Data Mining

Table des matières

1 Introduction

- Managing the data (Managing DataBases)
- Analysing the data (extract the information)
- Reporting the information (Business Intelligence)

What is data mining ?

Extract information that was not known before. This extraction is automatic or semi-automatic. The goal is to discover meaningful patterns.

It can be divided in 5 parts :

1. Input data
2. Data Preprocessing (feature selection, dimensionality reduction [1], normalization [2], data sub-setting [3])
3. Data Mining
4. Post Processing (filtering patterns, visualisation, pattern interpretation)
5. Information

What is not data mining ?

- Query web search engine
- Look up data in database

2 Data Mining Tasks

Prediction Methods : Use some variables to predict unknown or future values of other variables. (find a function for a correlation between multiples variables and use it to predict the future)

Description Methods : Find human interpretable patterns that describe the data

But also :

— Clustering
— Predictive Modeling
— Association Rules
— Anomaly Detection

1. Dimension is the number of variables. Visualising more than 3 can be complex, so we need to reduce it to 2 or 3. The most famous algorithm to do so are PCA, T-SNE, LDA
2. Normalisation is bringing all the value in the same scale 0-1(removing extremums and finding min and max)
3. Doing the analysing on batches of the database and not all the data at once. Ex : on the first 5 lines, then 5 next, ect ...

3 Machine Learning

- supervised learning
  - Classification
  - Regression
- Unsupervised learning
  - Clustering
- Reinforcement Learning

3.1 Predictive Modeling : Classification

First step is to split the DB in 2 : The train dataset and the test dataset (ex : 70% and 30%). The 2 sets are randomly selected.

!! In Science, splitting randomly is not enough. The model used is the cross-validation (k-fold) and the average of all the set is the result !!

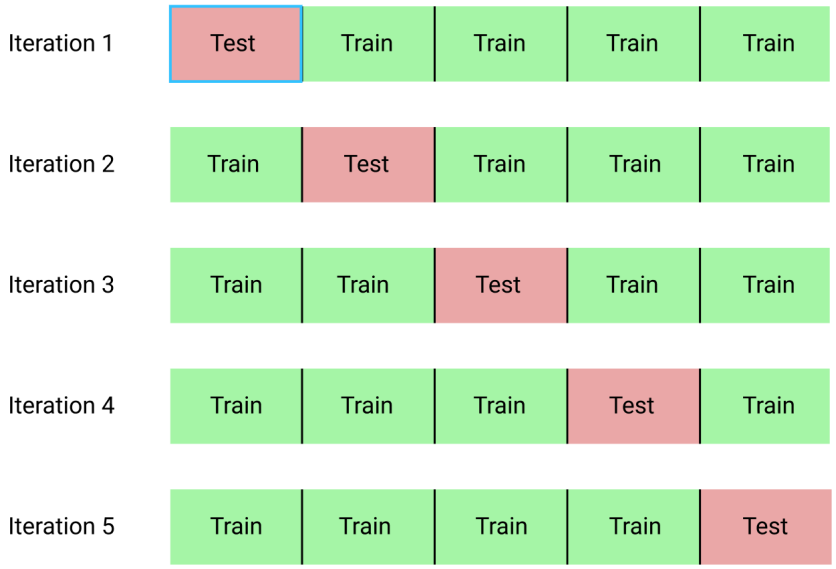

FIGURE 1 – 5-fold split

Exemple of application :

- classifying forest, water bodies, urban area,...
- identify intruder in cyberspaces
- predicting tumour cells as benign or malignant
- classifying proteins

3.2 Predictive Modeling : Regression

Predict a value of a given continuous variable based on other variables assuming a linear or non-linear model of dependency.
It's widely studied in statistics.

3.3 Clustering

Find group of objects such as distance between object of the same group is small and distance between groups is large.
It can be used to segment the market, group customer by their lifestyle, location, ...or group product by similarities.

4 AI-900 Azure AI Fundamentals

What is AI ?

It's a software that can imitate human capabilities and can make decision based on data and experience.

— Computer Vision
— Neural Language Processing
— Conversational AI = BOT
— Machine Learning
— Anomaly Detection

Challenges and risks

Challenge or Risk	Example
Bias can affect result	discrimination by gender via the data used for the training
Errors may cause harm	autonomous vehicle failure can kill someone
Data could be exposed	sensitive medical data are used to train the data
solution may not work on everyone	a predictive app do not provide audio output for visually deficient people
user must trust a complex system	we don't always understand the AI logic and believing it can be hard
who is liable for AI-driven decision ?	if someone is false accused, who is responsible ?

5 Machine Learning

For machine learning and computer vision, we have 2 sets of data : train and test. But for deep learning, we need 3 sets of data : train, validation (by back-propagation) and test.

5.1 Evaluating your model

To evaluate a model, you can use MAE (Mean absolute error, lower the better) or RMSE (Root mean squared error), RSE (Relative squared error, between 0 and 1, lower is better), R^2 (closest to 1 the best).

6 Computer Vision

[4]

Each image is divided in 3 matrix : RGB.
Each matrix is the size of the picture (pixel)and each value is between 0 and 255.

Applications

- Image classification (label for an image)
- Object detection (detect objects on a picture and draw a box arround it)
- Semantic segmentation = pixel classification (object detection pixel sized : the object is highlighted)
- image Analysis = image captioning (describe what is in the image)
- Face detection & recognition
- Optical character recognition (detect text on picture)

7 NLP (Natural Language Processing)

- Text Analysis and entity recognition
- Sentiment analysis
- Speech recognition and synthesis
- Machine translation
- Semantic language modeling

It can be decompose on different field :

Text Analytic	Language detection - Key phrase extraction - entity detection - sentiment analysis
Speech	Text to speech - speech to text - speech translation
Translator text	text translation

8 Conversation AI (BOT)

A solution that enable dialogue between AI and human. It's present in mail, web chat, social media, voice.

It must respect some principles

1. Be transparent about what the BOT can and cannot do
2. Male it clear that the user is communicating with a BOT
3. Enable the BOT to give hand to an human
4. Ensure the BOT respect cultural norms
5. Ensure the BOT is reliable
6. Respect user privacy
7. Handle data securely
8. Ensure the BOT meet accessibility standards
9. Assume accountability for the BOT action

4. see jupyter notebook