

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**ĐHQG TP.HCM**

-----🙞🙜🕮🙞🙜-----

**ĐỒ ÁN THỰC HÀNH**

**KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

*Chủ đề:* Tiếp thị ngân hàng

*Lớp:* CS313.K11.KHTN.1

*Giảng viên:* Ths. Nguyễn Thị Anh Thư

*Nhóm thực hiện:*

17520689 – Nguyễn Thùy Linh

17520210 – Lý Hồng Thiên Ân

17520730 – Phạm Viết Lưu

*Tp Hồ Chí Minh, ngày 20 tháng 10 năm 2019*.

Mục lục

[I. BANK MARKETING DATA SET 2](#_Toc23712380)

[1. bank-additional-full.csv 3](#_Toc23712381)

[2. Mục tiêu 4](#_Toc23712382)

[II. TIỀN XỬ LÝ DỮ LIỆU 4](#_Toc23712383)

[III. PHƯƠNG PHÁP SỬ DỤNG 4](#_Toc23712384)

[1. K-Nearest Neighbour 4](#_Toc23712385)

[2. Naïve Bayes 5](#_Toc23712386)

[3. Logistic Regression 6](#_Toc23712387)

[IV. KẾT QUẢ 7](#_Toc23712388)

[1. Thực nghiệm 1 7](#_Toc23712389)

[2. Thực nghiệm 2 8](#_Toc23712390)

[3. Thực nghiệm 3 10](#_Toc23712391)

# BANK MARKETING DATA SET

Dữ liệu liên quan đến các chiến dịch tiếp thi trực tiếp của một tổ chức ngân hàng Bồ Đào Nha. Các chiến dịch tiếp thị được thực hiện bằng các cuộc gọi điện thoại. Và thường có nhiều hơn một cuộc gọi cho cùng một khách hàng để kiểm tra xem khách hàng đó có gửi tiền có kỳ hạn vào ngân hàng hay không.

Trong Bank Marketing Data Set có 4 bộ dữ liệu:

(1) *bank-additional-full.csv* ( 41188 mẫu và 21 thuộc tính)

(2) *bank-additional.csv* (random 10% của bộ dữ liệu (1) )

(3) *bank-full.csv* ( phiên bản cũ của Data Set với ít thuộc tính hơn - 45211 mẫu và 18 thuộc tính)

(4) *bank.csv* (random 10% của bộ dữ liệu (3))

Trong đồ án lần này, chúng em thực hiện trên bộ dữ liệu (1).

## bank-additional-full.csv

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Thuộc tính** | **Kiểu** | **Mô tả** |
| 1 | age | numeric | Tuổi |
| 2 | job | nominal | Nghề nghiệp |
| 3 | marital | nominal | Tình trạng hôn nhân |
| 4 | education | nominal | Trình độ học vấn |
| 5 | default | nominal | Có thẻ tín dụng hay không |
| 6 | housing | nominal | Có khoản cho vay nhà ở hay không |
| 7 | loan | nominal | Có khoản cho vay cá nhân hay không |
| 8 | contact | nominal | Phương tiện liên lạc |
| 9 | month | nominal | Tháng liên lạc cuối cùng trong năm |
| 10 | day\_of\_week | nominal | Ngày liên lạc cuối cùng trong tuần |
| 11 | duration | numeric | Thời gian liên lạc |
| 12 | campaign | numeric | Số liên lạc được thực hiện trong chiến dịch |
| 13 | pdays | numeric | Số ngày trôi qua kể từ khi khách hàng được liên lạc lần cuối trong chiến dịch trước |
| 14 | previous | numeric | Số liên lạc được thực hiện trước chiến dịch này |
| 15 | poutcome | nominal | Kết quả của chiến dịch tiếp thị trước |
| 16 | emp.var.rate | numeric | Tỷ lệ biến động việc làm – chỉ số hàng quý |
| 17 | cons.price.idx | numeric | Chỉ số giá tiêu dùng – chỉ tiêu hàng tháng |
| 18 | cons.conf.idx | numeric | Chỉ số niềm tin tiêu dùng – chỉ số hàng tháng |
| 19 | euribor3m | numeric | Lãi suất euribor 3 tháng |
| 20 | nr.employed | numeric | Số lượng nhân công – chỉ số hàng quý |
| **21** | **y** | **nominal** | **Khách hàng có gửi tiền có kì hạn hay không** |

## Mục tiêu

Xây dựng mô hình dự đoán một khách hàng có tham gia gửi tiền có kỳ hạn vào ngân hàng hay không.

# TIỀN XỬ LÝ DỮ LIỆU

Bộ dữ liệu ban đầu có 21 thuộc tính (bao gồm cả thuộc tính đầu ra) với tỉ lệ nominal và numeric là 11:10. Để thuân tiện cho việc huấn luyện mô hình, chúng em tiến hành chuẩn hóa dữ liệu như sau:

* Với những thuộc tính numeric, tiến hành chuẩn hóa Min-max theo công thức:
* Với thuộc tính đầu ra, chúng em quy định giá trị ‘y’ và ‘no’ tương ứng với 1 và 0.
* Với những thuộc tính nominal, tìm những giá trị phân biệt và thay thế bằng chỉ số tương ứng, sau đó chuẩn hóa Min-max theo công thức trên.

**VD**: Thuộc tính *martial* có 4 giá trị phân biệt là *‘divorced’*, *‘married’*, *‘single’* và *‘unknown’*:

* Thay những giá trị *‘divorced’*, *‘married’*, *‘single’*, *‘unknown’* tương ứng bằng 0, 1, 2, 3.
* Chuẩn hóa Min-max: thay những giá trị 0, 1, 2, 3 tương ứng bằng 0, , , 1.

# PHƯƠNG PHÁP SỬ DỤNG

## K-Nearest Neighbour

**Đặc trưng:**

1 – Là một trong những thuật toán *supervised-learning* đơn giản nhất.

2 – Thuộc vào loại *lazy learning*.

3 – Label của một điểm dữ liệu mới được suy ra trực tiếp từ K điểm dữ liệu gần nhất trong training set.

4 – Không quan tâm đến việc có một vài điểm dữ liệu trong những điểm gần nhất này là nhiễu.

**Thuật toán:**

***B1***: Tính khoảng cách giữa điểm dữ liệu cần dự đoán và tất cả các điểm trong tập huấn luyện, khoảng cách có thể tính bằng cách lấy *norm 2*:

***B2***: Chọn k điểm gần với điểm dữ liệu cần dự đoán nhất trong tập huấn luyện và đánh trọng số cho mỗi điểm đó bằng cách lấy nghịch đảo khoảng cách vừa tính (). Trong trường hợp điểm dữ liệu cần dự đoán trùng với 1 điểm dữ liệu trong tập huấn luyện, tức khoảng cách *d(x’, x)* bằng 0, ta lấy luôn label của điểm dữ liệu huấn luyện đó.

***B3***: Gán điểm dữ liệu cần dự đoán vào lớp có tổng trọng số cao nhất vừa tính.

## Naïve Bayes

**Đặc trưng:**

1 – Thuộc nhóm *supervised-learning*.

2 – Giả thiết “*naïve*” – giả thiết về sự độc lập của các chiều dữ liệu.

3 – Dựa trên định lý Bayes trong lý thuyết xác suất.

4 – Đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê.

**Thuật toán:**

***B1***: Dựa vào tập dữ liệu huấn luyện, tính các phân phối *p(c)* và *p(xi|c)*. Để tránh trường hợp giá trị *p(xi|c) = 0* do không có mẫu nào trong dữ liệu huấn luyện thỏa mãn tử số, ta sử dụng phương pháp *Laplace smoothing*:

***B2***: Xác định lớp *c* của điểm dữ liệu mới *x* với:

## Logistic Regression

**Đặc trưng:**

1 – Mặc dù có tên là *Regression*, *Logistic Regression* lại được sử dụng nhiều trong các bài toán phân lớp (nhị phân).

2 – *Logistic function* là một đường cong S nhận các giá trị thực và ép chúng về các giá trị trong khoảng (0;1), phổ biến nhất là *Sigmoid function*.

3 – *Boundary* tạo bởi *Logistic Regression* có dạng tuyến tính: ***y = wTx***

4 – Phù hợp với loại dữ liệu mà 2 class là gần với *linearly separable*.

**Thuật toán:**

***B0***: Khởi tạo bộ trọng số *w*, learning rate , threshold

***B1***: Tính *z = wTx*

***B2***: Tính sigmoid

***B3***: Tính độ lỗi

***B4***: Cập nhật trọng số *w* bằng phương pháp *Stochastic Gradient Descent*:

* Điều kiện dừng: (wnew - wold)< threshold

# KẾT QUẢ

## Thực nghiệm 1

Dữ liệu huấn luyện: 37069 mẫu (90%)

Dữ liệu kiểm tra 1: 412 mẫu (10% của dữ liệu kiểm tra)

Dữ liệu kiểm tra 2: 1236 mẫu (30% của dữ liệu kiểm tra)

Dữ liệu kiểm tra 3: 2471 mẫu (60% của dữ liệu kiểm tra)

Kết quả thực nghiệm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **Kiểm tra** | **Precision** | **Recall** | **F-measure** |
| **K-Nearest Neighbour** | 1 | 0.87 | 0.89 | 0.87 |
| 2 | 0.86 | 0.88 | 0.86 |
| 3 | 0.86 | 0.88 | 0.87 |
| **Naïve Bayes** | 1 | 0.85 | 0.88 | 0.83 |
| 2 | 0.86 | 0.88 | 0.85 |
| 3 | 0.87 | 0.89 | 0.85 |
| **Logistic Regression** | 1 | 0.88 | 0.89 | 0.87 |
| 2 | 0.87 | 0.89 | 0.86 |
| 3 | 0.87 | 0.89 | 0.86 |

Biểu đồ theo F1-measure:

Nhận xét:

* Từ biểu đồ F-Measure ta thấy, *Naïve Bayes* cho kết quả thấp nhất trong 3 phương pháp đối với lần thực nghiệm này.
* Với *K-Nearest Neighbour*, khi dữ liệu kiểm tra càng nhiều, *precision* và *recall* đều giảm. Ngược lại, đối với *Naïve Bayes* thì *precision* và *recall* đều tăng. Trong khi, *Logistic Regression* cho kết quả là *precision* giảm và *recall* vẫn giữ nguyên.
* Tuy nhiên, mức độ chênh lệch giữa các con số khá nhỏ (<= 0.06) nên mọi đánh giá chỉ ở mức tương đối.
* Mặc dù vậy, cả 3 phương pháp đều cho kết quả khá tốt (nằm trong khoảng 0.83 – 0.89).

## Thực nghiệm 2

Dữ liệu huấn luyện: 28831 mẫu (70%)

Dữ liệu kiểm tra 1: 1236 mẫu (10% của dữ liệu kiểm tra)

Dữ liệu kiểm tra 2: 3707 mẫu (30% của dữ liệu kiểm tra)

Dữ liệu kiểm tra 3: 7414 mẫu (60% của dữ liệu kiểm tra)

Kết quả thực nghiệm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **Kiểm tra** | **Precision** | **Recall** | **F1-measure** |
| **K-Nearest Neighbour** | 1 | 0.86 | 0.88 | 0.86 |
| 2 | 0.86 | 0.88 | 0.87 |
| 3 | 0.87 | 0.89 | 0.87 |
| **Naïve Bayes** | 1 | 0.86 | 0.88 | 0.85 |
| 2 | 0.87 | 0.89 | 0.86 |
| 3 | 0.88 | 0.89 | 0.86 |
| **Logistic Regression** | 1 | 0.87 | 0.89 | 0.86 |
| 2 | 0.88 | 0.90 | 0.87 |
| 3 | 0.88 | 0.89 | 0.88 |

Biểu đồ theo F1-measure:

Nhận xét:

* Từ biểu đồ F-Measure ta thấy, *Naïve Bayes* cho kết quả thấp nhất trong 3 phương pháp đối với lần thực nghiệm này.
* Với cả 3 phương pháp, khi dữ liệu kiểm tra càng nhiều, *precision* và *recall* đều tăng.
* Tuy nhiên, mức độ chênh lệch giữa các con số khá nhỏ (<= 0.05) nên mọi đánh giá chỉ ở mức tương đối.
* Mặc dù vậy, cả 3 phương pháp đều cho kết quả khá tốt (nằm trong khoảng 0.85 – 0.90).
* Đặc biệt, khi giảm dữ liệu huấn luyện và tăng dữ liệu kiểm tra, thực nghiệm này lại cho kết quả cao hơn một chút so với lần thực nghiệm 1.

## Thực nghiệm 3

Dữ liệu huấn luyện: 20594 mẫu (50%)

Dữ liệu kiểm tra 1: 2059 mẫu (10% của dữ liệu kiểm tra)

Dữ liệu kiểm tra 2: 6178 mẫu (30% của dữ liệu kiểm tra)

Dữ liệu kiểm tra 3: 12357 mẫu (60% của dữ liệu kiểm tra)

Kết quả thực nghiệm:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Phương pháp** | **Kiểm tra** | **Precision** | **Recall** | **F1-measure** |
| **K-Nearest Neighbour** | 1 | 0.86 | 0.88 | 0.87 |
| 2 | 0.87 | 0.89 | 0.88 |
| 3 | 0.86 | 0.88 | 0.87 |
| **Naïve Bayes** | 1 | 0.86 | 0.88 | 0.84 |
| 2 | 0.88 | 0.90 | 0.87 |
| 3 | 0.88 | 0.89 | 0.86 |
| **Logistic Regression** | 1 | 0.87 | 0.89 | 0.86 |
| 2 | 0.89 | 0.90 | 0.88 |
| 3 | 0.88 | 0.90 | 0.87 |

Biểu đồ theo F1-measure:

Nhận xét:

* Từ biểu đồ F-Measure ta thấy, *Naïve Bayes* cho kết quả thấp nhất trong 3 phương pháp đối với lần thực nghiệm này.
* Khi dữ liệu huấn luyện và kiểm tra tương đương, cả 3 phương pháp đều cho kết quả dao động trong khoảng 0.84 – 0.90.