Course 2 Optional References

Machine Learning Data Lifecycle in Production

This is a compilation of resources including URLs and papers appearing in lecture videos. If you wish to dive more deeply into the topics covered this week, feel free to check out these optional references.

Overall resources:

Konstantinos, Katsiapis, Karmarkar, A., Altay, A., Zaks, A., Polyzotis, N., … Li, Z. (2020). Towards ML Engineering: A brief history of TensorFlow Extended (TFX). http://arxiv.org/abs/2010.02013

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2020). Challenges in deploying machine learning: A survey of case studies. http://arxiv.org/abs/2011.09926

Week 1: Collecting, Labeling and Validating Data

ML code fraction:

MLops

Data 1st class citizen

Runners app

Rules of ML

Bias in datasets

Logstash

Fluentd

Google Cloud Logging

AWS ElasticSearch

Azure Monitor

TFDV

Chebyshev distance

Sculley, D., Holt, G., Golovin, D., Davydov, E., & Phillips, T. (n.d.). Hidden technical debt in machine learning systems. Retrieved April 28, 2021, from Nips.cc https://papers.nips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf

Week 2: Feature Engineering, Transformation and Selection

Mapping raw data into feature

Feature engineering techniques

Facets

Embedding projector

Encoding features

TFX:

1. https://www.tensorflow.org/tfx/guide#tfx_pipelines
2. https://ai.googleblog.com/2017/02/preprocessing-for-machine-learning-with.html

Breast Cancer Dataset

Week 3: Data Journey and Data Storage

Data Versioning:

1. https://dvc.org/
2. https://git-lfs.github.com/

ML Metadata:

1. https://www.tensorflow.org/tfx/guide/mlmd#data_model
2. https://www.tensorflow.org/tfx/guide/understanding_custom_components

Chicago taxi trips data set:

1. https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/data
2. https://archive.ics.uci.edu/ml/datasets/covertype

Feast:

1. https://cloud.google.com/blog/products/ai-machine-learning/introducing-feast-an-open-source-feature-store-for-machine-learning
2. https://github.com/feast-dev/feast
3. https://blog.gojekengineering.com/feast-bridging-ml-models-and-data-efd06b7d1644

Week 4: Advanced Labeling, Augmentation and Data Preprocessing

Hand Labeling

Weak supervision

Snorkel

How do you get more data?

Advanced Techniques

Images in tensorflow

CIFAR-10

1. https://www.cs.toronto.edu/~kriz/cifar.html
2. https://www.tensorflow.org/datasets/catalog/cifar10

Weather dataset

Human Activity Recognition

Papers

Label Propagation:

Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2019). Label propagation for deep semi-supervised learning. https://arxiv.org/pdf/1904.04717.pdf

Slide 13 active learning:

Source: Original slides by Yale Cong