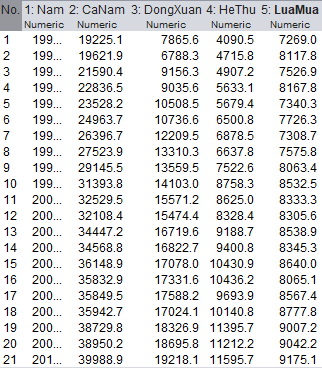
Sai số tuyệt đối tương đối: 42,6943%

Lỗi bình phương tương đối gốc: 43,2421%

Tổng số mẫu: 209

### Dự đoán sản lượng lúa

#### Mô tả dữ liệu



Bài toán: hồi quy

Kiểu dữ liệu: numeric

Số thuộc tính: 5

Số lượng mẫu: 21

Nguồn: dữ liệu thu thập thông tin: Số liệu này được lấy từ trang web của Tổng cục Thống kê và Bộ nông nghiệp Phát triển Nông thôn ngày 1/1/2012

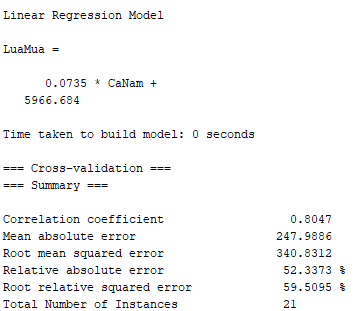
#### Mô tả bài toán

Bài toán 1: hồi quy tuyến tính tìm mối liên quan lúa mùa trong cả năm

Bài toán 2: hồi quy tuyến tính tìm mối liên quan sản lượng lúa các vụ qua cả năm

#### Kết quả và đánh giá

##### Bài toán 1



Phương trình hồi quy tìm được là: LuaMua = 0,0735\*CaNam + 5966,684

Hệ số tương quan: 0,8047

Sai số tuyệt đối trung bình: 247,9886

Lỗi bình phương gốc: 340,8312

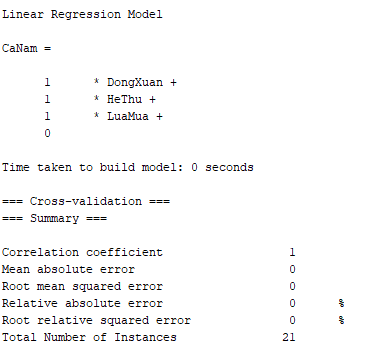
Sai số tuyệt đối tương đối: 52,3373%

Lỗi bình phương tương đối gốc: 59,5095%

Tổng số mẫu: 21

***KL: Hệ số tương quan bằng 0,8047 cho biết một sự tương quan thuận giữa hai biến, nghĩa là nếu giá trị lúa cả năm tăng thì giá trị lúa mùa cũng sẽ tăng***

##### Bài toán 2



Phương trình hồi quy tìm được là: CaNam = DongXuan + HeThu + LuaMua

Hệ số tương quan: 1

Sai số tuyệt đối trung bình: 0

Lỗi bình phương gốc: 0

Sai số tuyệt đối tương đối: 0%

Lỗi bình phương tương đối gốc: 0%

Tổng số mẫu: 21

***KL: Hệ số tương quan bằng 1 tuyệt đối cho biết một sự tương quan thuận giữa hai biến, nghĩa là nếu giá trị lúa các vụ đông xuân, hè thu, lúa mùa tăng thì sản lượng lúa của cả năm sẽ tăng***

# Mạng nơ ron

## **Dữ liệu Iris, so sánh MLP với 3 giải thuật đã học**

#### Mô tả dữ liệu

| ***Missing*** | ***Mẫu*** | ***Lớp*** | ***Thuộc tính*** | ***Kiểu dữ liệu*** |
| --- | --- | --- | --- | --- |
| 0% | 150 | 3 | 5 | Numeric |
|  |  | Iris Setosa |  |  |
|  |  | Versicolour |  |  |
|  |  | Virginica |  |  |
|  | **sepal,length** | **sepal,width** | **petal,length** | **petal,width** |
| **min** | 4,3 | 2 | 1 | 0,1 |
| **max** | 7,9 | 4,4 | 6,9 | 2,5 |
| **mean** | 5,84 | 3,06 | 3,76 | 1,20 |
| **std** | 0,83 | 0,44 | 1,77 | 0,76 |

#### So sánh thuật toán

|  |  | Thuật toán | | | |
| --- | --- | --- | --- | --- | --- |
| Độ đo | Tham số | MLP | SMO | IBK | J48 |
| Tỷ lệ huấn luyện/kiểm thử là 66/34 | Số mẫu phân lớp đúng | 50 | 49 | 49 | 49 |
| Số mẫu phân lớp sai | 1 | 2 | 2 | 2 |
| % Phân lớp đúng | 98.0392 % | 96.0784 % | 96.0784 % | 96.0784 % |
| % Phân lớp sai | 1.9608 % | 3.9216 % | 3.9216 % | 3.9216 % |
| Số lớp ẩn | 1 lớp ẩn có 3 nơ ron | Không có | | |
| Tốc độ học | 0,3 |
| Động lượng | 0,2 |
| batchsize | 100 |
| epoch | 500 |
| Ma trận nhầm lẫn | a b c <-- classified as  15 0 0 | a = Iris-setosa  0 19 0 | b = Iris-versicolor  0 1 16 | c = Iris-virginica | a b c <-- classified as  15 0 0 | a = Iris-setosa  0 19 0 | b = Iris-versicolor  0 2 15 | c = Iris-virginica | | |
|
|
|

Nhận xét: thuật toán MLP cho kết quả phân lớp tốt nhất

## **Dữ liệu breast-cancer, so sánh MLP với J48 và IBK**

#### Mô tả dữ liệu

Link data: <https://archive.ics.uci.edu/ml/datasets/breast+cancer>

* Nguồn: Matjaz Zwitter & Milan Soklic (physicians) Institute of Oncology University Medical Center Ljubljana, Yugoslavia.
* Kiểu dữ liệu: nominal
* Số thuộc tính: 9
* Số lượng mẫu: 286
* Missing Values: yes
* Số lượt truy cập: 561671
* Lĩnh vực: Y tế
* Đây là một trong ba tên miền do Viện Ung thư cung cấp, đã nhiều lần xuất hiện trong tài liệu học máy. (Xem thêm nội dung hạch bạch huyết và khối u nguyên phát.) Tập dữ liệu này bao gồm 201 cá thể của một lớp và 85 thể hiện của một lớp khác. Các trường hợp được mô tả bằng 9 thuộc tính, một số thuộc tính tuyến tính và một số thuộc tính danh nghĩa
* Thông tin thuộc tính

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.

5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.

6. node-caps: yes, no.

7. deg-malig: 1, 2, 3.

8. breast: left, right.

9. breast-quad: left-up, left-low, right-up, right-low, central.

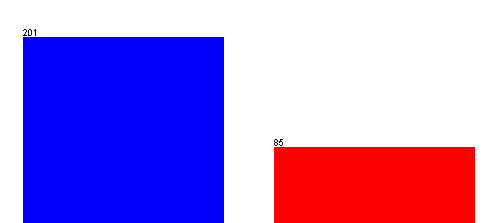
10. irradiat: yes, no.

#### Tiền xử lý dữ liệu

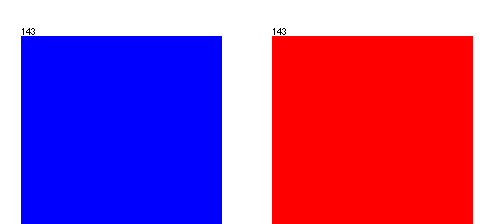
Dữ liệu không cần bằng giữa hai lớp phát bệnh và không phát bệnh

Vì vậy em sẽ tăng dữ liệu ở lớp không phát bệnh lên bằng kỹ thuật ClassBalancer đơn giản là tái lại trọng số sao cho tổng trọng số của tất cả các mẫu của lớp trong dữ liệu là như nhau. Không có trường hợp nào bị xóa hoặc thêm, vì vậy tổng số lượng mẫu của hai lớp vẫn không thay đổi.

Trước khi chuẩn hoá



Sau khi chuẩn hoá



#### So sánh thuật toán

|  | Test mode: 10-fold cross-validation | | |
| --- | --- | --- | --- |
| Tham số | J48 | ANN | KNN |
| Đặc trưng | cây đã được tỉa | 1 lớp ẩn có 26 nơ ron | k = 11 (tốt nhất) |
| Số lá: 27 | tốc độ học: 0,3 |  |
| Kích thước cây: 31 | quán tính: 0,2 |  |
|  | batchsize: 100 |  |
|  | epoch: 500 |  |
| Số mẫu phân lớp đúng | 170 | 176 | 190 |
| Số mẫu phân lớp sai | 116 | 110 | 96 |
| Thống kê kappa | 0,187 | 0,2336 | 0,3272 |
| Sai số trung bình tuyệt đối | 0,4728 | 0,3825 | 0,4317 |
| Lỗi bình phương trung bình gốc | 0,5192 | 0,5838 | 0,4735 |
| Sai số tuyệt đối tương đối | 94,54% | 76,4855% | 86,331 % |
| Ma trận nhầm lẫn | a b <-- classified as  102.45 40.55 | a = no-recurrence-events  75.71 67.29 | b = recurrence-events | a b <-- classified as  107.43 35.57 | a = no-recurrence-events  74.02 68.98 | b = recurrence-events | a b <-- classified as  112.41 30.59 | a = no-recurrence-events  65.61 77.39 | b = recurrence-events |
|
|

| Thuật toán | Tên phương pháp | TP Rate | FP Rate | Precision | Recall | F - Measure | ROC Area | Class |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| DT | J48 | 0,594 | 0,406 | 0,6 | 0,594 | 0,587 | 0,565 | Weighted Avg |
| ANN | MultilayerPerceptron | 0,751 | 0,518 | 0,592 | 0,751 | 0,662 | 0,646 | no-recurrence-events |
|  | 0,482 | 0,249 | 0,66 | 0,482 | 0,557 | 0,646 | recurrence-events |
|  | 0,617 | 0,383 | 0,626 | 0,617 | 0,61 | 0,646 | Weighted Avg |
| KNN | IBk | 0,664 | 0,336 | 0,674 | 0,664 | 0,659 | 0,692 | Weighted Avg |

KL: Với bộ dữ liệu này khi xử lý cân bằng dữ liệu bằng class balance và thực hiện thử nghiệm trên một vài thuật toán thì ta thu được kết quả trên. Thuật toán tốt nhất trong trường hợp này là KNN với k = 11

## 