**BÀI TẬP 1:**

1. **Phương pháp Hồi quy tuyến tính**
2. **Phương pháp K-NN**

**Yêu cầu:** Gửi file mã nguồn (.py hoặc .ipnb) và báo cáo kết quả chương trình đi kèm.

**Bài 1:** Bảng dữ liệu về chiều cao và cân nặng của 15 người như dưới đây:

| **Chiều cao (cm)** | **Cân nặng (kg)** | **Chiều cao (cm)** | **Cân nặng (kg)** |
| --- | --- | --- | --- |
| 147 | 49 | 168 | 60 |
| 150 | 50 | 170 | 71 |
| 153 | 51 | 173 | 63 |
| 155 | 53 | 175 | 64 |
| 158 | 54 | 178 | 66 |
| 160 | 56 | 180 | 67 |
| 163 | 58 | 183 | 70 |
| 165 | 59 |  |  |

Hãy xây dựng đường hồi quy tuyến tính của cân nặng theo chiều cao? Dự đoán cân nặng của 1 người khi có chiều cao là 159cm?

**Bài 2:** Theo dõi mức lãi suất (X) và tỷ lệ lạm phát (Y) ở một số quốc gia có số liệu:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 17.5 | 15.6 | 9.8 | 5.3 | 7.9 | 10 | 19.2 | 13.1 |
| Y | 14.2 | 11.7 | 6.4 | 2.1 | 4.8 | 8.1 | 15.4 | 9.8 |

1. Hãy xây dựng đường hồi quy tuyến tính của X theo Y
2. Dự đoán tỷ lệ lạm phát khi mức lãi suất là 11.

**Bài 3:** Sử dụng bộ data dùng để phân tích chất lượng của rượu winequality-red.csv (http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv) của đại học California-Berkeley. Các column của data

|  |  |
| --- | --- |
| fixed acidity | Nồng độ axit tartaric |
| volatile acidity | Tính axit |
| citric acid | Nồng độ axit Citric |
| residual sugar | Nồng độ đường dư |
| chlorides | Nồng độ clo |
| free sulfur dioxide | Nồng độ acid sulfurous tự do |
| total sulfur dioxide | Nồng độ acid sulfurus |
| density | Mật độ m/v(khối lượng/đơn vị thể tích) |
| Ph | pH |
| sulphates | Nồng độ sunfat |
| alcohol | Nồng độ alc |
| Quality | Điểm đánh giá chất lượng rượu từ 0-10/td |

1. Hãy xây dựng đường hồi quy tuyến tính của “alcohol” theo “density”
2. Hãy xây dựng đường hồi quy tuyến tính của “quality” theo các biến còn lại.

#### Bài 4: Bộ Dữ liệu Nhà ở Boston bao gồm giá nhà ở những nơi khác nhau ở Boston. Cùng với giá cả, tập dữ liệu cũng cung cấp thông tin như Tội phạm (CRIM), các khu vực kinh doanh không-bán-lẻ ở thị trấn (INDUS), tuổi chủ sở hữu ngôi nhà (AGE) và có nhiều thuộc tính khác có sẵn ở đây. Bộ dữ liệu chính nó có thể down từ <https://archive.ics.uci.edu/ml/datasets/Housing>. Chúng ta sử dụng scikit-learn, chúng ta có thể import nó từ scikit-learn. Hãy đưa ra đường hồi quy tuyến tính dự đoán giá nhà ở Boston dựa trên các thuộc tính còn lại.

Bài 5: Bảng dữ liệu :

|  |  |  |
| --- | --- | --- |
| X1 = Acid Durability(seconds) | X2 = Strength(kg/square meter) | Y |
| 7 | 7 | Bad |
| 7 | 4 | Bad |
| 3 | 4 | Good |
| 1 | 4 | Good |

1. Phân lớp bằng tay cho A(3,7) với k=3
2. Kiểm tra kết quả với python với python.

**Bài 6**: Cho bảng dữ liệu về chương trình cho vay ưu đãi cho các đối tượng mua chung cư với thông tin về lương và thời gian làm việc như sau:

|  |  |  |
| --- | --- | --- |
| Lương | Thời gian làm việc | Cho vay |
| 4 | 3 | 1 |
| 5 | 2.5 | 1 |
| 8 | 1 | 1 |
| 4 | 2.5 | 1 |
| 8 | 0.1 | 0 |
| 7 | 0.15 | 0 |
| 4 | 1 | 0 |
| 5 | 0.8 | 0 |

1. Phân lớp bằng tay cho A(6, 0.3) với k=3
2. Kiểm tra kết quả với python với python.

**Bài 7:** Iris flower dataset là một bộ dữ liệu bao gồm thông tin của ba loại hoa Iris (một loài hoa lan) khác nhau: Iris setosa, Iris virginica và Iris versicolor. Mỗi loại có 50 bông hoa được đo với dữ liệu là 4 thông tin: chiều dài, chiều rộng đài hoa (sepal), và chiều dài, chiều rộng cánh hoa (petal). Mỗi điểm dữ liệu trong tập này chỉ là một vector 4 chiều). Tách 150 dữ liệu trong Iris flower dataset ra thành 2 phần, gọi là training set và test set. Sử dụng thuật toán KNN sẽ dựa vào thông tin ở training set để dự đoán xem mỗi dữ liệu trong test set tương ứng với loại hoa nào. Dữ liệu được dự đoán này sẽ được đối chiếu với loại hoa thật của mỗi dữ liệu trong test set để đánh giá hiệu quả của KNN.

Ví dụ giải bài 4

* Tính khoảng cách

|  |  |  |
| --- | --- | --- |
| X1 = Acid Durability(seconds) | X2 = Strength(kg/square meter) | K/c (3,7) |
| 7 | 7 | (7-3)2+(7-7)2=16 |
| 7 | 4 | 25 |
| 3 | 4 | 9 |
| 1 | 4 | 13 |

* Sắp xếp

|  |  |  |  |
| --- | --- | --- | --- |
| X1 = Acid Durability(seconds) | X2 = Strength(kg/square meter) | K/c (3,7) | Y |
| 3 | 4 | 9 | Good |
| 1 | 4 | 13 | Good |
| 7 | 7 | (7-3)2+(7-7)2=16 | Bad |
| 7 | 4 | 25 | Bad |

Với k=3 🡪 (3,5) vào lớp Good