Nội dung 2: Khám phá dữ liệu với Python

1. python

Giới thiệu Python:

- https://youtu.be/HvVdgcLl9rc
- https://youtu.be/NZj6LI5a9vc

Hướng dẫn cài Python:

- https://www.youtube.com/watch?v=g5BdrxPhQU0&ab_channel=CodeXplore
- https://machinelearningcoban.com/faqs/

Cài đặt python và các thư viện trên Windows

- Cài đặt Python bằng Anaconda
  - Anaconda hỗ trợ rất nhiều thư viện giúp lập trình Python.
  - Để tải về Python và một số thư viện cần thiết, tải về Anaconda cho windows và cài đặt

  https://docs.continuum.io/anaconda/install/#anaconda-for-windows-install/
  - Sau khi cài đặt xong, bạn vào thư mục Scripts trong thư mục Anaconda vừa cài đặt, và khởi động Spyder.

10

Kiểm tra Libs

- Anaconda đã có sẵn khá là nhiều thư viện python như: Numpy, Scipy, Matplotlib, sklearn
- Để kiểm tra python của Anaconda đã có thư viện nào đó, chúng ta sẽ thử import nó trong Console.
  - >>> import numpy
  - >>> import sklearn

11

Cài đặt Libs bằng Anaconda

- Chúng ta sẽ bật cmd (Command Prompt) của windows gõ lệnh:
  - `conda install scikit-learn` hoặc
  - `pip install -U scikit-learn`
- Conda sẽ tự động tìm thư viện *sklearn* và cài vào đường dẫn Anaconda giúp chúng ta.

12

2. Khám phá dữ liệu với Python

- https://www.youtube.com/watch?v=HPGYTWYM13s
- https://kungfupandas.lhduc.com/gi%E1%BB%9Bi-thi%E1%BB%87u-pandas.html

13

- Yêu cầu: thực hành trên python, sử dụng: numpy, pandas,…
  - Đọc dữ liệu
  - Xem đặc điểm của dữ liệu
  - Thêm hàng/cột
  - Xóa hàng/cột
  - Gom nhóm
  - Trích xuất dữ liệu
  - Xử lý dữ liệu thiếu…

14

3. Chuẩn hóa dữ liệu

- **Tham khảo:**
- https://www.geeksforgeeks.org/data-pre-processing-wit-sklearn-using-standard-and-minmax-scaler/

- **Yêu cầu:**
  - Thực hành biến đổi dữ liệu (vd chuyển đơn vị đo)
  - Chuẩn hóa dữ liệu:
    - Min-max
    - Z-score

15

Chuẩn hóa min-max

- Sử dụng MinMaxScaler (thư viện scikit-learn)

```
# Ví dụ về scale sử dụng MinMaxScaler
from sklearn.preprocessing import MinMaxScaler
# Load dữ liệu
data = ...
# tạo bộ scaler, mặc định chuẩn hóa về [0,1]
scaler = MinMaxScaler()
#nếu muốn chuẩn hóa về miền bất kỳ, ví dụ [1,10]
scaler = MinMaxScaler(feature_range=(1,10))
# fit scaler vào data
model = scaler.fit(data)
# Thực hiện scale
normalized = model.transform(data)
# Nếu muốn quay lại miền giá trị cũ
inverse = model.inverse_transform(normalized)
```

16

Chuẩn hóa z-Score

- Chuẩn hóa dữ liệu bằng thư viện **scikit-learn** với **StandardScaler**

```
# ví dụ chuẩn hóa z-score với sklearn
from sklearn.preprocessing import StandardScaler
# load data
data = ...
# create scaler
scaler = StandardScaler()
# fit scaler on data
model=scaler.fit(data)
# apply transform
standardized = scaler.transform(data)
# Nếu muốn quay lại dữ liệu cũ
inverse = model.inverse_transform(standardized)
```

17

4. Trực quan hóa dữ liệu

- **Tham khảo:**
- https://phamdinhkhanh.github.io/deepai-book/ch_appendix/index_matplotlib.html