Machine Learning for Business Applications
46-887

Carnegie Mellon University
Tepper School of Business

Homework 1

Overview

The goals of this first assignment are:

1. To choose a business problem that your team wants to address during the remainder of the course using Machine Learning (ML).

2. To find *at least* 1 source (from a textbook, from a blog post, ...) where this problem has already been solved using ML.

3. To find *at least* 1 relevant dataset that will allow your team to build a Proof Of Concept (POC) ML system to address the business problem of your choice over the next few weeks. There is a good chance that relevant datasets are already mentioned in the sources that your team will find as part of the work associated with the previous goal.

In one sentence: **by the end of this first assignment, your team needs to be clear on what it will be working on for the next few weeks, and it needs to be confident that the chosen business problem has *some* existing baseline solution.**

The best case scenario is that by the end of the project your team will have implemented a different or improved solution. But your team also needs to mitigate the risk of the worst case scenario - i.e., not being able to deliver *any* reasonably decent POC ML system by the end of the course - as early as possible. This is precisely why your team will find sources that already addressed the chosen business problem as part of this assignment.

Deliverable

Your team will submit a project proposal of no more than 3 pages describing:

1. The business problem that your team will focus on.

2. How this problem has already been solved in the past using ML (from the sources you found).

3. The datasets that your team has identified.

Your team is welcome to include screenshots and other figures as needed. These do not count towards the 3 page limit.

Your team can submit the project proposal as a PDF document or as a Word document.

Checklist

Here is a checklist to help your team make sure that the most important items are covered in the project proposal.

- ☐ Include a description of the business problem that your team wants to address and why it is an interesting problem to solve from a business perspective. Does solving this problem help making money? Saving money? Saving time? For instance: *Given a collection of text documents, we want to help users identify the most frequent topics represented in the collection. This is useful in a number of practical settings. For example: what are the characteristics of a product that are most frequently discussed by its users in their reviews of that product? Developing a better understanding of the perception and experiences that the users have with the product can help a company make money by introducing new features for which the users explicitly indicate interest, and it can save time and money because it provides immediately actionable user feedback without the need of conducting more extensive and more expensive user research.*

- ☐ Include at least 1 source where this business problem has been solved using ML. For instance: *This type of problem - generally referred to as "topic modeling" - has been addressed in the past in a few ways. For example, SOURCE uses a Latent Dirichlet Allocation (LDA) model to learn the topics represented in COLLECTION OF DOCUMENTS. Similarly, OTHER SOURCE uses BERTTopic to learn the topics represented in OTHER COLLECTION OF DOCUMENTS.*

- ☐ Explain the characteristics of the ML models/techniques that were used in each of the references that your team has cited. For instance: *At a high level, LDA works as follows: HIGH LEVEL EXPLANATION OF LDA. On the other hand, BERTTopic works as follows: HIGH LEVEL EXPLANATION OF BERTTopic. In general, the advantages of LDA are that ADVANTAGES OF LDA, whereas it has the disadvantages that DISADVANTAGES OF LDA. On the other hand, BERTTopic has the advantages that ADVANTAGES OF BERTTopic, whereas its disadvantages are that DISADVANTAGES OF BERTTopic.*

- ☐ Include at least 1 reference to a dataset that will allow your team to build a similar ML system during the course. For instance: *We could use the dataset DATASET mentioned in SOURCE to build a POC ML system for topic modeling. The data is available at LINK/WEBSITE. Alternatively, we could also use OTHER DATASET, mentioned in OTHER SOURCE. This other dataset can be downloaded at LINK/WEBSITE.*

- ☐ Include a description of the data available in each of the datasets that your team has cited. For instance: *The data in DATASET include NUMBER OF DOCUMENTS IN DATASET for a total of NUMBER OF WORDS IN DATASET. Additional information in DATASET includes ADDITIONAL INFORMATION IN DATASET. Similarly, the data in OTHER DATASET includes NUMBER OF DOCUMENTS IN OTHER DATASET for a total of NUMBER OF WORDS IN OTHER DATASET. OTHER DATASET also inludes ADDITIONAL INFORMATION IN OTHER DATASET.*

Other notes

- Keep in mind that your team's goal in the course is to build a *POC* ML system. It is advisable to keep things simple - or at least manageable - when it comes to the choice of the data and of the models that your team will experiment with. A good rule of thumb: if your team can't build its POC ML system on a modern laptop - e.g., because the dataset your team selected is too large for it, because the model your team is trying to train/fit requires more computing power, ...- it likely means that the POC ML system your team is trying to build is too complex (for this course, anyway).

- For this project, it is advisable that your team chooses a business problem that can be addressed using an "online" ML system. Business problems that can be framed in terms of e.g., regression, classification, recommendation of contents, search and retrieval, ... tend to lend themselves better to online inference than business problems that can be instead framed in terms of e.g., clustering, dimensionality reduction, ... This is not a strict requirement for the project. It will just make your team's life a little easier in the next few weeks.

- Your team is free to manage the project's code and other assets in any way the team sees fit. However, it is advisable to make good use of a version control tool such as git. Consider creating a repository on GitHub for your team's work. If you have never used a version control tool in the past, this is a good opportunity to learn the basics.