1.3.2 | Crafting a Data Mining Problem Statement Part 2

In the previous session, you have learned the steps in solving a data mining problem.

In this session, you will learn the five primary questions that data mining asks in general.

At the end of the session, you should be able:

1. Make distinctions of the five primary questions that data mining asks in solving problems;
2. Identify the type of a data mining question; and
3. Craft a data mining question.

What was the data mining problem that you have formulated in the previous session? Does it fall under the following question categories? If not, then you should refine to align it to any of them.

How much? How many?

Is it type A or type B?

How is this organized

Is it a weird behavior?

What should be done next?

💪Let's test your understanding of the five types of questions.💪

Formulate five questions about any prevailing condition or problem that you want data mining to solve.

Tap the button to share your answer to the other learners.

(https://drive.google.com/drive/folders/14zb87CGI6DvXOwwuZXk4EXBBrK6K9eWR?usp=sharing)

Question 1. How much? Or How many?

More companies are trying to get into the mask-making business as everybody scrounge for protection confronting the COVID 19 pandemic. Suppose that you are working in one of these companies, and you are thinking of striking the need of protective gears while the demand is very high. Will you suggest that your company:

1. Should venture into the production of other types of personal protective equipment? or
2. Should your company not venture into the production of personal protective equipment and venture into AI-driven machines such as the Danish Sanitizing Robot instead?

To answer these questions, you need to formulate a "How much" question which can result in numeric prediction.

Try to formulate "How much" type of question. Save it for next session's learning activity.

Tap the button for a hint.

Isn't it that the expected sales or budget is a numeric value? What data are needed to answer the questions?

Collate relevant data that can answer the question, share it to the course bank. Collaborate in aggregating your dataset.

(https://drive.google.com/drive/folders/1SLGNtfia18myADsPJb8_wgLU_h4xkcln?usp=sharing)

Regression

To solve a numeric prediction problem, create a regression model. This is done by finding the relationship between variables. It is vital to include variables that you believe are significant for the model. However, you can also include variables that may not affect the model.

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Models for Numeric Prediction

The following model can mathematically represent the problem of regression.

```
Model : Response = f(Variables)
```

When you model the function between response and other variables, you can make use of the following types of regression modelling methods.

- Linear Regression Models
  - Univariate as a function of a single variable
  - Multivariate as a function of multiple variables
- Tree Models for Regression
  - Decision Trees based on rules
  - An ensemble like Random Forest
- Neural Network for Regression
  - Prediction based on a complex interaction of variables

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Question 2. Is it a Type A or Type B?

Practical Example

Suppose you want to answer a question, "Will the type of personal protective equipment pass the standard or not? There are two possible outcomes to this question.

1. Standard
2. Not Standard

This is a Type A or Type B question. Type A is the Standard outcome, and Type B is the Not Standard outcome. This problem is called "**Prediction of Classes.**"

Other examples of this type of question are:

1. Is the email spam or legit?
2. Is the sentiment positive or negative?

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Can you think of another "Type A or Type B" problem?

Submit your answer to the course bank

(https://drive.google.com/drive/folders/1CYN2yl-Jscx5ZqbVNgHQ8iVU6jwSdNYt?usp=sharing)

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Models for Class Prediction

The following model can mathematically represent the problem of classification.

```
Model : P (Class) = f(Variables)
```

For example, what is the probability of standard of a type of personal protective equipment in terms of other variables such as the manufacturer?

The probability is between 0 and 1.

- Probability = 0 means Not Standard
- Probability = 1 means Standard

It is always good to look into the effect of other variables in the outcome. For instance, the manufacturer will increase the probability of the personal protective equipment standard.

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

In class prediction modelling, the following methods can be considered:

- Logistic Regression
  - Univariate with probability as a function of a single variable
  - Multivariate with a probability depending on multiple variables
- Tree-Based Classification
  - A decision tree that is based on rules
  - The ensemble of trees like Random Forest
- Neural Network Classification
  - Prediction based on the complex interaction of variables.

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xIqFy2LTRekhzPzuj?usp=sharing)

Question 3. How is this organized?

Practical Example

Suppose you want to understand how data of a car company is organized in terms of purchases. By knowing the structure, you will know a specific customer group to target for a specific promotional campaign.

This type of problem can be answered using the clustering method. Clustering is about finding groups of data points that are close together but are far from other groups of points.

The concept of clustering depends on distance of data points. Some common methods to use are Euclidean distance and Cosine distance. The following are the common algorithms that are used for clustering.

- KMeans Clustering
  - Choose the number of clusters and the initial cluster centres
  - Include points in clusters based on their distance from centres.
- Hierarchical Clustering
  - Connect points to one another based on their mutual distance
  - Connect groups pf points step-by-step to find optimal clusters

---

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xIqFy2LTRekhzPzuj?usp=sharing)

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xIqFy2LTRekhzPzuj?usp=sharing)

Question 4. Is it a weird behaviour?

In this type of problem, you are looking for irregularities in data through the behaviour of the data points.

In data mining, these irregularities are called anomalies. Anomaly is an observation that greatly deviates from most of the other observations. Anomaly detection flags unexpected or unusual behaviours that can serve as the basis for the detection of problems.

Intrusion detection is commonly applied in computer networks, social media, but can also be applied in other fields such as healthcare, usually in monitoring conditions of patients.

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Models to Perform Anomaly Detection

The most common methods used for anomaly detection in data mining are clustering and finding of structures. These are performed through the following modelling methods.

- Cluster Analysis-based Detection
  - Create clusters in data to identify regular behaviour. Anomalies are those points that are far from all regular clusters.
- Nearest Neighbour Detection Model
  - Compute distances of all points to their nearest neighbours. Anomalies are the points that are sparse within the neighbours.
- Support Vector Machine Detection
  - This method is based on decision boundaries and works well for large data with complex structure.

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Question 5. What should be done next?

This is commonly referred to as reinforcement or adaptive learning. Reinforcement-Learning is learning how to best react to situations, through trial and error. In machine learning, reinforcement learning is researched with respect to artificial decision-makers, referred to as agents.

Reinforcement learning enables an agent to learn from experience and adapt to new situations, without human intervention. Reinforcement learning agent is able to learn to make decisions on the sampling of the environment which provides the data.

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Models for Adaptive Learning

The following are the common approaches used for reinforcement learning.

- Monte Carlo
- State-Action-Reward
- Q-Learning
- Deep Reinforcement

Read this article on Reinforcement Learning in Self Driving Cars (https://towardsdatascience.com/reinforcement-learning-towards-general-ai-1bd68256c72d)

(https://drive.google.com/drive/folders/1eMJvs6xz2PMgX22xlqFy2LTRekhzPzuj?usp=sharing)

Craft data mining questions. Tap the button to submit your answer in the course bank.

(https://drive.google.com/drive/folders/18_KOlp8-7p0lynXWPBLtgnifQ1oLXF6T?usp=sharing)

(https://drive.google.com/drive/folders/1p6_XnSiBdyucrd5C09RGOgfNStyBpQVe?usp=sharing)

Tap the button to assess your understanding of crafting data mining problem statement.

Assignment

(https://tip.instructure.com/courses/9953/assignments/156975)