[03-05-2025], Lecture Note: Representation Bais

[Niket Pathak, Purva Tandel]

CS 516: Responsible Data Science and Algorithmic Fairness; Spring 2025
Abolfazl Asudeh; www.cs.uic.edu/∼ *asudeh/teaching/archive/cs516spring25/*

1 Introduction to Representation Bias

1.1 Definition

Representation bias occurs when certain groups or regions in a dataset are underrepresented, leading to biased outcomes in data-driven decision-making systems. This bias can manifest in various forms, such as demographic, geographic, or temporal underrepresentation.

1.2 Importance

Addressing representation bias is crucial for ensuring fairness, equity, and accuracy in AI and machine learning models. Biased models can lead to unfair treatment of underrepresented groups, especially in critical applications like healthcare, criminal justice, and recruitment.

1.3 Scope

The focus is on identifying and resolving representation bias in both structured (tabular) and unstructured (image, text, graph) data. The goal is to ensure that datasets are balanced and representative of the entire population they aim to model.

2 Measuring Representation Bias

2.1 Representation Rate

The representation rate measures the proportion of data points belonging to different demographic or categorical groups. It ensures equitable data distribution across different segments.

$$R = \frac{|G_1|}{|G_2|} \tag{1}$$

where:

- $|G_1|$: Number of samples from group G_1
- $|G_2|$: Number of samples from group G_2
- $R \approx 1$: Balanced dataset
- $R \gg 1$ or $R \ll 1$: Imbalanced dataset

2.2 Data Coverage

Data coverage measures how well a dataset represents the full range of conditions or groups within a feature space. Insufficient data coverage can lead to representation bias, where certain groups or characteristics are underrepresented or missing, causing the model to favor well-represented areas. By assessing data coverage, sparse or missing data can be identified, allowing for targeted data collection or augmentation to ensure fairness and better generalization. This helps reduce bias and improves model performance across diverse real-world scenarios.

2.3 Query Condition

A query point q is considered covered if:

$$\{x \in D \mid \text{similarity}(x, q) \leq R\} \geq K \tag{2}$$

where:

- R: Similarity radius.
- K: Minimum threshold of similar points required for coverage.

3 Identifying Uncovered Regions

3.1 Problem Definition

An uncovered query point refers to a data point that lacks a sufficient number of neighboring data points in its vicinity. Identifying such regions is crucial in various domains, including fairness-aware machine learning, where data sparsity can lead to biased model predictions. It also plays a significant role in anomaly detection, as outliers often exist in low-density regions of the data space. Additionally, detecting these uncovered areas helps in dataset bias correction by ensuring that underrepresented regions receive adequate coverage for more balanced and accurate modeling.

3.2 Voronoi Diagrams

A Voronoi diagram is a spatial partitioning structure that divides space into regions based on the nearest data point. Each region, known as a Voronoi cell, contains all locations that are closer to its respective center point than to any other. This structure is widely used in computational geometry, clustering, and spatial analysis to understand proximity relationships in a dataset. Voronoi diagrams help in identifying decision boundaries, optimizing resource allocation, and analyzing spatial distributions in various real-world applications.

- Construction complexity: $O(n \log n)$.
- Querying complexity: $O(\log n)$.

3.3 k-Voronoi Diagrams

An extension of Voronoi diagrams partitions space based on the k-nearest neighbors rather than a single closest point, providing a more flexible representation of spatial relationships. Instead of assigning each region to just one data point, this approach considers multiple nearby points, allowing for smoother and more adaptive boundaries. This method is particularly useful in identifying uncovered areas, as it accounts for local density variations and reduces sensitivity to noise. By incorporating multiple neighbors, it enhances accuracy in applications like clustering, anomaly detection, and spatial modeling.

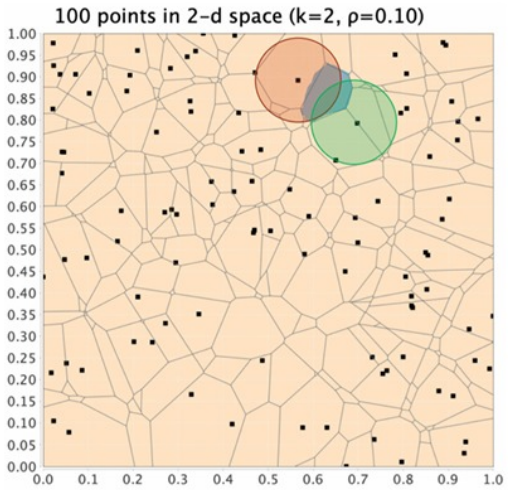
3.3.1 Algorithm Steps

1. Construct the k-Voronoi diagram for the dataset.
2. Iterate through Voronoi cells to assess coverage based on k-nearest neighbors.
3. Merge uncovered regions to define overall gaps in the dataset.

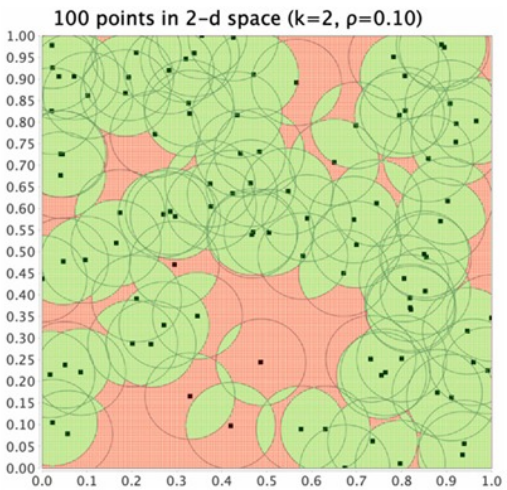
4 Case Study: Breast Cancer Detection

4.1 Real-World Example

In Chicago, breast cancer disproportionately affects Black women, with higher mortality rates compared to their white counterparts. This disparity is driven by multiple factors, including socioeconomic barriers, limited access to quality healthcare, and delays in diagnosis and treatment.



Identifying the covered region in the gray Voronoi cell.



Covered region in dataset $\mathcal{D}$ marked in green. The covered region is the union of all the covered points in each Voronoi cell.

Despite the prevalence of breast cancer in Black women, early detection models are often trained on predominantly white patient data, leading to biases in risk assessment and screening recommendations. These models may fail to accurately identify high-risk cases in Black women, further exacerbating healthcare inequities. Addressing these biases requires more inclusive datasets, equitable healthcare policies, and targeted outreach to ensure early detection and improved outcomes for all racial groups.

4.2 Bias in Training Data

- **White women**: Greater access to routine mammography screenings, leading to early-stage detection data in datasets.
- **Black women**: Due to socioeconomic and systemic factors, undergo screenings less frequently, resulting in late-stage detection data.

4.3 Consequences of Biased Data

Models trained on imbalanced datasets often struggle to generalize across diverse demographic groups, leading to inaccurate predictions and biased outcomes. When a dataset is dominated by one group—such as white patients in medical imaging or diagnostic models—the trained model becomes highly attuned to patterns in that group while performing poorly for underrepresented populations. In healthcare, this imbalance can result in higher misdiagnosis rates for minority groups, such as Black and Hispanic patients, who may have different genetic risk factors, disease progression patterns, or symptom presentations. These biases can lead to delayed diagnoses, inappropriate treatments, or even missed critical conditions, potentially causing severe health consequences. To mitigate these risks, it is essential to incorporate diverse and representative datasets, apply fairness-aware machine learning techniques, and continuously evaluate model performance across different demographic groups.

5 Strategies for Data Curation

5.1 Collect More Data from Underrepresented Groups

- Deploy mobile mammography units to underserved communities to enhance early detection efforts.
- Partner with healthcare providers to increase the collection of diverse data.

5.2 Financial Incentives for Data Collection

Data Markets: Establishing frameworks that allow hospitals and medical institutions to sell anonymized, diverse datasets to research organizations can significantly improve the quality and fairness of machine learning models in healthcare. These frameworks should ensure that data is ethically sourced, securely anonymized to protect patient privacy, and representative of diverse demographic groups, including different races, ethnicities, and socioeconomic backgrounds. By enabling the sale or controlled sharing of such datasets, researchers and AI developers can train more balanced models that generalize better across populations, reducing biases in diagnosis and treatment recommendations. Additionally, these frameworks should include regulatory oversight to prevent misuse, promote transparency, and incentivize hospitals to participate without compromising patient confidentiality. Implementing such initiatives can enhance the accuracy of predictive models, improve healthcare equity, and drive innovations in personalized medicine.

5.3 Synthetic Data Generation

Data Augmentation: Utilizing Generative Adversarial Networks (GANs) and other machine learning techniques to generate synthetic data can help balance datasets and improve model fairness. GANs consist of two neural networks—a generator and a discriminator—that work together to create realistic synthetic data points that mimic real-world patterns. In healthcare, GANs can be trained on limited datasets to produce high-quality synthetic patient records, medical images, or diagnostic data that preserve the statistical properties of real data while ensuring privacy. This approach is particularly useful for addressing data imbalances in underrepresented groups, such as minority populations in medical studies, where collecting sufficient real-world data may be challenging due to privacy concerns, low participation rates, or historical biases. By incorporating synthetic data into training sets, machine learning models can improve their generalization ability, reducing disparities in predictions and enhancing diagnostic accuracy across diverse demographic groups. However, careful validation is required to ensure that synthetic data faithfully represents real-world variations and does not introduce unintended biases.

6 Optimized Data Collection Using Computational Techniques

6.1 Problem Setup

- **Data Sources**: Multiple institutions provide demographically diverse datasets.
- **Objective**: Minimize data collection costs while ensuring the dataset meets demographic representation requirements.
- **Unknown Distributions**: Each data source has different probabilities of providing specific demographic samples, which must be inferred through active querying.

6.2 Algorithmic Approaches

6.2.1 Dynamic Programming Approach

$$f(c_1, c_2) = \min_{D_i} \left(\sum P_{ij} f(c_1 - 1, c_2) + \sigma_i \right) \tag{3}$$

where:

- P_{ij}: Probability of getting a sample from group j from data source D_i.
- σ_i: Cost of querying data source D_i.

Time Complexity: $O(n^k)$.

6.2.2 Reinforcement Learning Approach (Multi-Armed Bandit)

- **Upper Confidence Bound (UCB)**: Prioritizes sources with high potential but uncertain returns.
- **Thompson Sampling**: Uses Bayesian inference to dynamically balance exploration and exploitation.

7 Key Takeaways and Applications

- **Voronoi Diagrams:** Efficiently partition space and identify uncovered data regions.
- **Dynamic Programming:** Provides optimal solutions for data collection but can be computationally expensive.
- **Greedy Heuristics:** Offer practical, approximate solutions in simplified settings.
- **Multi-Armed Bandit Algorithms:** Aid in decision-making when data distributions are unknown.
- **Applications:** These computational techniques are widely used in machine learning, data science, and computational geometry.

8 Conclusion

- Representation bias is a critical issue in data-driven decision-making, and addressing it requires a combination of theoretical and practical approaches.
- Techniques like Voronoi diagrams, dynamic programming, and multi-armed bandit algorithms provide powerful tools for identifying and mitigating representation bias.
- Future work should focus on extending these techniques to more complex datasets and real-world applications, ensuring fairness and equity in AI systems.

Reference

Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput. Surv.* 55, 13s, Article 293 (July 2023), 39 pages. https://doi.org/10.1145/3588433