Topics in Machine Learning

AU STAT-427/627

Emil Hvitfeldt

2021-10-18

We have in this class talked about how models work, and how the machine learning process is structured

But these technical questions are only a small part of the whole machine learning pipeline

We start at the top (simplified), developing a question we want to get an answer to. Consult exports, collect data, validation. Many steps go into the process.

Should we even be doing this?

This is always the first step. Machine learning algorithms involve math and data, but that does not mean they are neutral. They can be used for purposes that are helpful, harmful, or even unethical.

What bias is already in the data?

It is important to note that bias can be seen in both statistical and non-statistical

Does the data has deviations from the "normal"?

This can be due to many different reasons including data collection

Data collection is an inherently human task and the biases of the data collectors will be represented in data

Can the code and data be audited?

Consider how accessible your code and data are to internal and external stakeholders.

What are the error rates for sub-groups?

This is often described in terms of race, but any kind of stratification matters

If you think of this in terms of residuals, you should preferably want equal distribution of residuals across sub-groups if the subgroups are unrelated to the question at hand

What is the accuracy of a simple rule-based alternative?

Straightforward heuristics are easy to implement, maintain, and audit, compared to machine learning models; consider comparing the accuracy of models to simpler options.

What processes are in place to handle appeals or mistakes?

If you build a model and put it into production by your organization, what would happen if a complaint was classified incorrectly? We as data practitioners typically (hopefully) have a reasonable estimate of the true positive rate and true negative rate for models we train, so processes to handle misclassifications can be built with a good understanding of how often they will be used.

How diverse is the team that built it?

If you are working with healthcare data; have healthcare professionals in your team

If you are working with language data; have people on the team speaking those languages

Does the data we are using for training represent the data we want to apply the model on?

This should be your main concern when working with Machine Learning Models

Please Watch This!

Final Project

Search for data!

Ideas:

- https://github.com/rfordatascience/tidytuesday

- https://www.data-is-plural.com/

- https://www.kaggle.com/