

CSC 301: Analysis of Algorithms

Lecture 23: Bias in Algorithms

William Rebelsky
[rebelskyw]

Grinnell College

Wednesday 13th March, 2024

Reminders

- Homework 5 is due tonight
- Midterm is due tonight
- The Essay has been released, it is due after break

Table of Contents

1 Objectives

Objective Function

- In most modeling approaches, the goal is to minimize the error between the predicted value and the actual value of the objective
- For example, when fitting a line to a set of data, we generally minimize the mean squared error (ℓ_2 norm) of the predicted value $\hat{y}$ vs the actual value y
- The choice of error function is important (e.g. we could just as easily minimize the absolute error (ℓ_1 norm) or the worst case error (ℓ_∞ norm))
- However, today we will talk about the difference between the goal of the designers, and the objective in the data

Matching Objectives vs Objective Mismatch

Given a dataset of dogs and cats, can you build a model that when given an image:

- Tells you whether it is a dog or a cat? In this case, the objective (dog or cat) is captured in the data, so the overall objective is solvable with an objective based on the data.
- Tells you whether it is a rat or not a rat? In this case the objective is not captured in the data, so there is a mismatch.
- Tells you whether or not the image is a dog? In this case, the objective appears to be in the data (since we have images of dogs), but there may not be enough information in the cat images to properly identify a rat as not a dog.

Objective Mismatch

Objective mismatch: fit a model to an objective that is different than the one you wish to optimize. The less correlated the objectives are the worse that this gets. This is especially evident in models where we can't actually measure the objective we want to optimize.

- Policing models (where should we send more police):
 - Goal: Send more police to places that are more likely to have crimes occur
 - Data that can be found: Places that are more likely to have someone arrested
 - Metric: Arrest counts
- Healthcare (how should finite resources be allocated):
 - Goal: Minimize cost of hospital stays
 - Data: Cost of hospital stays by person
 - Metric: Minimize length of stay since each additional day is costly

Objective Mismatch

- Healthcare (how should finite resources be allocated):
 - Goal: determine who is most likely to need medical intervention, apply appropriate interventions in a cost efficient manner
 - Data: People who have received interventions
 - Metric: Minimize dollars spent while maximizing care
- Generative AI for healthcare:
 - Goal: Use generative AI to summarize and provide up to date information during the pandemic
 - Data: Internet
 - Result: Misinformation
- Amazon hiring. Goal: Hire people who will do well for Amazon. Data: People who have done well in the past. Result: Continuation of biased practices

Choice of Objective

- Goodhart's Law: When a measure becomes a target, it ceases to be a good measure. Or (Jon Danielsson) Any statistical relationship will break down when used for policy purposes.
- People respond to metrics by
 - improving the system
 - distorting the system (e.g.)
 - distorting the data (cheating on exams, teaching to the test)
- Examples of distortions:
 - Cobra bounty: If there is a bounty on cobras, raise them and sell corpses
 - Exams: If school funding is based on test scores, teach to the test
 - 5/10 point scales for service (e.g. Uber): Companies want high ratings → anything lower than a 5 is failing

Choice of Objective

- Paperclip Maximizer - Nick Bostrom
- Survivorship bias: airplane example
- Recent lawsuits:
 - USDA crop insurance: Loss of crops (goal) vs lack of rain (measure). Distortion: covering buckets.
 - ADM: Maximize profit (goal) vs maximize profit of new division (measure). Distortion: moving money to new division by buying from other divisions at a loss.

Today's Discussion

- Readings:
 - Racial Bias in Health Care
 - Data Before Algorithms
- Objective Mismatch: Are there other examples of objective mismatch in the readings this week? In popular news? From other sources you know about?
- How would you more appropriately model the objective for some of today's examples?