Data modelling in Machine Learning

machine learning pipeline can be broken down into three major steps. Data collection, data modelling and deployment. Lets see what are the steps involved in Data Modelling

Step 1. Problem definition

- Rephrase your business problem as a machine learning problem. It may be a supervised or unsupervised or Semi-supervised Problem.
- Let's take an example problem statement and rephrase it as ML problem. You have a CCTV cameras around your house and want to check whether the persons who are entering into the house are known before or not.

- You have hours of video footages of people moving in and out. Lets phrase it as ML problem, When a person enters you need to classify as known person or unknown person.
- Here classifying is the keyword to define it is a classification problem.

Data modelling in Machine Learning

Step 2. Data

- The data you have or need to collect will depend on the problem you want to solve. It may be different types
- **Structured data:** Think a table of rows and columns, an Excel spreadsheet of customer transactions, a database of patient records. Columns can be numerical, such as average heart rate, categorical, such as sex, or ordinal, such as chest pain intensity.
- **Unstructured data:** Anything not immediately able to be put into row and column format, images, audio files, natural language text.
- **Static data:** Existing historical data which is unlikely to change. Your companies customer purchase history is a good example.
- **Streaming data:** Data which is constantly updated, older records may be changed, newer records are constantly being added.

Data modelling in Machine Learning

Step 3. Evaluation

- There are different evaluation metrics for classification, regression and recommendation problems. Which one you choose will depend on your goal. (We already explained about performance metrics in our previous posts, please check them.)

Step 4. Features

- The three main types of features are categorical, continuous (or numerical) and Derived(Features you create from the data. Often referred to as feature engineering). Some important things to remember when it comes to features. Keep them the same during experimentation (training) and production (testing), Work with subject matter experts, Are they worth it? and feature leakage problem.

Data modelling in Machine Learning

Step 5. Modelling

- Modelling breaks into three parts, choosing a model, improving a model, comparing it with others.
- When choosing a model, you'll want to take into consideration, interpretability and ease to debug, amount of data, training and prediction limitations.
- improving a model involves changing hyperparameters such as learning rate or optimizer, etc...
- Comparing models

Data modelling in Machine Learning

Step 6. Experimentation

- This step involves all the other steps. Because machine learning is a highly iterative process, you'll want to make sure your experiments are actionable.
- Your biggest goal should be minimising the time between offline experiments and online experiments.
- Offline experiments are steps you take when your project isn't customer-facing yet. Online experiments happen when your machine learning model is in production.