ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Học máy cơ bản

Quy trình xây dựng hệ thống học máy

Ngô Văn Linh

Nội dung môn học

- Buổi 1: Giới thiệu về Học máy
- **Buổi 2: Quy trình xây dựng hệ thống học máy**
- Buổi 3: Hồi quy tuyến tính
- Buổi 4: Học dựa trên láng giềng gần nhất (KNN)
- Buổi 5: Cây quyết định và Rừng ngẫu nhiên
- Buổi 6: Naïve Bayes
- Buổi 7: Máy vector hỗ trợ (SVM)
- Buổi 8: Đánh giá hiệu quả của mô hình học máy
- Buổi 9: Phân cụm
- Buổi 10: Kiểm tra giữa kỳ và trình bày ý tưởng làm dự án cuối kỳ

Mục lục

1. Giới thiệu về Scikit-learn, Keras/Pytorch

Giới thiệu thư viện Scikit-Learn

- Scikit-learn: Là thư viện mạnh mẽ về các thuật toán học máy và khai phá dữ liệu bằng ngôn ngữ Python
- Bắt nguồn từ một dự án của Google, Sklearn được xây dựng chủ yếu dựa trên Python, trên nền tảng của một số thư viện khác: Numpy, Scipy, Matplotlib, IPython, Sympy,

Giới thiệu thư viện Scikit-Learn

- Nhóm các thuật gói (https://scikit-learn.org/stable/user_guide.html):
  - Supervised learning
  - Unsupervised learning
  - Model selection and evaluation
  - Datasets
  - Dataset transformations
  - Visualization
  - …

Cài đặt Sklearn

- Trước khi cài Sklearn, cần cài trước các thư viện sau:
- Python (>= 2.6 or >= 3.3),
- NumPy (>= 1.6.1),
- SciPy (>= 0.9).
- Sau đó, có thể dùng pip/ conda gọi câu lệnh cài đặt sklearn:
  - *!pip install -U scikit-learn*
  - *!conda install scikit-learn*
- !!!Khi sử dụng, ta luôn cần lệnh ***import sklearn***...

Một số ví dụ

- Tiền xử lý: **sklearn.preprocessing**

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
sc.fit(X_train)
```

- Chia dữ liệu train/test: **sklearn.model_selection**

```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
sc.fit(X_train)
```

Một số ví dụ

- Huấn luyện mô hình tuyến tính: `sklearn.linear_model`

```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error
lr = LinearRegression()
lr_lasso = Lasso()
lr_ridge = Ridge()
```

- Huấn luyện và đánh giá:

```
lr.fit(X_train, y_train)
lr_score = lr.score(X_test, y_test) # with all num var 0.7842744111909903
lr_rmse = rmse(y_test, lr.predict(X_test))
lr_score, lr_rmse
```

```
(0.7837532911322952, 65.91685277030534)
```

Một số ví dụ

- Mô hình SVM và Rừng ngẫu nhiên

1.2.2 Support Vector Machine

```
from sklearn.svm import SVR
svr = SVR()
svr.fit(X_train,y_train)
svr_score=svr.score(X_test,y_test)
svr_rmse = rmse(y_test, svr.predict(X_test))
svr_score, svr_rmse
```

```
(0.24613512350275257, 123.07460910013376)
```

1.2.3 Random Forest

```
from sklearn.ensemble import RandomForestRegressor
rfr = RandomForestRegressor()
rfr.fit(X_train,y_train)
rfr_score=rfr.score(X_test,y_test)
rfr_rmse = rmse(y_test, rfr.predict(X_test))
rfr_score, rfr_rmse
```

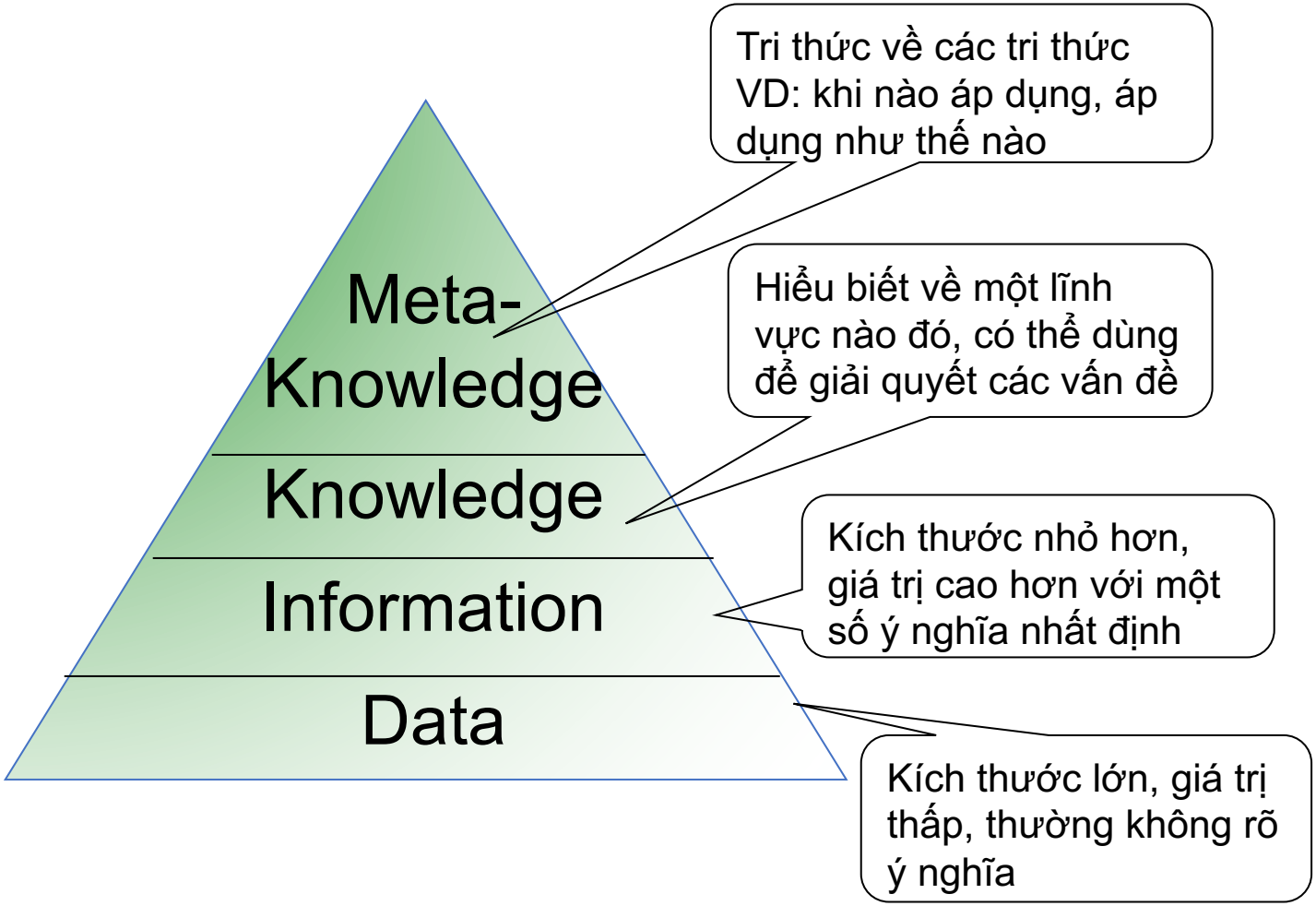
```
(0.8922988135841156, 46.51916964240559)
```

Keras/Pytorch

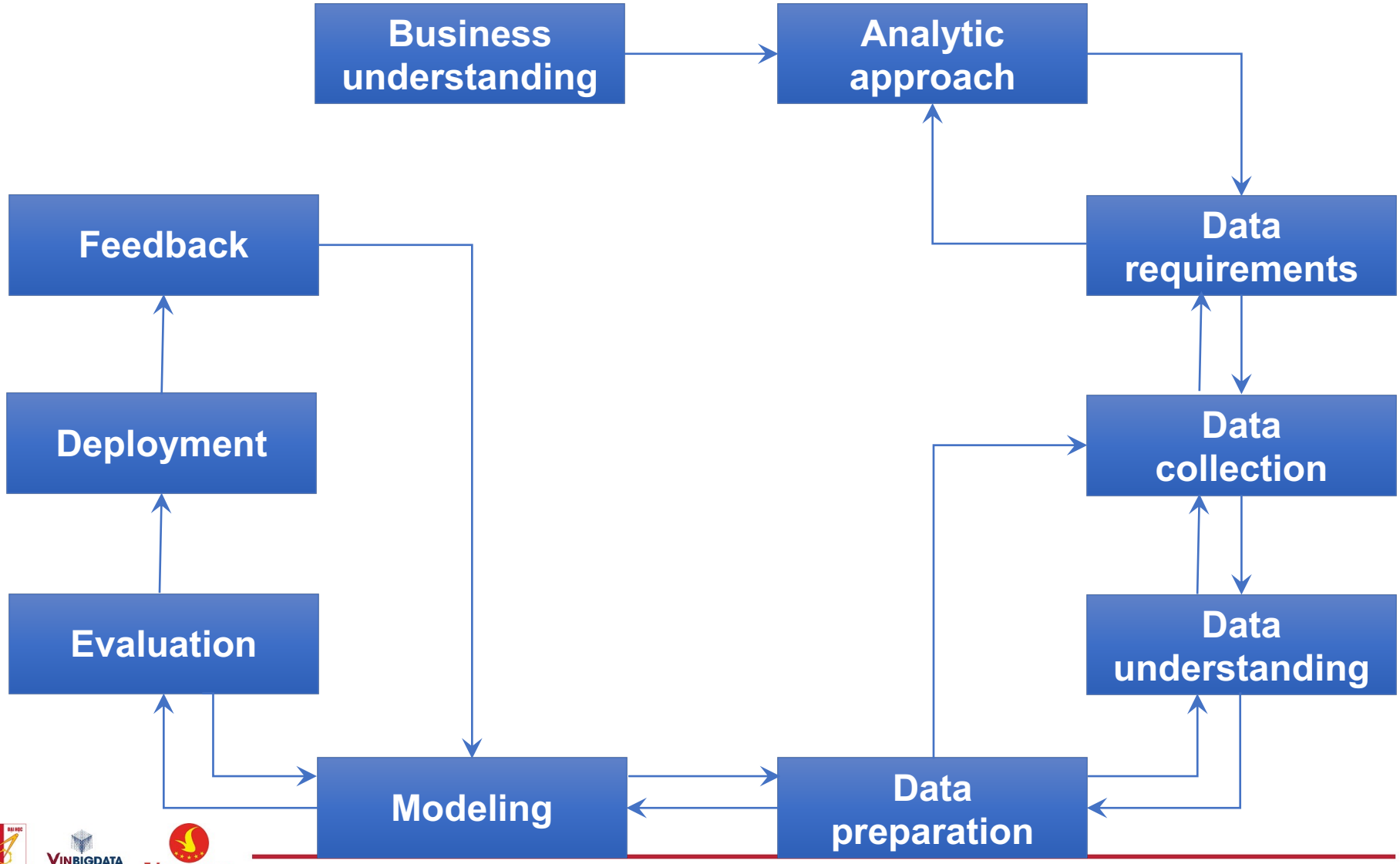
- Thư viện hỗ trợ lập trình giải bài toán học sâu: Keras, Pytorch, Caffee, TensorFlow, etc.
- ⇒ Giúp việc lập trình cho bài toán học sâu trở nên đơn giản và hiệu quả hơn.

- Học trong phần học sâu

2. Quy trình xây dựng hệ thống học máy, khai phá dữ liệu



Quy trình thực hiện: hướng sản phẩm



(http://www.theta.co.nz/)

3. Tiền xử lý dữ liệu

- Đặt vấn đề
- Làm sạch dữ liệu
- Tích hợp
- Giảm chiều

ĐẶT VẤN ĐỀ

- Khai phá dữ liệu là một qúa trình phân tích dữ liệu theo nhiều khía cạnh và tổng hợp nó lại để có được thông tin hữu ích hay tri thức.

ĐẶT VẤN ĐỀ

- Các bước của quá trình phát hiện tri thức gồm
  1. Thu thập, lựa chọn dữ liệu
  2. Tiền xử lý dữ liệu
  3. Chuyển đổi
  4. Khai phá dữ liệu
  5. Giải thích/Đánh giá

ĐẶT VẤN ĐỀ

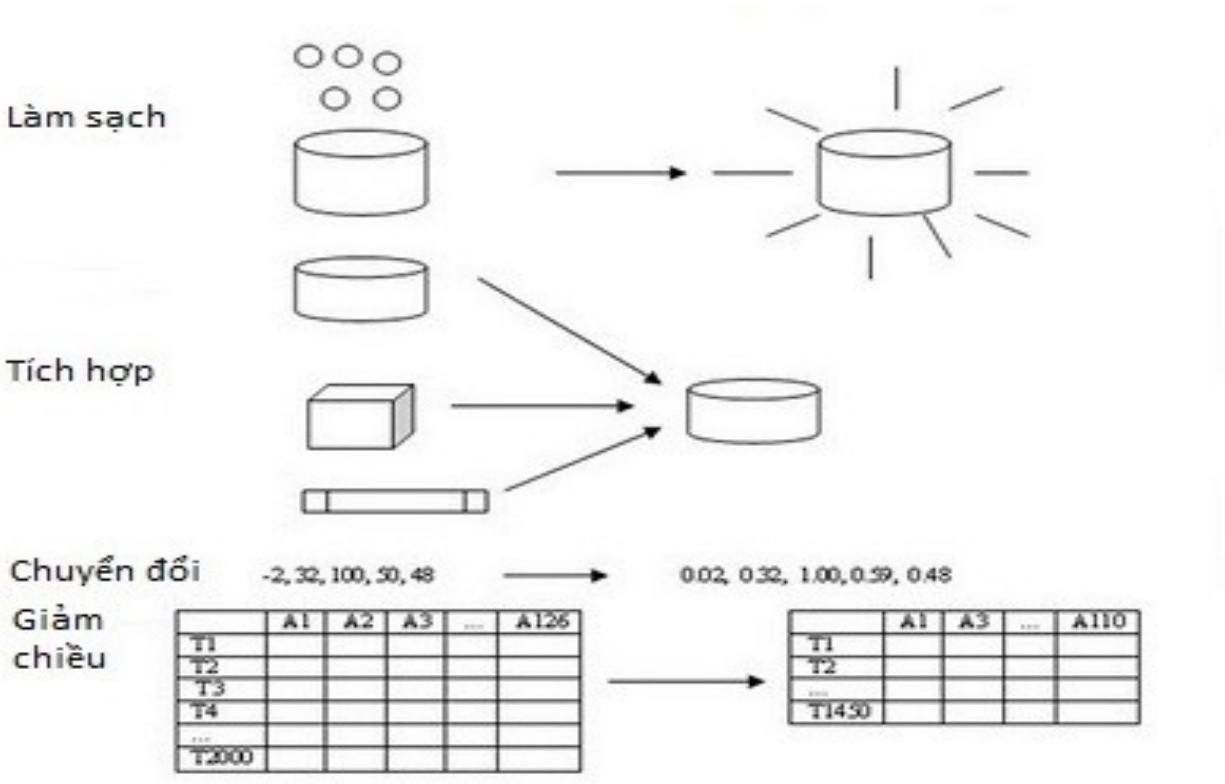
- Vì sao phải tiền xử lý dữ liệu ?
  - Không đầy đủ (Incomplete) : thiếu một vài giá trị thuộc tính ....
  - Nhiễu (Noisy) : xuất hiện giá trị lỗi, lỗi chủ quan người nhập dữ liệu ....
  - Không nhất quán (Inconsistent) : sự khác biệt trong cách phân loại, phân biệt hay đơn vị của dữ liệu ...

ĐẶT VẤN ĐỀ

- Quy trình tiền xử lý dữ liệu
  1. Làm sạch : Loại bỏ các giá trị sai, kiểm tra tính nhất quan của dữ liệu.
  2. Tích hợp : Dữ liệu có nhiều nguồn nên cần lưu theo một cách thức thống nhất.
  3. Chuyển đổi : Chuẩn hóa và tập hợp dữ liệu.
  4. Giảm chiều : Mô tả dữ liệu trong kích thước nhỏ nhưng không làm mất kết quả cần kiết xuất.

ĐẶT VẤN ĐỀ

- Quy trình tiền xử lý dữ liệu



MỤC LỤC

- Đặt vấn đề
- Làm sạch dữ liệu
- Tích hợp
- Giảm chiều

LÀM SẠCH DỮ LIỆU

- Đây là thủ tục quan trọng gồm ba bước chính
  1. Điền đầy các giá trị bị mất
  2. Chuốt dữ liệu để loại nhiễu
  3. Kiểm tra và sửa tính không nhất quán

LÀM SẠCH DỮ LIỆU

- **Bước 1**: Điền đầy các giá trị bị mất, có thể chọn một trong các phương pháp
  - Bỏ không xét đến bộ dữ liệu bị mất giá trị
  - Điền lại giá trị bằng tay
  - Gán cho giá trị nhãn đặc biệt hay ngoài khoảng biểu diễn
  - Gán giá trị trung bình cho nó.
  - Gán giá trị trung bình của các mẫu khác thuộc cùng lớp đó.
  - Tìm giá trị có xác suất lớn nhất điền vào chỗ bị mất (hồi quy, suy diễn Bayes, cây quyết định qui nạp)

LÀM SẠCH DỮ LIỆU

- **Bước 2**: Chuốt dữ liệu loại nhiễu, có thể chọn một trong các phương pháp
  - Hồi quy (Regression) : sẽ dành chương riêng
  - Phân cụm (Cluster) : sẽ dành chương riêng

LÀM SẠCH DỮ LIỆU

- **Bước 3**: kiểm tra và sửa tính không nhất quán trong dữ liệu.
  - Đề phát hiện kiểm tra sự bất thường trong giá trị dữ liệu
  - Dùng để sửa tính không nhất quán dữ liệu
    - Công cụ chà dữ liệu (Data scrubbing tools) : dùng cho một lĩnh vực cụ thể.
    - Công cụ kiểm toán dữ liệu (Data auditing tools) : dùng cho việc phân tích dữ liệu, xác định quan hệ, xác định các luật.

MỤC LỤC

- Đặt vấn đề
- Làm sạch dữ liệu
- Tích hợp
- Giảm chiều

TÍCH HỢP

- **Ý nghĩa tích hợp và chuyển đổi dữ liệu, chuẩn hóa để tiến hành khai phá/học máy**
  - Tập hợp : các giá trị dữ liệu tạo thành bộ hay khối
  - Tổng quát hóa (generalization) : các dữ liệu "thô" được thay bằng các khái niệm đã chuẩn hóa
  - Chuẩn hóa (normalization) : nếu phạm vi dữ liệu lớn thì đưa nó về phạm vi chuẩn
  - Xây dựng thuộc tính (attribute construction) : thuộc tính mới thêm vào giúp quá trinh khai phá dữ liệu

MỤC LỤC

- Đặt vấn đề
- Làm sạch dữ liệu
- Tích hợp
- Giảm chiều

GIẢM CHIỀU

- Ý nghĩa: Việc giảm kích thước của dữ liệu cần đồng thời giữ được tính phân tích dữ liệu, tăng tốc qúa trình khai phá/học máy

GIẢM CHIỀU

- Các chiến lược giảm kích thước dữ liệu
  - Lựa chọn tập con các thuộc tính : trong đó các thuộc tính không liên quan, dư thừa hoặc các chiều cũng có thể xóa hay loại bỏ
  - Giảm chiều : trong đó cơ chế mã hóa được sử dụng để giảm kích cỡ tập dữ liệu
  - Rời rạc hóa và trừu tượng khái niệm : trong đó các giá trị dữ liệu thô được thay thế bằng các khái niệm trừu tượng đã rời rạc hóa.

TỔNG KẾT

- Scikit-learn, pytorch, keras là những thư viện phổ biến sử dụng trong học máy
- Quy trình xây dựng hệ thống học máy: Thu thập dữ liệu, tiền xử lý, xây dựng mô hình, lựa chọn tham số và đánh giá mô hình, ứng dụng thực tế
- Vấn đề khi tiến hành thu thập dữ liệu dùng cho bài toán khai phá dữ liệu/học máy.
- Cần theo các bước của quy trình thu thập và tiền xử lý dữ liệu.
- Cần hiểu ý nghĩa trong từng bước của quy trình.

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict