Một số thuật toán trong KHDL

BUỔI THỰC HÀNH 7

Trần Mạnh Tuấn
Bộ môn Hệ thống thông tin, Khoa CNTT
Trường đại học Thủy Lợi

NỘI DUNG

- Phân lớp dữ liệu trên python
- Phân cụm dữ liệu trên python

Phân lớp dữ liệu trên python

Giới thiệu về phân lớp dữ liệu

- **Mục đích:** để dự đoán những nhãn phân lớp cho các bộ dữ liệu/mẫu mới
- **Đầu vào:** một tập các mẫu dữ liệu huấn luyện, với một nhãn phân lớp cho mỗi mẫu dữ liệu
- **Đầu ra:** mô hình (bộ phân lớp) dựa trêntập huấn luyện và những nhãn phân lớp

Các bước phân lớp dữ liệu

- **Bước 1: Xây dựng mô hình** từ tập huấn luyện:
  - ✓ Mỗi bộ/mẫu dữ liệu được phân vàomột lớp được xác định trước
  - ✓ Lớp của một bộ/mẫu dữ liệu được xácđịnh bởi thuộc tính gán nhãn lớp
  - ✓ Tập các bộ/mẫu dữ liệu huấn luyện - tập huấn luyện tập huấn luyện được dùng để xây dựng mô hình
  - ✓ Mô hình được biểu diễn bởi các phương pháp phân lớp
- **Bước 2: Sử dụng mô hình** - kiểm tra tính đúng đắn của mô hình và dùng nó để phân lớp dữ liệu mới:
  - ✓ Phân lớp cho những đối tượng mới hoặc chưa được phân lớp
  - ✓ Đánh giá độ chính xác của mô hình
    - lớp biết trước của một mẫu/bộ dữ liệu đem kiểm tra được so sánh với kết quả thu được từ mô hình
    - tỉ lệ chính xác = phần trăm các mẫu/bộ dữ liệu được phân lớp đúng bởi mô hình trong số các lần kiểm tra

Các mô hình phân lớp dữ liệu

- Cây quyết định
- Naïve Bayes
- Mô hình thống kê
- Mạng nơ ron
- Mô hình SVM
- Mô hình KNN
- Các mô hình khác

Phân lớp dựa trên Naïve Bayes

- **Thư viện sử dụng:**
  - **from sklearn import** datasets
  - **from sklearn import** metrics
  - **from sklearn.naive_bayes import** GaussianNB
- **Đọc dữ liệu (bộ dữ liệu Iris):**
  - dataset = datasets.load_iris()
  - dataset.data[0:6]

Lưu ý: C:\Users\DELL\Anaconda3\Lib\site-packages\sklearn\datasets\data

- **Sử dụng Naive Bayes của Scikit-learn để xây dựng mô hình dự đoán**

Phân lớp dựa trên Naïve Bayes

- **Sử dụng Naive Bayes của Scikit-learn để xây dựng mô hình dự đoán:**
  - model = GaussianNB()
  - model.fit(dataset.data, dataset.target)
  - **print**(model)
- **Xem kết quả phân lớp:**
  - expected = dataset.target
  - predicted = model.predict(dataset.data)
  - print(metrics.classification_report(expected, predicted))

Phân lớp dựa trên Naïve Bayes

- **Kết quả thu được:**

	precision	recall	f1-score	support
0	1.00	1.00	1.00	50
1	0.94	0.94	0.94	50
2	0.94	0.94	0.94	50
avg / total	0.96	0.96	0.96	150

Phân lớp dựa trên Naïve Bayes

- **Kết quả dự báo trên các lớp:**

```
print(metrics.confusion_matrix(expected, predicted))
```

```
[[50  0  0]
 [ 0 47  3]
 [ 0  3 47]]
```

Phân lớp dựa trên các phương pháp khác

- Cây quyết định
- Naïve Bayes
- AdaBoostM1
- SVM
- RadomForest
- KNN

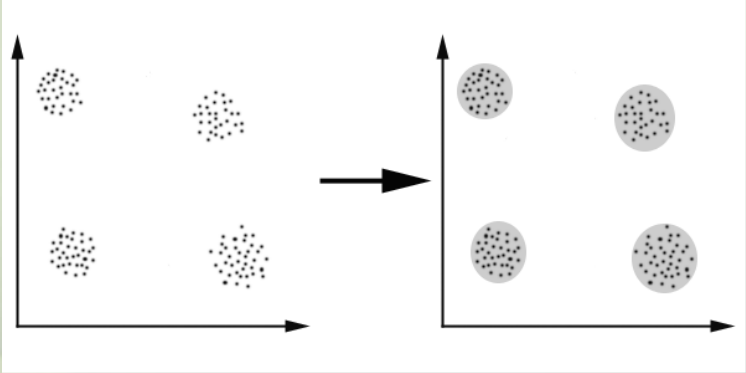
Lưu ý: Các phương pháp phân lớp khi cài đặt nằm trong thư mục (Phụ thuộc vào người cài khi đặt thư mục ở đâu):

"Anaconda3\pkgs\scikit-learn-0.19.1-py36h53aea1b_0\Lib\site-packages\sklearn"

Phân cụm dữ liệu trên python

Giới thiệu về phân cụm dữ liệu

Phân cụm dữ liệu



- Phân cụm rõ: các điểm dữ liệu được chia vào các cụm, trong đó mỗi điểm dữ liệu thuộc vào chính xác một cụm.
- Phân cụm mờ: các điểm dữ liệu có thể thuộc vào nhiều hơn một cụm với độ thuộc tương ứng.

Phân cụm dựa trên Kmeans

- **Thư viện sử dụng:**
  - **from sklearn import datasets**
  - **from sklearn import metrics**
  - **from sklearn.cluster import Kmeans**
- **Đọc dữ liệu (bộ dữ liệu Iris):**
  - dataset = datasets.load_iris()
  - dataset.data[0:6]

Lưu ý: C:\Users\DELL\Anaconda3\Lib\site-packages\sklearn\datasets\data

Phân cụm dựa trên Kmeans

- **Xây dựng mô hình phân cụm:**
  - kmeans = KMeans(n_clusters=3).fit(X)
- **Xem kết quả phân cụm:**
  - pred_label = kmeans.predict(X)

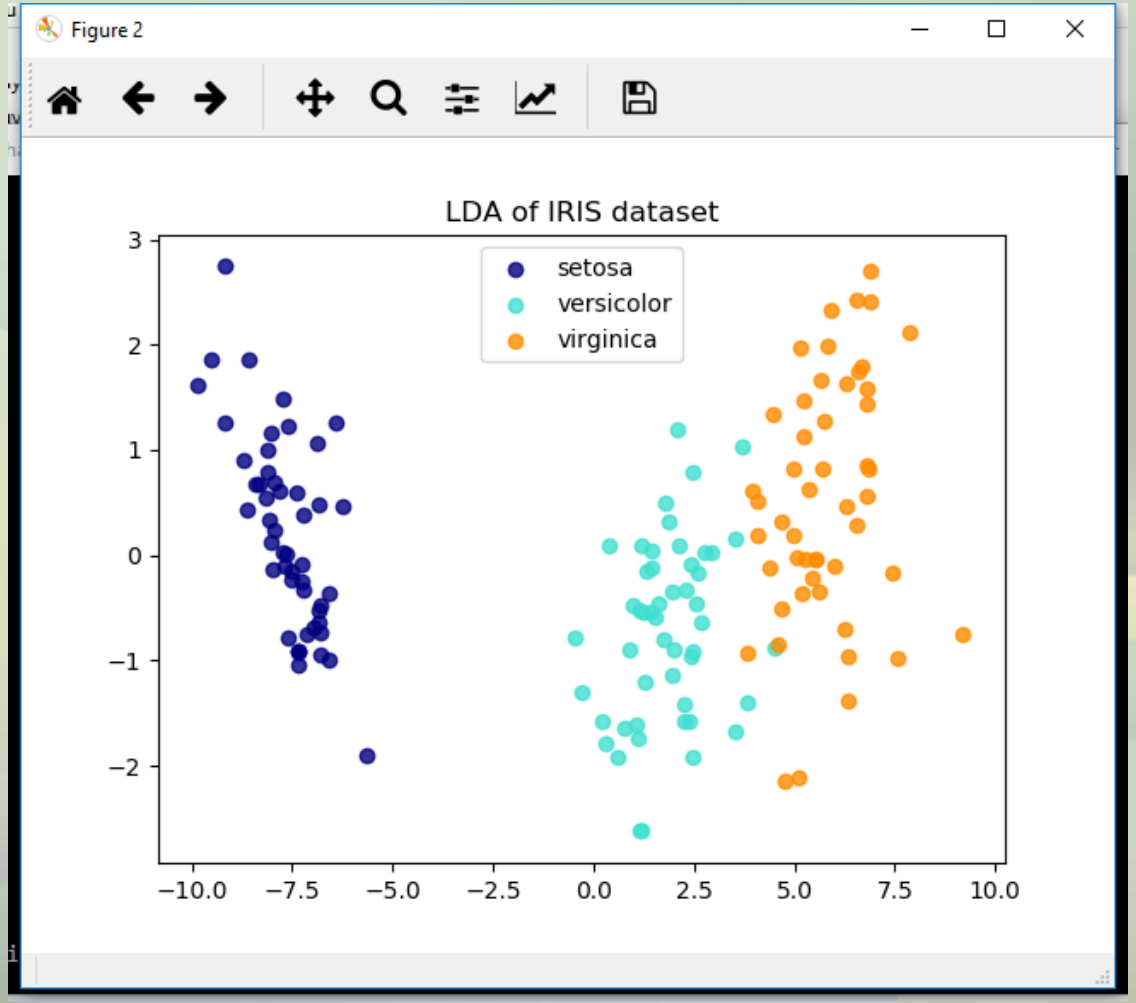
Phân cụm dựa trên Kmeans

- **Kết quả thu được:**

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 2 2 2 2 1 2 2 2 2
 2 2 1 1 2 2 2 2 1 2 1 2 1 2 2 1 1 2 2 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2
 2 1]
```

Phân cụm dựa trên Kmeans

- **Kết quả thu được:**



THỰC HÀNH