Training Data

Introduction

- This chp details data from data science prespective
- Discuss Common Challenges :
  1. Creating training data
  2. Label multiplicity problem
  3. lack of labels
  4. class imbalance
  5. tech in data augmentation
- As model evolves, training data as well evolves
- Caution: Data is full of potential bias

Sampling

- Where sampling happens:
  1. Sample training data from real world dataset
  2. Sample DS to train, validation, test splits
  3. Sample events for monitoring
- Scenarios:
  1. Dont have access to all possible data
  2. Infeasible to process all data
  3. Accomplish task faster and cheaper
- Families of Sampling
  - Non Probability Sampling
    - Caution: Will have sampling bias
    - Convenience Sampling
      - Sample data based on availability
    - Snowball Sampling
      - Future sample based on existing sample . Ex: Twitter followers scrapes
    - Judgment Sampling
      - Subject matter experts decides
    - Quota Sampling
      - Sample based on quota of slices
    - Example: Available text for language modeling, Sentiment analysis on general review text, Self-driving cars data
  - Probability Sampling
    - Simple Random Sampling
      - Pros: easy to Implement
        cons: Rare cases might not appear
    - Stratified Sampling
      - Cons: Impossible to divide sample into groups always
    - Weighted Sampling
      - Domain expertise knowledge
    - Reservoir Sampling
      - Sampling in streams of data
    - Importance Sampling
      - Used in Policy based reinforcement learning

Labeling

- Most ML models are supervised, so we need labels
- Stragegies
  - Hand Labels
    - Cons:
      1. Expensive
      2. Threat to data privacy
      3. Slow
    - May give rise to Label Multiplicity
    - Maintain label data lineage for debugging
  - Natural Labels
    - Automatically generated. Ex: google maps, recommender systems
    - Will have Implicit or Explicit feedback
    - Need to consider feedback loop length
- Handling Lack of labels
  - Weak supervision
    - Based on heuristics, noisy labels
    - Can use aggregation strategy to reduce noisy labels effect
  - Semi Supervision
    - Leverages structural assumptions
      High prob score are assumed correct self-training
    - Based on co-occurrence characteristics
      Based on Clustering
      Based on Perturbation
  - Transfer Learning
    - First model is used as a starting point for downstream task
      First model is trained with/without label data
  - Active Learning
    - Choose the data point to learn from
      1. Based on prob score
      2. Based on disagreement
      3. Based on gradient or loss update

Class Imbalance

- Example:
  Classification - Normal cell vs Cancerous Cell
  Regression - Estimating health care bills
- Challenges of Class Imbalance
  - ML algo works well for balanced data and usually not well for unbalanced data
  - Reason 1: Insufficient signal from minority classes
    Reason 2: Stuck at non optimal solutions
    Reason 3: Leads to Asymmetric cost of error
- Classical examples : Fraud detection, churn prediction, disease screening
- Other Reason for Class imbalance : Due to Sampling bias, labelling errors
- Handling Class Imbalance
  - Using right evaluation metrics
    - Acc and Error rate gives equal importance for all classes
    - Define metrics based on specific classes : F1, Precision, recall, Accuracy on individual classes
    - AUC- ROC for 0-1 regression problems
  - Data-level methods - Resampling
    - Under sampling - Totem links
    - Oversampling - SMOTE
  - Algorithm level methods
    - Cost sensitive learnings
      Drawback: Manually choosing cost
    - Class balanced loss
    - Focal loss

Data Augmentation

- Add more data to improve model performance
- Types
  - Simple Label Preserving Transformations
    - By manipulation
  - Perturbations
    - Adding noise to training data so that model generalizes well against adversarial attacks
  - Data Synthesis