Data Acquisition & Management

Introduction

- Applications of ML are now much more broad
  - Speech Recognition – Image Classification – Traffic Prediction, etc.
- Better Computational Power & Larger Amounts of available Data
- Cost of preparing data is expensive : collecting, cleaning, analyzing, visualizing data + feature engineering
- Data collection is extremely relevant in today
  - New applications → less available training data

Data Acquisition approaches

- Data Discovery : Sharing & Searching new datasets
- Data Augmentation: Enhancing/Extending existing datasets through external data
- Data Generation: …

Data Discovery

- Data Sharing: Datasets must be indexed and published for sharing
- Data Searching: Searching data for a specific ML task
  - Challenges:
    - Scaling search
    - Determining relevancy of search

Data Discovery : Data Sharing

- Collaborative Analysis:
  - Environment where data scientists collaboratively analyze and share different versions of data sets
  - DataHub:
    - Dataset version control system
    - UI platform provides → search, cleaning, integration & visualization of data.
- Web
  - Google Fusion Tables:
    - Cloud-base data management service → fusion tables used to gather data and support basic operations
    - Crawled by web-search engines
  - Data Marketplace: CKAN – DataMarket, etc.

- Both:
  - Kaggle

Data Discovery: Data Searching

- **Data Lakes**: Large repositories of data (often) stored in its raw format
  - Popular in corporate environments → (merge scattered data in common repo)
  - Tools to process & organize datasets inside data lakes have been deployed
    - IBM: Introduced system taking care of data wrangling
    - Google: GOODS → catalogues metadata of billions of datasets from many different storage systems
    - Data Civilizer: linkage graph to illustrate PK-FK relationships
  - Technical challenges:
    - (different data formats, computing relevancy, etc.)

Data Discovery: Data Searching

- **WEB**:
  - Much more diverse data → Table extraction techniques
    - Keyword searching
    - Row-subset queries & column seatch
    - Entity-attribute queries
  - WebTables: System used to extract structured data published online in the form of HTML tables

Data Augmentation

- Enhance existing dataset with external data

- *3 popular approaches:*
  - ***Deriving Latent Semantics***:
    - generate & use embeddings representing words, entities, etc.
    - Word2Vec: CBOW & Skip-Gram
  - ***Entity Augmentation***:
    - InfoGather: Uses web pages & fills in missing values of attributes.
  - ***Data integration***:
    - Hamlet System: determines if KFK joins are necessary → reduces total run-time

Data Generation : Crowdsourcing

- **Crowdsourcing** → Manual
  - Workers are rewarded to complete tasks and gather bits of information that will then become a generated dataset
    - Amazon Mechanical Turk

Data Generation : Crowdsourcing

- Data gathering
  - Closed world vs Open world data
  - AutoMan, Turkit, CrowdDB
  - Data augmentation & Gathering complementarity → CrowdFill
- Data preprocessing
  - Follows data gathering
  - Crowdsourced preprocessing tasks: Data curation, entity resolution & joining dataset
  - Data Tamer: Curation system uses crowd to accomplish attribute ID & Entity resolution
- Challenges: Quality Control & Maximizing worker productivity and background

Data Generation: Synthetic Data Generation

- Increase in popularity of SDG techniques due to lower cost (vs Crowdsourcing)
- **Generative Adversarial Networks (GANS):**
  - Synthetic images – videos – patient records, etc.
  - 2 contesting Neural Networks trained concurrently:
    - Generative Network → Learns to maps from latent space to a data distribution
    - Discriminative Network → Discriminates candidates produces by the generative network according to the true distribution.
- **Data Specific techniques:**
  - Synthetic image generation: used in object detection
    - Transformations to known data to create new entries
    - Synthetic text images → vary font, font-weight, font-size, etc.
  - Synthetic text generation:
    - Paraphrasing popular approach to generating new text
    - Syntactically controlled paraphrase networks (SCPN)
    - Semantically equivalent adversarial rules for text (SEARs