Deployment of Machine Learning Models

Section 3.4 Notes

Pickle challenges:

As per the scikit-learn docs:

"Since a model internal representation may be different on two different architectures, dumping a model on one architecture and loading it on another architecture is not supported."

See: https://scikit-learn.org/stable/modules/model_persistence.html

On the risks of the pickle format:
https://www.youtube.com/watch?v=7KnfGDajDQw

The below resources are on more advanced topics that we will not be covering in the course

If you are interested in reading more about the streaming approach (pattern 3 from the lecture), this is a useful primer on machine learning with Apache Kafka:
https://www.confluent.io/blog/using-apache-kafka-drive-cutting-edge-machine-learning

And here are some code examples:
https://github.com/kaiwaehner/kafka-streams-machine-learning-examples

Here is a tutorial on working with Apache Spark for large scale data processing:
https://towardsdatascience.com/deep-learning-with-apache-spark-part-1-6d397c16abd

Here are some more advanced architecture discussions from larger companies:

Netflix on architecture for recommendation systems:
https://medium.com/netflix-techblog/system-architectures-for-personalization-and-recommendation-e081aa94b5d8

Google's TFX Paper: https://ai.google/research/pubs/pub46484

Uber's (very complex!) Michelangelo System: https://eng.uber.com/michelangelo/

Deployment of Machine Learning Models

Testing Machine Learning Systems: https://ai.google/research/pubs/pub45742