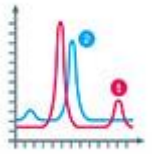
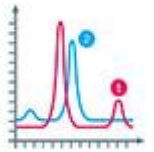


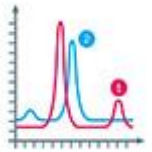
NHẬP MÔN LẬP TRÌNH KHOA HỌC DỮ LIỆU

Bài 11: Thư viện scikit-learn

Nội dung



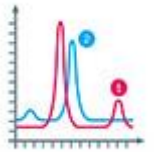
1. Mối quan hệ giữa Khoa học Dữ liệu và Học máy
2. Một số loại bài toán học máy
3. Thư viện học máy scikit-learn
4. Bài tập



Phần 1

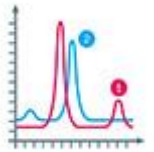
Mối quan hệ giữa Khoa học Dữ liệu và Học máy

Khoa học dữ liệu là gì?



- Hầu hết các ngành khoa học từ xưa đến nay đều giải quyết vấn đề dựa trên lập luận và tri thức
  - Ngành toán: dựa trên các mệnh đề, công thức, lập luận… để chứng minh bài toán
  - Ngành vật lý: dựa trên các quan sát, thực nghiệm, tính toán,… kiểm chứng các giả thiết
  - Ngành hóa học:…
  - …
  - Ta gọi các ngành khoa học này là “knowledge-driven” (dẫn dắt bởi tri thức)
- Có ngành có chút ngoại lệ, ví dụ: ngành xác suất

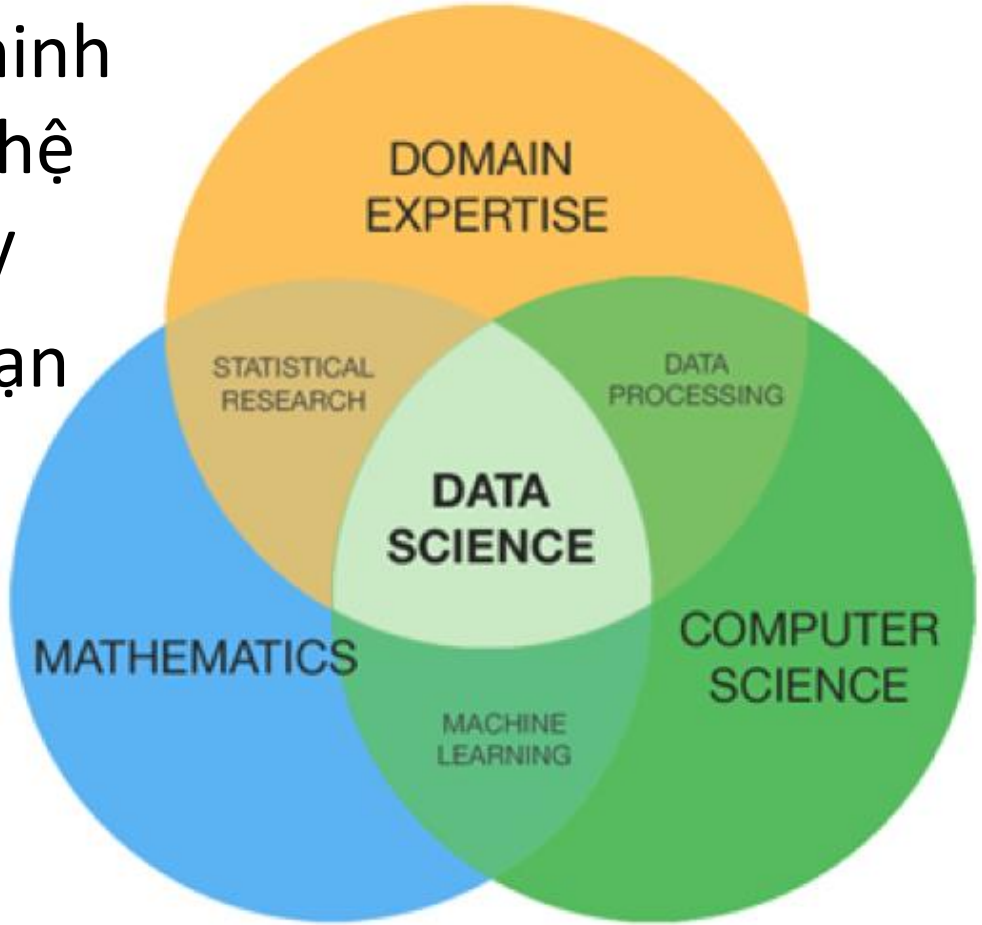
Khoa học dữ liệu là gì?



- Với quan điểm như vậy, tất cả những quan sát mà không được chứng minh chặt chẽ thường được cho là "không khoa học"
  - Chẳng hạn: chuồn chuồn bay thấp thì mưa
- Khoa học dữ liệu ≠ Khoa học thông thường ở quan điểm: tìm tri thức từ dữ liệu (dẫn dắt bởi dữ liệu – "data-driven")
  - Chúng ta rút ra tri thức bằng việc tìm tòi từ dữ liệu chứ không nhất thiết phải chứng minh nó
  - Tất nhiên tri thức tìm ra phải có tính ổn định (luôn có cùng kết quả nếu sử dụng cùng một phương pháp)

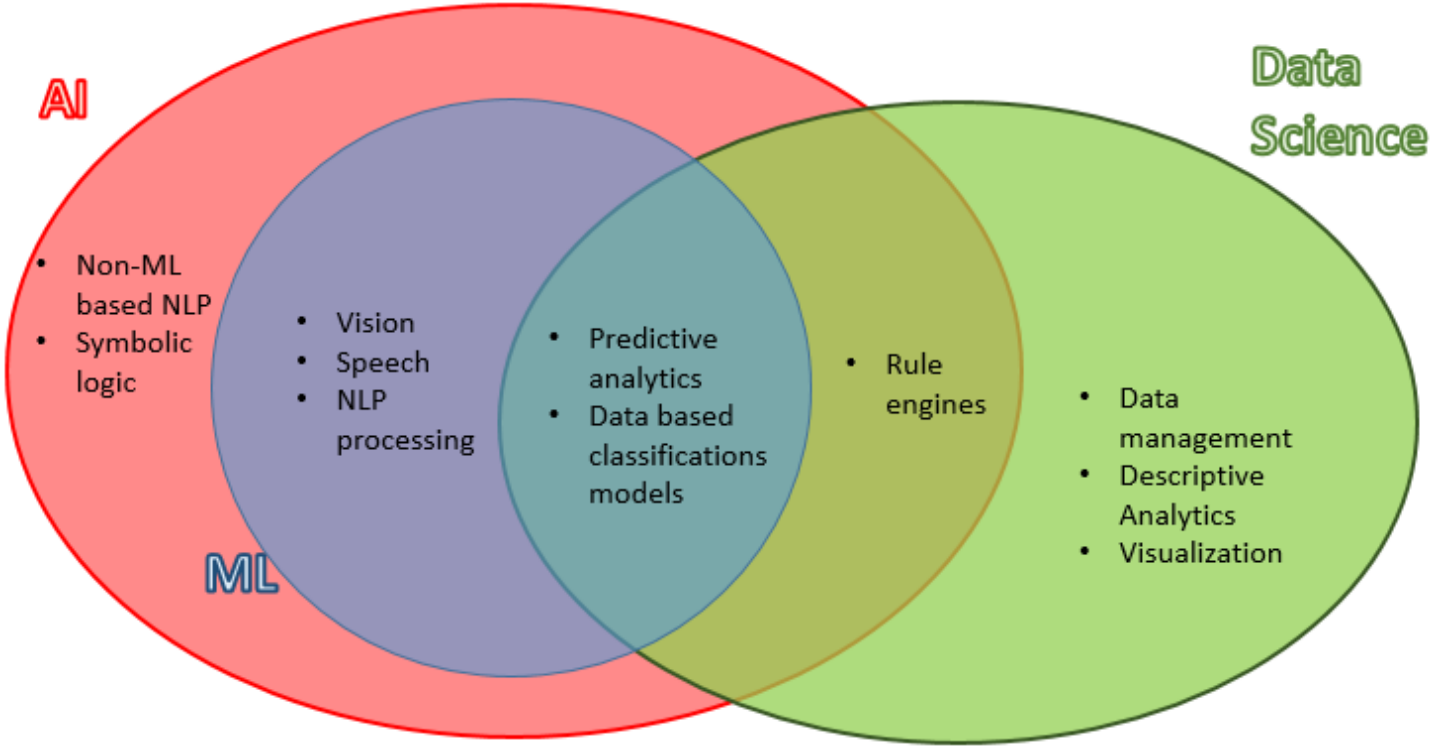
Khoa học Dữ liệu và Học máy

- Không có sơ đồ nào minh họa đầy đủ mối quan hệ giữa hai khái niệm này
- Nhiều người (chẳng hạn như Nate Silver) cho rằng ngành khoa học dữ liệu chỉ là một dạng thống kê

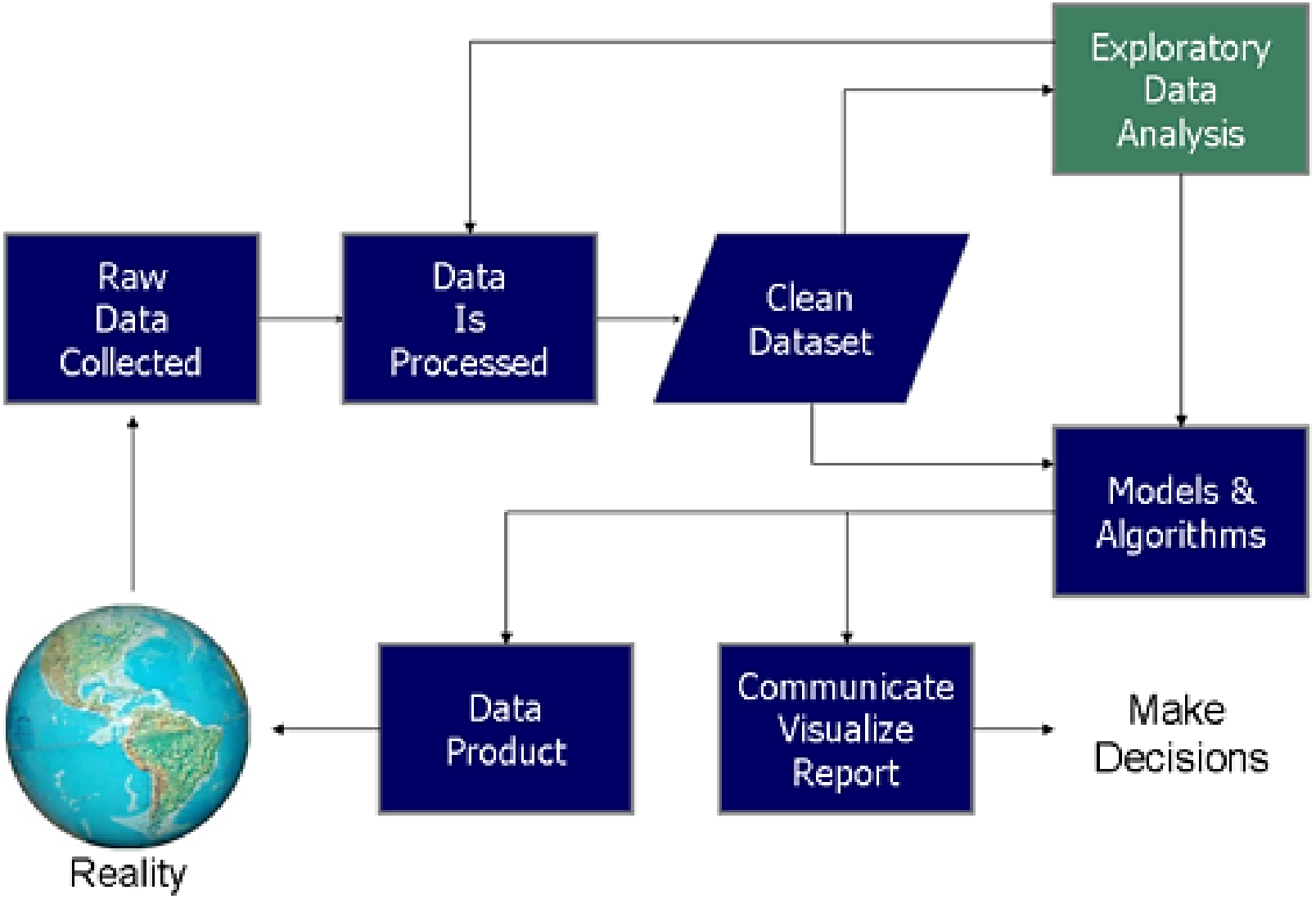


Khoa học Dữ liệu và Học máy

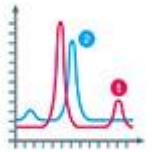
- Học máy là phương pháp quan trọng để xử lý dữ liệu trong ngành data science, bên cạnh những phương pháp truyền thống khác



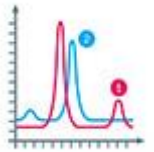
Quá trình xử lý của khoa học dữ liệu



Ví dụ: hệ thống phát hiện thư rác



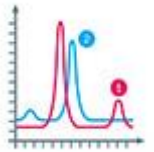
1. Thu thập mẫu thư (gồm cả thư rác và thư thường)
2. Xác định đề bài (phân lớp hay đánh giá)
3. Xử lý dữ liệu
4. Chọn mô hình học máy phù hợp với bài toán phân loại thư rác
5. Huấn luyện mô hình
6. Hiệu chỉnh, tinh chỉnh mô hình
7. Áp dụng thực tế (chạy trên email server thực)
8. Tiếp tục cập nhật theo phản hồi của người dùng



Phần 2

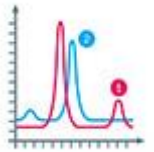
Một số loại bài toán học máy

Một số bài toán thực tế



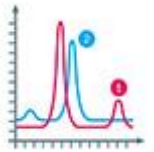
- Hệ thống phân loại email
- Nhận dạng chữ viết từ ảnh
- Ước lượng giá cả của sản phẩm
- Dự báo thời tiết
- Đánh giá trạng thái của người qua ảnh/video
- Trả lời tự động (chat bot)
- Gợi ý sản phẩm phù hợp với nhu cầu khách hàng
- Tự động chơi trò chơi
- Mô phỏng giọng nói của một người nào đó

Các lớp bài toán cơ bản



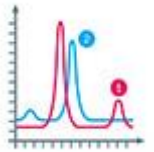
- Học có giám sát (supervised learning): học cách tiên đoán đầu ra theo mẫu cho trước
  - Tập mẫu cho trước, cho cả đầu bài và kết quả
    - Cho email, chỉ rõ trước đâu là spam, đâu không phải spam
  - Mô hình được huấn luyện trên tập mẫu
  - Thử nghiệm bằng cách cho đầu bài, mô hình tiên đoán kết quả, mô hình đoán càng chính xác càng tốt
    - Cho một email mới, máy tính đoán xem có phải spam không?
  - Có 2 loại cơ bản:
    - Hồi quy (regression): đầu ra là số hoặc vector
    - Phân lớp (classification): đầu ra thường là xác suất dự báo

Các lớp bài toán cơ bản



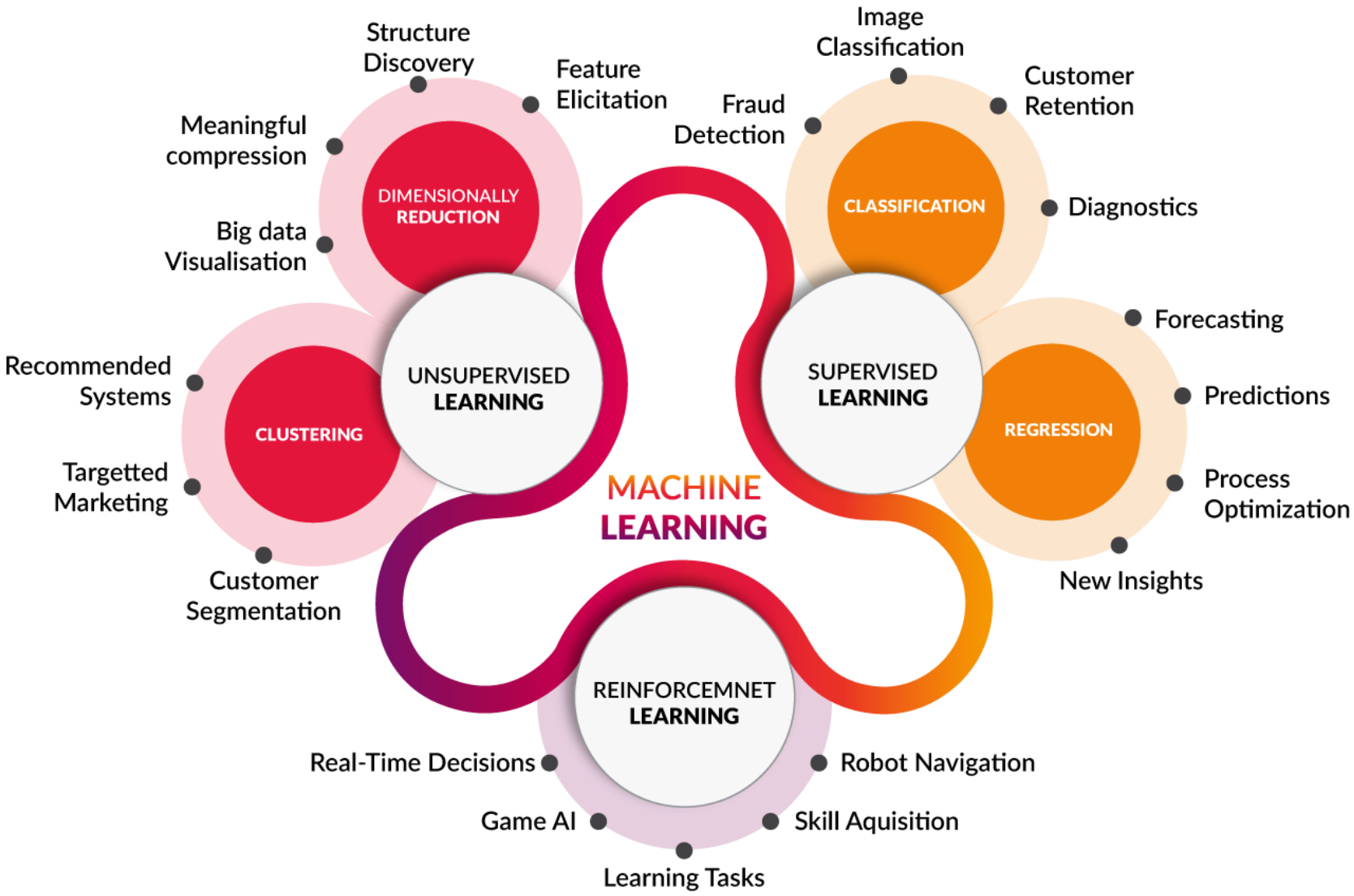
- Học không giám sát (unsupervised learning): tự khai phá các đặc trưng nội tại hợp lý của đầu vào
  - Chỉ cho mẫu vào, không cho biết đầu ra
    - Cho tập băng ghi âm lời nói của một người
  - Hệ thống tự học trên các mẫu mà không có định hướng
    - Tạo ra một đoạn phát âm theo ngữ điệu của người đã cho
  - Một vài chiến lược cơ bản:
    - Biến đổi dữ liệu đầu vào có số chiều cao thành dữ liệu có số chiều thấp hơn
    - Dữ liệu có số chiều cao nhưng các đặc trưng thành phần có tính "kinh tế" (economical) hơn
    - Gom cụm dữ liệu đầu vào

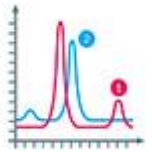
Các lớp bài toán cơ bản



- Học tăng cường (reinforcement learning): hiệu chỉnh các siêu tham số (hyperparameter) để cực đại hóa lợi ích trong tương lai
  - Cho bối cảnh và các quy tắc
    - Bàn cờ Vây và các quy tắc của trò chơi cờ Vây
  - Ứng với mỗi hành động (hoặc chuỗi hành động), có một phần thưởng tương ứng
    - Đặt một quân sẽ bị mất điểm, không được hoặc được điểm
  - Hệ thống tự điều chỉnh chuỗi hành động sao cho được phẩn thưởng lớn nhất
    - Hệ thống học cách chơi để thắng người chơi giỏi nhất

Các lớp bài toán cơ bản

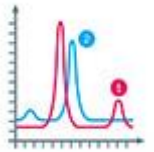




Phần 3

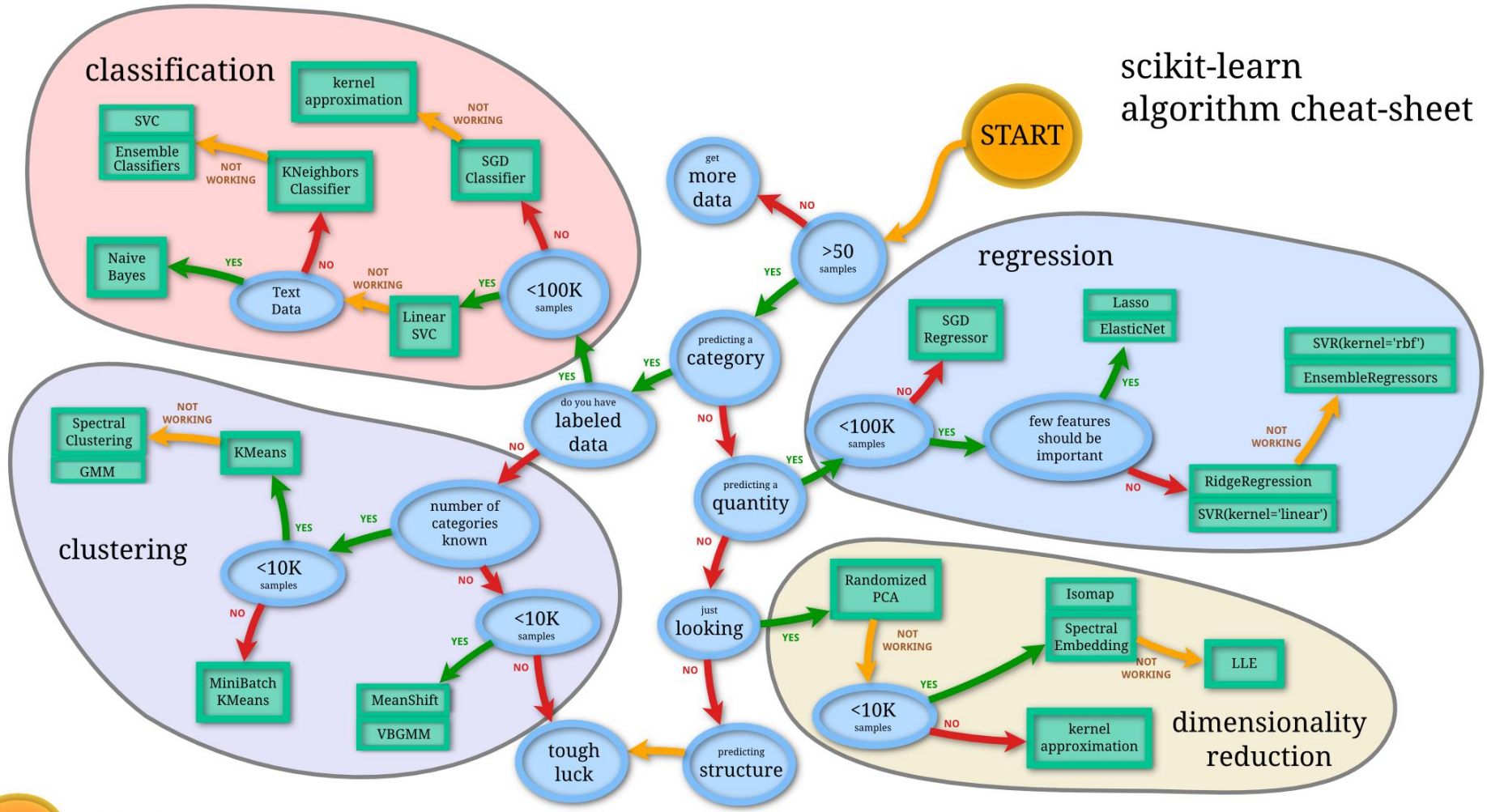
Thư viện học máy scikit-learn

Thư viện học máy scikit-learn

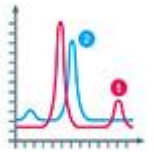


- Scikit-learn xuất phát là một dự án trong một cuộc thi lập trình của Google vào năm 2007, người khởi xướng dự án là David Cournapeau
- Sau đó nhiều viện nghiên cứu và các nhóm ra nhập, đến năm 2010 mới có bản đầu tiên (v0.1 beta)
- Scikit-learn cung cấp gần như tất cả các loại thuật toán học máy cơ bản (khoảng vài chục) và vài trăm biến thể của chúng, cùng với đó là các kĩ thuật xử lý dữ liệu đã được chuẩn hóa
- Cài đặt: **pip install scikit-learn scipy**

Chọn thuật toán học máy phù hợp



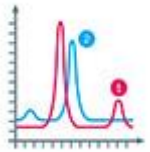
Ví dụ: dự báo cân nặng của người



- Tập mẫu quan sát có n người
  - Gồm tên, chiều cao, cân nặng
  - Và nhiều loại chỉ số khác nữa
- Xây dựng một mô hình dự báo về cân nặng người, dựa trên các chỉ số còn lại
  - Trong trường hợp bài toán của ta, chúng ta cố gắng dự báo cân nặng từ chiều cao
  - Thực tế thì cân nặng phụ thuộc vào nhiều thông số khác nữa, như giới tính, vòng eo,…

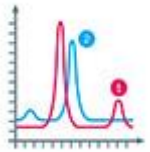
	A	B	C
1	Ten	Cao	Nang
2	A	147	49
3	B	150	50
4	C	153	51
5	D	155	51
6	E	168	60
7	F	170	62
8	G	173	68
9	H	175	65
10	I	178	66
11	J	180	71
12	K	183	68
13	L	165	59
14	M	163	58
15	N	160	56
16	O	158	54
17	P	169	62
18	Q	172	63
19	S	170	62
20	T	176	62
21	U	180	69

Dự báo sử dụng hồi quy tuyến tính



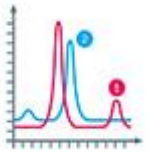
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn import linear_model, metrics

# đọc dữ liệu từ file csv
df = pd.read_csv("nguoi.csv", index_col = 0)
print(df)

# vẽ biểu đồ minh họa dataset
plt.plot(df.Cao, df.Nang, 'ro')
plt.xlabel('Chiều cao (cm)')
plt.ylabel('Cân nặng (kg)')
plt.show()
```

Dự báo sử dụng hồi quy tuyến tính



```
# sử dụng hồi quy tuyến tính
X = df.loc[:, ['Cao']].values                    # X là dữ liệu đầu vào
y = df.Nang.values                               # y là dữ liệu đầu ra
model = linear_model.LinearRegression()          # loại mô hình
model.fit(X, y)                                  # tập huấn trên dữ liệu

# in một số thông tin về mô hình
mse = metrics.mean_squared_error(model.predict(X), y)
print("Tổng bình phương sai số trên tập mẫu:", mse)
print("Hệ số hồi quy:", model.coef_)
print("Sai số:", model.intercept_)
print(f"Công thức: [Nặng] = {model.coef_} x [Cao] +
{model.intercept_}")
```

Dự báo sử dụng hồi quy tuyến tính



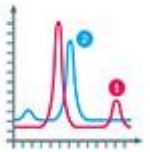
```
# vẽ lại sơ đồ
plt.scatter(X, y, c='b')
plt.plot(X, model.predict(X))
plt.show()

# dự báo một số tình huống
while True:
      x = float(input("Nhập chiều cao (nhập 0 để dừng): "))
      if x <= 0: break
      print("Người cao", x, "cm, dự báo cân nặng",
model.predict([[x]]))
```

Mở rộng: thêm cột giới tính

- Vẫn dữ liệu cũ, bổ sung thêm cột giới tính (Nam/Nu)
- Sử dụng phương pháp cũ, để xem giới tính ảnh hưởng như thế nào đến cân nặng

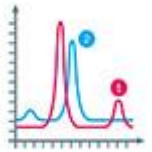
	A	B	C	D
1	Ten	Gioitinh	Cao	Nang
2	A	Nu	147	49
3	B	Nu	150	50
4	C	Nu	153	51
5	D	Nam	155	51
6	E	Nu	168	60
7	F	Nam	170	62
8	G	Nu	173	68
9	H	Nam	175	65
10	I	Nam	178	66
11	J	Nam	180	71
12	K	Nam	183	68
13	L	Nam	165	59
14	M	Nu	163	58
15	N	Nu	160	56
16	O	Nu	158	54
17	P	Nam	169	62
18	Q	Nam	172	63
19	S	Nu	170	62
20	T	Nam	176	62
21	U	Nam	180	69

Dự báo sử dụng hồi quy tuyến tính



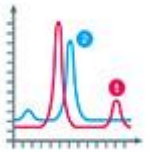
```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
from sklearn import linear_model, metrics

# đọc dữ liệu từ file csv
df = pd.read_csv("nguoi2.csv", index_col = 0)
print(df)

# thêm cột mới, giới tính Nam = 1, giới tính Nữ = 0
df['GT'] = df.Gioitinh.apply(lambda x: 1 if x=='Nam' else 0)
print(df)
```

Dự báo sử dụng hồi quy tuyến tính

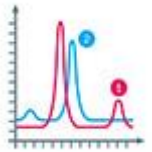


```
# sử dụng hồi quy tuyến tính
X = df.loc[:, ['Cao', 'GT']].values          # X là dữ liệu đầu vào
y = df.Nang.values                           # y là dữ liệu đầu ra
model = linear_model.LinearRegression()      # loại mô hình
model.fit(X, y)                              # tập huấn trên dữ liệu

# in một số thông tin về mô hình
mse = metrics.mean_squared_error(model.predict(X), y)
print("Tổng bình phương sai số trên tập mẫu:", mse)
print("Hệ số hồi quy:", model.coef_)
print("Sai số:", model.intercept_)
print(f"Công thức: [Nặng] = {model.coef_} x [Cao, Giới tính] +
{model.intercept_}")
```

Dự báo sử dụng hồi quy tuyến tính



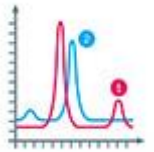
```
# dự báo một số tình huống
while True:
      x = float(input("Nhập chiều cao (nhập 0 để dừng): "))
      if x <= 0: break
      print("Nam giới cao", x, "cm, dự báo cân nặng",
model.predict([[x, 1]]))
      print("Nữ giới cao", x, "cm, dự báo cân nặng",
model.predict([[x, 0]]))
```



Phần 4

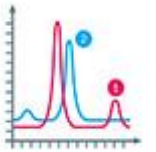
Bài tập

Bài tập



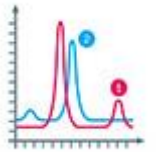
1. Tải về file winequality.csv về các số đo của rượu vang và chất lượng của rượu
   - Liên kết: http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv
   - Đây là bộ data của đại học California-Berkeley
   - Bộ data gồm 1599 mẫu rượu vang, mỗi mẫu gồm 11 loại chỉ số và đánh giá của chuyên gia về chất lượng rượu (cột quality, điểm số từ 0 đến 10)
   - Chú ý:
     - Dữ liệu sử dụng dấu chấm phẩy (;) để ngăn giữa các cột
     - Tên các cột có chứa dấu cách

Bài tập



2. In ra dữ liệu vừa tải về, ý nghĩa các cột thuộc tính
   - 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 sulfurus tự do
   - total sulfur dioxide — Nồng độ acid sulfurus
   - density — Mật độ (khối lượng/đơn vị thể tích)
   - pH — Độ pH
   - sulphates — Nồng độ sunfat
   - alcohol — Nồng độ chất alcohol

Bài tập



3. Sử dụng các cột thuộc tính “alcohol” để tìm tương quan giữa thuộc tính này và điểm chất lượng rượu
4. Sử dụng tất cả 11 thuộc tính để tìm tương quan giữa các thuộc tính với điểm chất lượng rượu
5. (*) Loại bỏ những thuộc tính “không quan trọng”, chọn 3 thuộc tính quan trọng nhất và xây dựng tương quan tuyến tính giữa 3 thuộc tính đó với điểm chất lượng rượu