Model Prototyping:
Research vs. Production

About This Reading

Once an ML model has been developed by a research team, there are a lot of steps required to convert it into a production-ready model. This is because many of the characteristics of an ML model are different in a production environment. These include characteristics related to the data, model training, overall goal, priority of metrics, and additional constraints that are not present in the research environment. In this reading, we will explore some of these differences.

	Academic/Research ML	**Production ML**
Data	Static	Dynamic - shifting
Priority for Design	Highest overall accuracy	Fast inference, good interpretability
Model Training	Optimal tuning and training	Continuously assess and retrain
Fairness	Very important	Crucial

Differences

Data. The most obvious difference in the production environment is the dynamicity of data. Data is often being collected continuously, which also means the data distribution will be continuously changing. Furthermore, the overall data size will get larger over time. Thus, our ML models have to be configured to take in and take advantage of new data, while ensuring it is able to handle changes in the data distribution which might affect the served predictions.

Priority for Design. The goals in the research environment and production environment are often misaligned. The priority in an academic context is often a single metric, which is typically accuracy. Skim through half a dozen ML papers and you'll typically see accuracy being compared between a variety of models on a particular task. However, this single-mindedness often does not translate into the production environment, where things like speed, efficiency, and interpretability become important. After all, what good is an algorithm that gives 100% accuracy if it is impossibly slow and also violates the GDPR rule on the right to the explanation? If a model is planning to be productized, these additional metrics should be considered alongside accuracy so that the goals in the research and production environments are appropriately aligned.

Model Training. In a research context, the models are often only being developed to obtain a specific metric to be placed in a journal article, conference paper, or PowerPoint presentation. This is fundamentally different from a production ML environment, where the goal is high performance at a minimal cost. This means that instead of excessively fine-tuning a single model, we care about continuously assessing a model's performance over time and retraining it once a significant amount of new data is available or the performance metrics begin to dip below an acceptable limit. Setting up an environment for continuous monitoring and retraining requires a lot of additional patience, expertise, and resources on top of the initial development of the model.

Fairness. We already touched on this slightly, but in some jurisdictions - most notably the European Union - there are stipulations that ML models must be able to "explain" how they arrived at their predictions. This is often used in circumstances such as when deciding if an individual is given a bank loan for transparency purposes, giving the affected party the opportunity to see on what basis their bank loan application was rejected. For models in production, ensuring algorithmic fairness is crucial for legal compliance as well as ensuring equal opportunities to all potential algorithmic users. In a research context, less emphasis is generally placed on algorithmic fairness and interpretability since this often comes at the expense of performance. For example, the performance of a neural network is typically far superior to linear regression, but while linear regression is highly interpretable, neural networks are decidedly less interpretable.

Summary

As we have just discussed, research and production environments for ML are quite different in their characteristics. Typically, data scientists and machine learning engineers are responsible for the research aspect, while MLOps/DevOps teams and production engineers are responsible for the production environment. In an ideal world, these two sides would work together and align their practices with the end-user in mind. Next time you are developing a model, try to think more critically about how it will be implemented in a real-world application, whether the way you are designing and evaluating the performance of your model is correctly aligned with this, and how the production environment will vary from your current research environment.