Analyze Datasets and Train ML Models using AutoML

Week 1

Data Ingestion:

- AWS Glue
- Data Wrangler: https://github.com/aws/aws-sdk-pandas/blob/main/tutorials/003%20-%20Amazon%20S3.ipynb

Data Visualization

- Pandas, numpy
- Matplotlib
- Seaborn

Week 2

Statistical Bias:

- Training data does not comprehensively represent the problem space.
- Some elements of a dataset are more heavily weighted or represented. For example: Fraud detection or product reviews.
- Reasons : Activity bias, societal bias, selection bias(man or woman etc.), Data drift
- Metrics :
    - Class Imbalance
    - Difference in Proportions of Label(DPL) : Measures the imbalance of positive outcomes between different facet values. For Example : Does a product_category has disproportionately higher ratings than others.
    - For all available metrics : https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-measure-data-bias.html
- Tools :
    - Amazon SageMaker Data Wrangler : It's only using subset of dataset for process.
    - Amazon SageMaker Clarify(SDK) : Eligible for big data. It's scalable.

Feature Importance:

- Explains the features that make up the training data using a score (importance).
- How useful or valuable the feature is relative to other features.
- Predict the sentiment for a product → Which features play a role?

Open Source Framework - SHapley Additive exPlanations(Shap)

- Shapley values based on game theory.
- Explain predictions of a ML model

        - Each feature value of training data instance is a player in a game
        - ML prediction is the payout
- Local vs global explanations
- SHAP can guarantee consistency and local accuracy.

In order to calculate feature importance:

1. Go to Sagemaker studio.
2. Select new data flow.
3. Select your data: from S3 or Athena.
4. Select your data file and import.
5. Click “plus” sign and choose add analysis.
6. Select Quick Model from analysis type.
7. Select your label. It uses %0.7 of data for training.
8. We can save the analysis.

Week 3

AutoML

Amazon SageMaker Autopilot eliminates the heavy lifting of building ML models. You simply provide a tabular dataset and select the target column to predict, and SageMaker Autopilot will automatically explore different solutions to find the best model. You then can directly deploy the model to production with just one click or iterate on the recommended solutions to further improve the model quality.

Week 4

Text Analytics Algorithms

Word2Vec
- Convert text into vectors called “embeddings”
- 300-dimensional vector space
- Perform machine learning on the vectors

Glove

FastText
- Extension of word2vec
- Breaks the word into character sets of length n (n-grams):
    - "amazon" => "a", "am", "ama", "amaz", "amazo", "amazon"
- Embedding for a word is the aggregate of the embedding of each n-gram within the word

Transformer
- Attention is all you need!

BlazingText
- Scales and accelerates Word2Vec using multiple CPUs or GPUs for training

- Extends FastText to use GPU acceleration with custom CUDA kernels
- Creates n-gram embeddings using CBOW and skip-gram
- Saves money by early-stopping a training job
    - when the validation accuracy stops increasing
- Optimized I/O for datasets stored in Amazon S3

ElMo(Embeddings from language models)

GPT

Bert

Resources

- Slides : https://community.deeplearning.ai/t/pds-course-1-lecture-notes/48242