Section 1.4: CRISP-DM ML Process

Monday, September 5, 2022 17:15

[00:24] Plan

- CRISP-DM -- methodology for organizing ML projects.
- From problem understanding to deployment.
- Spam detection example.

CRISP-DM stands for "CRoss-Industry Standards Processing for Data Mining."

This lesson serves to take a step back from the machine learning process and get a larger picture as to the process's methodology.

[00:57] ML Projects:

- Understand the problem.
- Collect the data.
- Train the model.
- Use it.

[1:19] Spam Detection:

Recall the spam example, where you get some mail; we extract some **features** from the email, we put them in the **model**, and the **model** generates a **score** for each email. If the **model declares the message to be scored lower than say, 0.5, the message does not go to the spam folder**, but if the **model declares the message score is greater than or equal to 0.5, the message does go to the spam folder; it goes to the inbox**.

[02:03] CRISP-DM:

CRISP DM is an industry-wide methodology standard which outlines exactly how machine learning projects should be organized. The methodology was invented in the 1990s by IBM, but is still useful today with almost no modifications.

There are {6} steps:

- {1} Business Understanding
  - Identify the problem.
  - Try to understand the problem.
  - Understand how we measure the success of the project.
  - Consider: Do we actually need ML here?
- {2} Data Understanding
- {3} Data Preparation
- {4} Modeling
  - Where we train our model.
- {5} Evaluation
- {6} Deployment
  - When our model is used.

[04:02] {1} Business Understanding:

- Our users complain about spam.
- Analyze to what extent it is a problem.
    - Is it a lot of users who complain, or perhaps one user who complains a lot?
    - Perhaps consider how impactful our project is.
    - Motivate investing time into the project.
- Will machine learning help?
    - Consider the possibility that we may be fine just with developing a rule-based system; or developing some sort of heuristic without investing a lot of time and resources into developing a machine learning system.
- If not: propose an alternative solution.

If we decide to go with ML:

- Define the goal:
    - Reduce the amount of spam messages; or
    - Reduce the amount of complaints about spam messages.
- The goal has to be measurable:
    - For example, you want to reduce the amount of spam by 50%.

If you come up with a metric for measuring spam messages, you can develop a quantifiable goal for how much you want to reduce spam messages. Consider how success for the project is measured.

Once we determine that ML is the right tool for this problem, we transition to the next step in the CRISP-DM methodology. We do this by trying to understand what data is available to us to solve the problem. We need to have data; no data, no ML. In any case, data needs to be collected, bought, etc.

[06:45] {2} Data Understanding:

Identify the data sources:

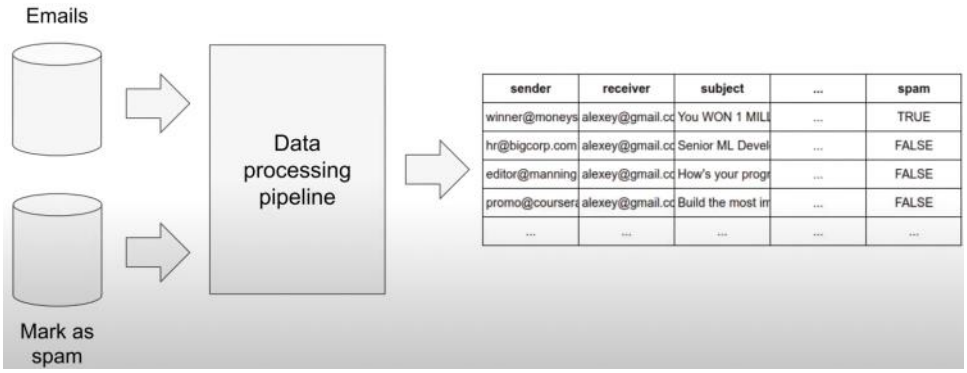
- We have a report spam button.
- Is the data behind this button good enough?
- Is it reliable?
- Do we track it correctly?
- Is the dataset large enough?
- Do we need to get more data?
- It may influence the goal.
- We may go back to the previous step and make adjustments.

After data understanding, we move on to data preparation. This is where we transform the data so it can be put into a ML algorithm. We have already assessed the data to be usable, reliable, and sufficient to be incorporated into the ML process.

[10:03] {3} Data Preparation:

- Clean the data; remove all the excess noise.
- Build the pipelines.
    - A sequence of steps that takes the raw data that cleans it and produces clean, ready-to-use data.

- Convert into tabular format.
    - Tabular format means something that can be put into an ML model, which occurs during the data processing pipeline.



Once the data is prepared, we transition to training the models: the actual machine learning happens here.

[12:13] {4} Modeling:

- Try different models and select the best one.
- There are an assortment of models from which to choose:
    - Logistic regression
    - Decision tree
    - Neural networks
    - Among other types…

The next lesson will cover more about how to choose a model; however, sometimes we may have to go back to data preparation:

- Add new features.
- Fix data issues.

After we select the best possible model, we need to evaluate how well the model performs in terms of solving the problem.

[13:37] {5} Evaluation:

Since evaluation assessment is important, we must begin to ask some questions:

- Is the model good enough?
    - Have we reached the goal?
    - Do our metrics improve?
- Goal: reduce the amount of spam by 50%.
    - Have we reduced it? By how much?
    - (Evaluate on the test group)
- Do a retrospective:
    - Was the goal achievable?
    - Did we solve/measure the right thing?
- After that, we may decide to:
    - Go back and adjust the goal.
    - Roll the model to more users/all users.

    - Stop working on the project.

Evaluation and deployment are often executed concurrently.

- Online evaluation: evaluation of live users.
- It means: deploy the model, evaluate it.

[16:13] {6}Deployment:

- Roll the model to all users.
- Proper monitoring.
- Ensuring the quality and maintainability.

[17:05] Iterate!:

ML projects require many iterations! You do not stop at the deployment, you learn from the process, make adjustments, and continue gaining an understanding about the project as a whole. You do this so you converge closer towards your objective.

Start simple and learn from feedback so you can improve. You do two or three iterations and resume you journey.

[18:53] Summary:

- Business understanding: define a measurable goal.
    - Ask: do we need ML?
- Data understanding: do we have the data?
    - Is it good?
- Data preparation: transform the data into a table, so we can put it into ML.
- Modeling: to select the best model, use the validation set.
- Evaluation: validate that the goal is reached.
- Deployment: roll out to production to all the users.
- Iterate: start simple, learn from feedback, improve.

Notes by Community:

CRISP-DM is a methodology for organizing ML projects. It was invented in the 90s by IBM. The steps of this procedure are:

- **Business understanding:** An important question is if do we need ML for the project. The goal of the project has to be measurable.
- **Data understanding:** Analyze available data sources, and decide if more data is required.
- **Data preparation:** Clean data and remove noise applying pipelines, and the data should be converted to a tabular format, so we can put it into ML.
- **Modeling:** training Different models and choose the best one. Considering the results of this step, it is proper to decide if is required to add new features or fix data issues.
- **Evaluation:** Measure how well the model is performing and if it solves the business problem.
- **Deployment:** Roll out to production to all the users. The evaluation and deployment often happen together - **online evaluation**.

It is important to consider how well maintainable the project is.
In general, ML projects require many iterations.