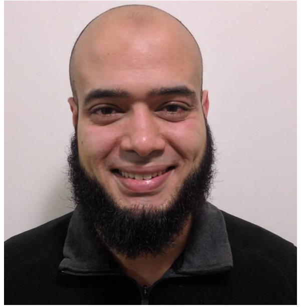
Machine Learning

Sampling Techniques

Mostafa S. Ibrahim
Teaching, Training and Coaching for more than a decade!

Artificial Intelligence & Computer Vision Researcher
PhD from Simon Fraser University - Canada
Bachelor / MSc from Cairo University - Egypt
Ex-(Software Engineer / ICPC World Finalist)



Sampling

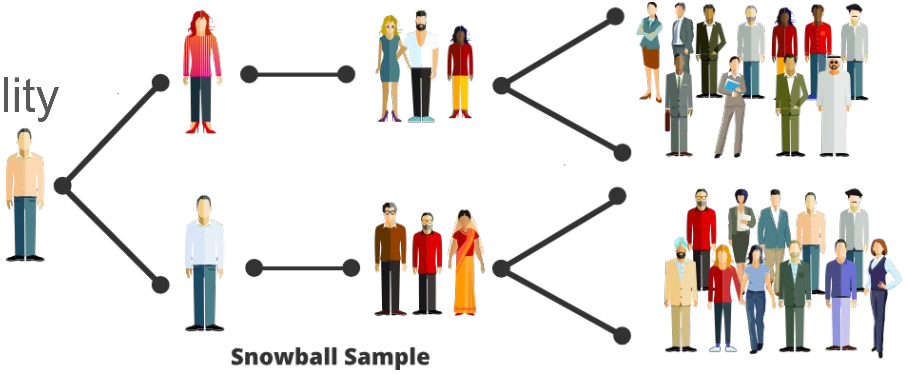
- Sampling techniques in machine learning are strategies used to select a **subset** of data from a larger dataset, typically for the purpose of **training or evaluation.**
- Different sampling techniques serve various **objectives**, such as balancing class distribution (oversampling / undersampling), reducing computational time, or improving model performance
- Sampling techniques may create a **bias** (such as examples selection in active Learning).
  - One has to carefully understand what kind of bias I am introducing and its effect

Convenience Sampling

- Convenience sampling is a **non-probability** sampling technique where subjects are **selected** because of their convenient accessibility.
- This is the most common case for building early iterations of your dataset for POC
    - Easy and cheap to build, but clearly biased and can’t generalize
    - For example, you collected data for in-capin automotive in the summer using the workers in nearby factory who are mainly black males

Snowball Sampling

- Non-probability sampling: Start from a few samples and from them identify more samples and so on
- For example, start from a few twitter accounts and use their connections to find more
  - Or Start from a page of links and crawl more pages
  - You always limited to the possible scope of early sample
  - Good for Social Network Analysis
  - Again biased
- Many datasets for follows non-probability sampling, such as crawling wiki or reviews (amazon/IMDM)



Img src

Random Sampling

- **Simple** Random Sampling
  - Each instance has an equal chance of being selected.
  - Very common
- **Stratified** Random Sampling
  - Preserves the same class distribution as in the original dataset.
  - Typically in imbalance datasets
- **Weighted** sampling (with/out Replacement): assign weight for samples
  - **import** random
  - items = [**'apple'**, **'banana'**, **'cherry'**, **'date'**]
  - weights = [0.5, 0.2, 0.2, 0.1]
  - sampled_items = random.choices(items, weights, k=5)

Reservoir Sampling

- Imagine a stream of data coming to production (e.g. video frames from self-driving)
- We want to fine-tune our model on randomly selected K frames, but we don't know the total size (N)
- RS algorithms can pick random subsets from a stream of data
    - **Interview** Question
    - Given a stream of numbers, sample 1 number uniformly
        - Harder version: sample k numbers uniformly

Importance Sampling

- **Importance** sampling is a statistical technique used primarily to estimate properties of a particular distribution, while only having samples generated from a **different distribution** rather than the distribution of interest
- Applications in Machine Learning
  - Rare Event Simulation: Sample the 'important' parts of a space where rare events occur.
  - Monte Carlo Methods: improve the efficiency of Monte Carlo simulations.
  - Reinforcement Learning: In scenarios like off-policy learning
  - Variational Inference: approximate intractable integrals in Bayesian inference
- Study in the future when you do deeper with probabilistic modeling

"Acquire knowledge and impart it to the people."

"Seek knowledge from the Cradle to the Grave."