Machine Learning

Bộ môn Khoa học dữ liệu
Khoa Công nghệ thông tin
Trường Đại học Công nghiệp thành phố Hồ Chí Minh-IUH

Bài 1:

Cho dữ liệu như sau:

Data =

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

Hãy viết code thực hiện các yêu cầu sau:

1.1 Tính trung bình từng chiều
1.2 Chuẩn hóa dữ liệu
1.3 Tính ma trận hiệp phương sai
1.4 Tính giá trị riêng và vector riêng
1.5 Giảm chiều dữ liệu xuống còn 1 chiều
1.6 Chuyển dữ liệu qua hệ trục mới và in kết quả

Bài 2:

Giải quyết bài toán ở trên sử dụng thư viện PCA và in ra kết quả

Bài 3:

Cho dữ liệu ở đường link sau:

https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data
Thông tin về dữ liệu này như sau:

a. Number of Attributes : 13

1) Alcohol	2) Malic acid	3) Ash
4) Alcalinity of ash	5) Magnesium	
6) Total phenols	7) Flavanoids	
8) Nonflavanoid phenols	9) Proanthocyanins	
10) Color intensity	11)Hue	
12) OD280/OD315 of diluted wines		13)Proline

b. Number of Instances

class 1 59
class 2 71
class 3 48

c. For Each Attribute:

All attributes are continuous
No statistics available, but suggest to standardise
variables for certain uses (e.g. for us with classifiers
which are NOT scale invariant)
NOTE: 1st attribute is class identifier (1-3)

d. Missing Attribute Values: None

e. Class Distribution: number of instances per class

class 1 59
class 2 71
class 3 48

Hãy thực hiện các yêu cầu sau:

3.1. Sử dụng Bộ dữ liệu trên, chia với tỷ lệ Train/Test là 7:3
3.2. Vẽ một biểu đồ thể hiện tỉ lệ phương sai (variance ratios) của trị riêng
3.3. Chọn 2 principal component đầu và trực quan dữ liệu này
3.4. Áp dụng thư viện PCA của sklearn để giảm chiều dữ liệu
3.5. Áp dụng giải thuật SVD để giảm chiều dữ liệu, trực quan hóa kết quả
3.6. Áp dụng giải thuật CUR để giảm chiều dữ liệu, trực quan hóa kết quả
3.7. So sánh, nhận xét về kết quả của 3 giải thuật PCA, SVD và CUR
3.8. Sử dụng Logistic Regression để huấn luyện mô hình
3.9. Phân loại dữ liệu trên tập Train chỉ với dữ liệu 2 chiều
3.10. Phân loại dữ liệu trên tập Test chỉ với dữ liệu 2 chiều