Một số bài luyện tập

1. Mối quan hệ giữa tuổi, cân nặng và huyết áp ngưỡng trên (systolic blood pressure) của một số người được cho trong bảng sau (Nguồn : Data for Multiple Linear Regression) :

X1	X2	X3
132	52	173
143	59	184
153	67	194
162	73	211
154	64	196
168	74	220
137	54	188
149	61	188
159	65	207
128	46	167
166	72	217

Trong đó :

- X1 : huyết áp ngưỡng trên
- X2 : tuổi
- X3 : cân nặng (pound)

Xây dựng mô hình Linear Regression để ước tính huyết áp theo tuổi và cân nặng.

2. Mối quan hệ giữa tình trạng sức khỏe của một cộng đồng dân cư và các yếu tố xã hội được cho trong file dữ liệu Health.xls (Nguồn : Data for Multiple Linear Regression)

X1	X2	X3	X4	X5
8	78	284	9.100000381	109
9.300000191	68	433	8.699999809	144
7.5	70	739	7.199999809	113
8.899999619	96	1792	8.899999619	97
10.19999981	74	477	8.300000191	206
8.300000191	111	362	10.89999962	124
8.800000191	77	671	10	152
8.800000191	168	636	9.100000381	162
10.69999981	82	329	8.699999809	150
11.69999981	89	634	7.599999905	134
...	...	...	...	...

Trong đó :

- X1 : tỉ lệ tử vong trên 1000 cư dân

- X2 : số bác sĩ trên 100000 dân
- X3 : số bệnh viện trên 100000 dân
- X4 : thu nhập bình quân đầu người (x $1k/năm)
- X5 : mật độ dân cư (số người/mile^2)

Xây dựng mô hình Linear Regression để ước tính tỉ lệ tử vong của cộng đồng dân cư theo các yếu tố xã hội trong bảng.

3. "Customer churn" là tình trạng khách hàng của một công ty ngừng sử dụng sản phẩm/dịch vụ của công ty đó. Để tránh tình trạng mất khách hàng, các công ty thường xuyên theo dõi các thông tin về hoạt động của khách hàng để dự đoán trước những trường hợp có nguy cơ ngừng sử dụng sản phẩm/dịch vụ của họ, từ đó có các biện pháp chăm sóc phù hợp để giữ khách quen của mình.

Một bảng dữ liệu về tình hình khách hàng của một công ty được cho tại file Churn_Modelling.csv. (Nguồn : Build your First Deep Learning Neural Network Model using Keras in Python).

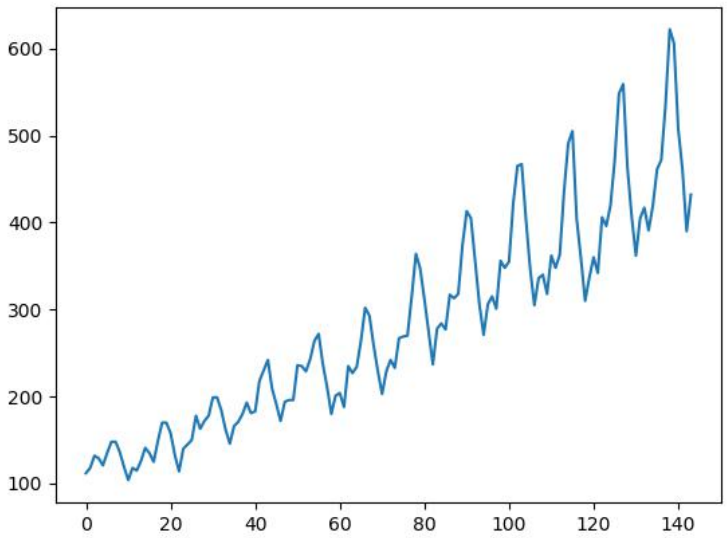
Trong bảng có các thông tin về từng khách hàng với các trường : tên, tuổi, giới tính, số dư tài khoản, ước tính thu nhập …, cột cuối cùng trong bảng dữ liệu cho biết khách hàng có quyết định ngừng dùng dịch vụ hay không (0/1).

Hãy xây dựng mô hình cho phép dự đoán nguy cơ khách hàng ngừng dùng dịch vụ dựa trên các trường thông tin đã có. Có thể sử dụng một trong các mô hình:

- Logistic Regression
- Decision Tree
- K-nearest Neighbours
- SVM
- Neural network

Lưu ý phân chia bảng dữ liệu thành các tập training và test để kiểm chứng độ chính xác mô hình trên tập dữ liệu test.

4. Lượng khách sử dụng hàng không trên toàn thế giới từ năm 1949 đến năm 1960 được cho trong file dữ liệu international-airline-passengers.csv. Đồ thị lượng hàng khách theo từng tháng được thể hiện trong hình dưới:



Lượng hành khách hàng không trên thế giới thống kê theo từng tháng (từ năm 1949-1960)

Hãy xây dựng mô hình mạng Neuron cho phép dự đoán trước lượng hành khách của thế giới trong vòng một năm. Cụ thể, dùng dữ liệu có từ năm 1949 đến năm 1959 để đưa ra ước đoán lượng hành khách trong 12 tháng của năm 1960.

Gợi ý:
Có nhiều cách xây dựng mô hình dự đoán với mức độ chính xác khác nhau. Ở ví dụ này, có thể dùng mô hình tương đối đơn giản sau:

- Chọn đầu vào :
  - Biến thời gian, tính theo từng tháng. Ví dụ tháng 1/1949 ~ t=0, tháng 2/1949 ~ t=1, tháng 1/1950 ~ t=12, tháng 2/1950 ~ t=13. Biến này đại diện cho sự tăng trưởng số hành khách theo thời gian
  - Tháng của năm, thể hiện bằng mã hóa *onehot*, gồm 12 giá trị 0/1. Tháng 1 tương ứng với (1,0, 0 , …0), tháng 12 tương ứng vơi (0, 0, …, 1). Lí do sử dụng mã hóa *onehot* là lượng hành khách giữa các tháng chênh lệch khá nhiều. Bộ 12 giá trị này đại diện cho sự tuần hoàn trong năm của số lượng hành khách
  - Giá trị trung bình của số lượng hành khách trong năm liền trước thời điểm đang xét. Ví dụ, với năm 1950 thì dùng số lượng hành khách trung bình trong năm 1949, năm 1951 thì dùng số lượng hành khách trung bình trong năm 1950. Giá trị này đại diện cho tính lịch sử của dữ liệu, theo đó giá trị hiện tại sẽ có xu hướng biến động xung quanh giá trị đã có trong quá khứ.

  Số đầu vào tổng cộng : 1 + 12 + 1 = 14

- Chọn mô hình:
  Đơn giản nhất là chọn mô hình mạng neuron 2 lớp, lớp ẩn có từ 10-50 neuron với hàm activation là relu, lớp cuối có một đầu ra với hàm activation là identity