UIT
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

NHẬN DẠNG
PATTERN RECOGNITION

MACHINE LEARNING & ML Life Cycle

Recap: Máy Học Là Gì ?

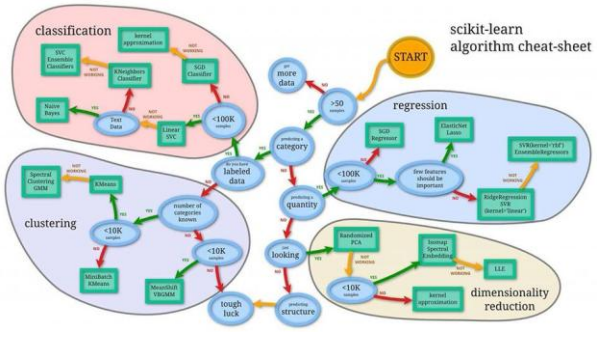
- Ta nói một máy tính *có khả năng học* nếu nó tự cải thiện hiệu suất hoạt động ***P*** cho một công việc ***T*** cụ thể, dựa vào kinh nghiệm ***E*** của nó.
- Như vậy *một bài toán học máy* có thể biểu diễn bằng 1 bộ (***T, P, E***)
  - ***T:*** một công việc (nhiệm vụ)
  - ***P:*** tiêu chí đánh giá hiệu năng
  - ***E:*** kinh nghiệm

(from Eric Xing lecture notes)

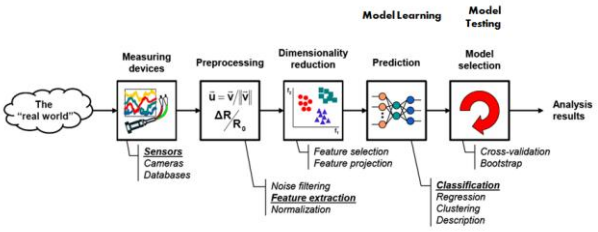
Recap: Phân loại dựa vào phương thức học

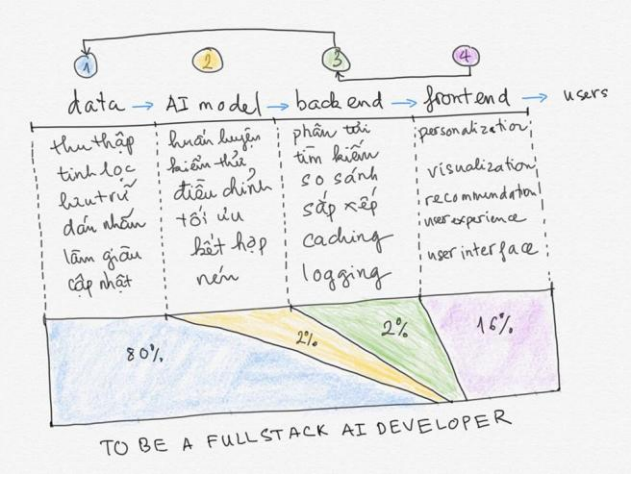
- **Supervised Learning** (Học có giám sát)
  - Classification (Phân loại, phân lớp)
  - Regression (Hồi quy)
- **Unsupervised Learning** (Học không giám sát)
  - Clustering (phân nhóm, cụm)
  - Association (luật)
- **Semi-Supervised Learning** (Học bán giám sát)
- **Reinforcement Learning** (Học Củng Cố)

Recap: Phân loại dựa vào phương thức học



Recap: QUÁ TRÌNH HOC MÁY CƠ BẢN

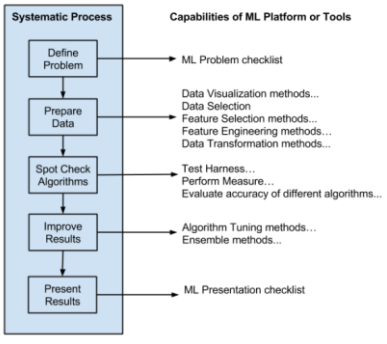




NỘI DUNG

- Định nghĩa bài toán
- Tiền xử lý dữ liệu
- Feature and Feature Enginering

GiẢI QUYẾT MỘT BÀI TOÁN BẰNG ML

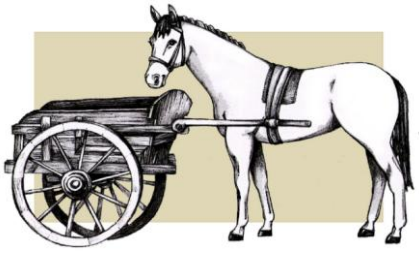


ĐỊNH NGHĨA BÀI TOÁN

- **What is the problem?**
  - Define the problem **informally and formally**.
  - List the **assumptions about the problem** (e.g. about the data).
  - List **known problems** similar to your problem.
- **Why does the problem need to be solved?**
  - Describe the **motivation for solving the problem**.
  - Describe the **benefits of the solution** (model or the predictions).
  - Describe **how the solution will be used**.
- **How could the problem be solved manually?**
  - Describe how the problem is **currently solved** (if at all).
  - Describe how a subject matter expert would **make manual predictions**.
  - Describe how a **programmer might hand code a classifier**.

9

ĐỊNH NGHĨA BÀI TOÁN



10

What is the problem?

Informal description

- Describe the problem **as though you were describing it to a friend or colleague**
- For example: *I need a program that will tell me which tweets will get retweets.*

Formalism

- Use this formalism to define the *T*, *P*, and *E* for your problem.
- For example:
  - **Task** (*T*): Classify a tweet that has not been published as going to get retweets or not.
  - **Experience** (*E*): A corpus of tweets for an account where some have retweets and some do not.
  - **Performance** (*P*): Classification accuracy, the number of tweets predicted correctly out of all tweets considered as a percentage.

11

What is the problem?

Assumptions

Create a list of assumptions about the problem and it's phrasing. These may be rules of thumb and domain specific information that you think will get you to a viable solution faster.

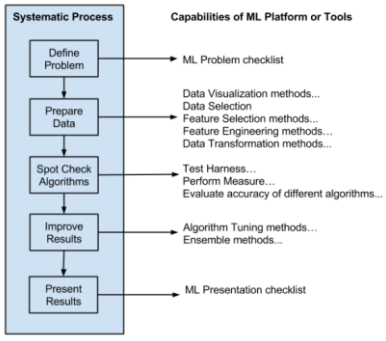
- For example:
  - The specific words used in the tweet matter to the model.
  - The specific user that retweets does not matter to the model.
  - The number of retweets may matter to the model.
  - Older tweets are less predictive than more recent tweets

Similar problems

- Other problems can inform the problem you are trying to solve by highlighting limitations in your phrasing of the problem
- Point to algorithms and data transformations that could be adopted to spot check performance.
- **For example:** A related problem would be email spam discrimination that uses text messages as input data and needs binary classification decision.

12

GiẢI QUYẾT MỘT BÀI TOÁN BẰNG ML



13

GATHERING DATA

Data for:

- Model training.
- Model evaluation.
- Model tuning.
- Model validation.

Question

- Has anybody done it before?
- What is the domain of your problem? is it related to Computer Vision, Natural Language Processing, Sensor data, or some XYZ?

14

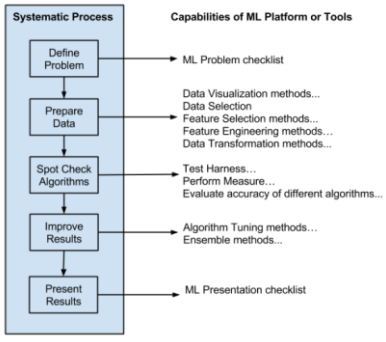
How Much Training Data is Required for Machine Learning?

- **It Depends; No One Can Tell You**
  - **The complexity of the problem**
  - **The complexity of the learning algorithm**
- **Reason by Analogy :** look at studies on problems similar to yours as an estimate for the amount of data that may be required.
- **Use Domain Expertise :** Use your domain knowledge, or find a domain expert
- **Use a Statistical Heuristic**
  - **Factor of the number of classes**
  - **Factor of the number of input features**
  - **Factor of the number of model parameters**
  - **Nonlinear Algorithms Need More Data**

Get More Data (No Matter What!?)

15

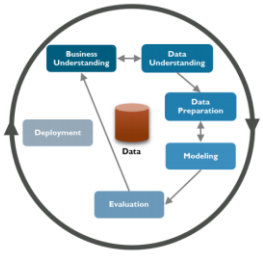
GiẢI QUYẾT MỘT BÀI TOÁN BẰNG ML



16

DATA PREPROCESSING

1. Data Engineering – 80%
   - Data extraction
   - Data cleaning
   - Data transformation
   - Data normalization
   - Feature extraction
2. Machine Learning – 20%
   - Model fitting
   - Hyperparameters tuning
   - Model evaluation

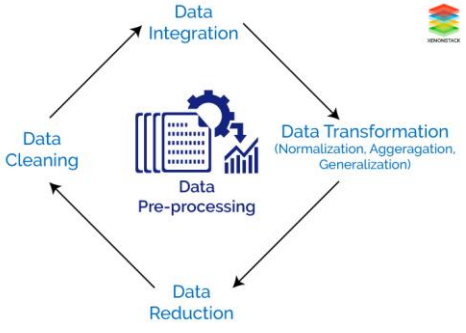


CRISP-DM

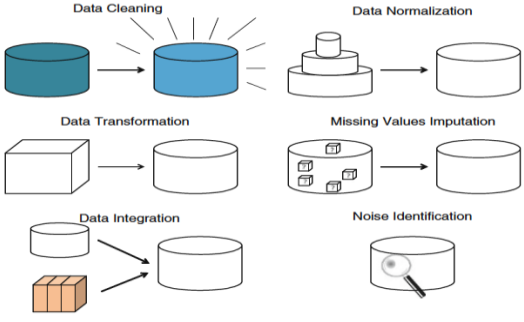
DATA PREPROCESSING

- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - e.g., occupation=""
  - noisy: containing errors or outliers
    - e.g., Salary="-10", Age="222"
  - inconsistent: containing discrepancies in codes or names
    - e.g., Age="42" Birthday="03/07/1997"
    - e.g., Was rating "1,2,3", now rating "A, B, C"
    - e.g., discrepancy between duplicate records
    - e.g., ***Endereço:*** travessa da Igreja de Nevogilde ***Freguesia:*** Paranhos

DATA PREPROCESSING



DATA PREPROCESSING



WHY DATA PREPROCESSING

- Preparing data also prepares the miner so that when using prepared data the miner produces better models, faster
- GIGO- good data is a prerequisite for producing effective models of any type

DATA PREPROCESSING

- **Data Cleaning**
  - Correct bad data, filter some incorrect data out of the data set and reduce the unnecessary detail of data.

Original data

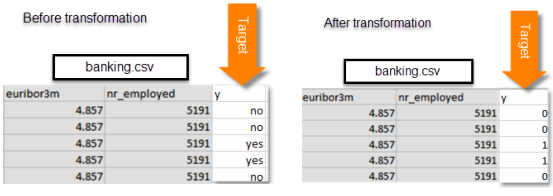
Name	Date	Duration (s)	Genre	Plays
Highway star	1984-05-24	-	Rock	139
Blues alive	1990/03/01	281	Blues	239
Lonely planet	2002-11-19	5:32s	Techno	42
Dance, dance	02/23/1983	312	Disco	N/A
The wall	1943-01-20	218	Reagge	83
Offside down	1965-02-19	4 minutes	Techno	895
The alchemist	2001-11-21	418	Bluesss	178
Bring me down	18-10-98	328	Classic	21
The scarecrow	1994-10-12	269	Rock	734

Cleaned data

Name	Date	Duration (s)	Genre	Plays
Highway star	1984-05-24		Rock	139
Blues alive	1990-03-01	281	Blues	239
Lonely planet	2002-11-19	332	Techno	42
Dance, dance	1983-02-23	312	Disco	
The wall	1943-01-20	218	Reagge	83
Offside down	1965-02-19	240	Techno	895
The alchemist	2001-11-21	418	Blues	178
Bring me down	1998-10-18	328	Classic	21
The scarecrow	1994-10-12	269	Rock	734

DATA PREPROCESSING

- **Data Transformation (chuyển đổi)**
  - The data is consolidated so that the mining process result could be applied or may be more efficient.

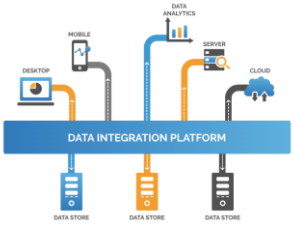


euribor3m	nr_employed	y
4.857	5191	no
4.857	5191	no
4.857	5191	yes
4.857	5191	yes
4.857	5191	no

euribor3m	nr_employed	y
4.857	5191	0
4.857	5191	0
4.857	5191	1
4.857	5191	1
4.857	5191	0

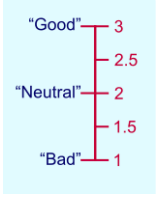
DATA PREPROCESSING

- **Data Integration**
  - Merging of data from multiple data stores.



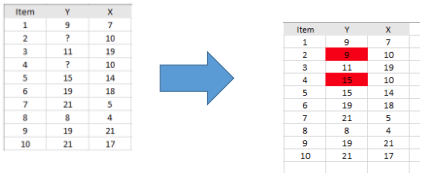
DATA PREPROCESSING

- **Data Normalization**
  - To express data in the same measurements units, scale or range.



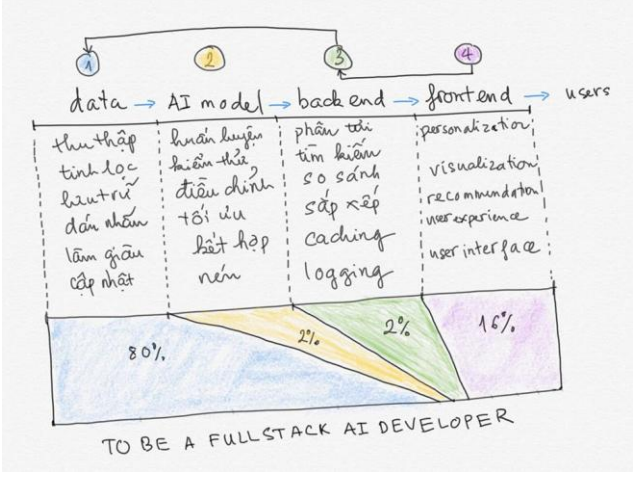
DATA PREPROCESSING

- **Missing Data Imputation**
  - To fill the variables that contain missing values with some intuitive data.



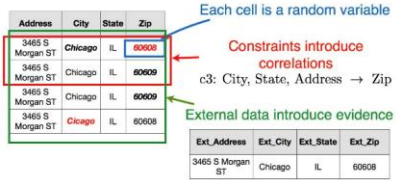
Item	Y	X
1	9	7
2	?	10
3	11	19
4	?	10
5	15	14
6	19	18
7	21	5
8	8	4
9	19	21
10	21	17

Item	Y	X
1	9	7
2	9	10
3	11	19
4	15	10
5	15	14
6	19	18
7	21	5
8	8	4
9	19	21
10	21	17



DATA PREPROCESSING

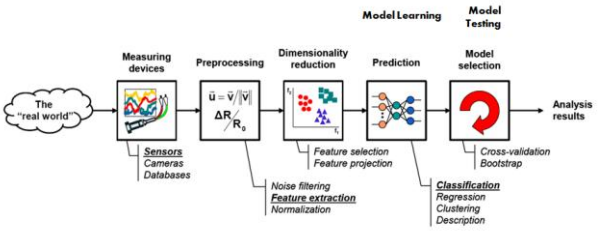
- **Noise Identification**
  - To detect random errors or variances in a measured variable.



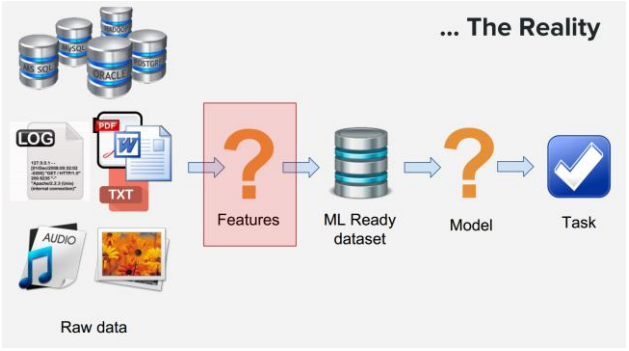
Address	City	State	Zip
3465 S Morgan ST	Chicago	IL	60608
3465 S Morgan ST	Chicago	IL	60609
3465 S Morgan ST	Chicago	IL	60609
3465 S Morgan ST	Cicago	IL	60608

Ext_Address	Ext_City	Ext_State	Ext_Zip
3465 S Morgan ST	Chicago	IL	60608

QUÁ TRÌNH HOC MÁY CƠ BẢN

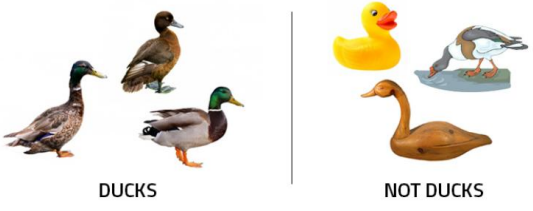


29

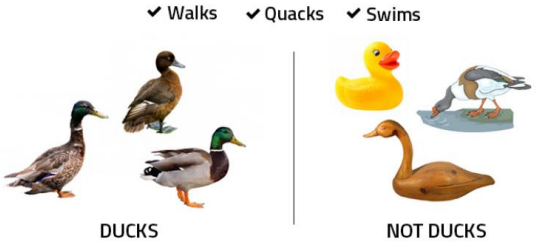


30

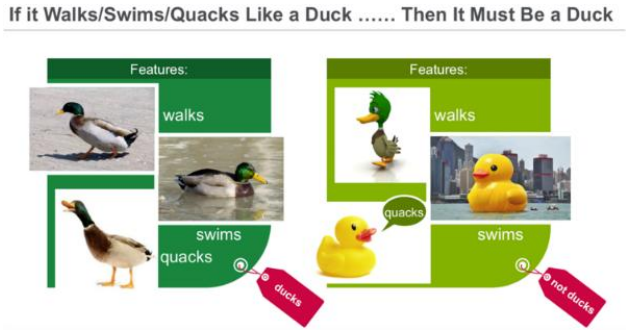
31



32



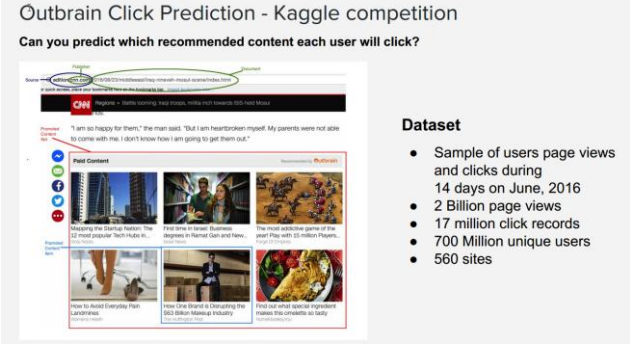
33



34

Feature ?

- *An attribute or group of attributes that constitute a characteristic property or set of properties which is* ***unique, measurable and differentiable*** .
- Feature is something that **makes a class different from another class**.



36

Feature Engineering

"Feature engineering is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data."
– Jason Brownlee

Feature Engineering

- Là qúa trình chuyển **đổi tập dữ liệu thô đầu vào** thành tập **các thuộc tính (features)** có thể giúp biểu diễn tập dữ liệu tốt hơn.
- Feature engineering cố gắng **biểu diễn tốt nhất** tập dữ liệu ban đầu sao cho **tương thích với mô hình dự đoán** đang sử dụng.

Feature Engineering

Bài toán: Phân loại sinh viên có học tốt hay không

- Sử dụng tất cả **các thuộc tính liên quan đến sinh viên** đó để áp dụng cho bài toán phân lớp, mà các thuộc tính này **thường rất nhiều từ 20-50 cột thuộc tính.**

➔**Thời gian để máy traing rất lâu**, đồng thời **kết qủa dự đoán có độ chính xác thấp.**

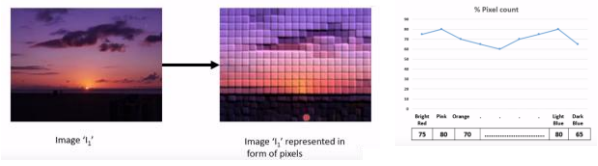
➔Sử dụng kĩ thuật feature engineering để **chọn ra một số thuộc tính phù hợp hơn như có vay mượn để đóng học phí không, số điểm đầu vào là bao nhiêu, qúa trình tiến bộ trong học tập là bao nhiêu**,...

Một số bài toán trong Feature Engineering

- **Bài toán rút trích đặc trưng (feature extraction)**
- **Bài toán đánh giá độ hữu dụng của các feature**
- **Bài toán lựa chọn đặc trưng (feature selection)**
- **Bài toán xây dựng đặc trưng mới (feature construction)**
- **Bài toán xác định feature thông qua training dữ liệu (feature learning).**

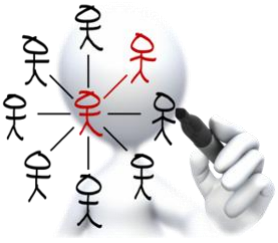
Feature extraction

- Là tiến trình tự động hoá được dùng để **giảm số chiều dữ liệu** sao cho dữ liệu ban đầu được chuyển đổi sang dạng **đơn giản và nhỏ hơn dữ liệu ban đầu,** trước khi đưa vào mô hình dự đoán



Feature Selection

- Tự động **hoá lựa chọn tập con trong số các feature** ban đầu sao cho các **feature được lựa chọn này phù hợp với bài toán hiện tại**



Chủ đề seminar

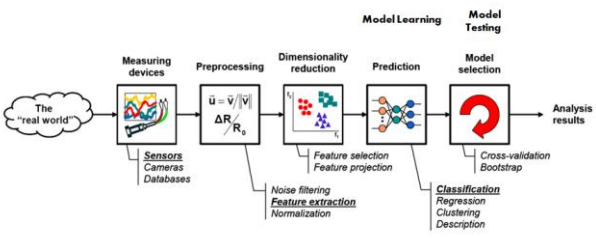
Trình bày một số thuật toán thường được sử dụng trong quá trình Feature selection:

- **Pearson Correlation**
- **Chi-Squared**
- **Recursive Feature Elimination**
- **Lasso**
- **Tree-based**

Tài liệu tham khảo

- https://www.coursera.org/learn/machine-learning?
- https://machinelearningcoban.com/
- https://ongxuanhong.wordpress.com/2015/10/29/feature-engineering-la-gi/
- https://machinelearningmastery.com/discover-feature-engineering-how-to-engineer-features-and-how-to-get-good-at-it/
- Lê Công Thành - InfoRe Technology

QUÁ TRÌNH HOC MÁY CƠ BẢN



45

46