15.095 Machine Learning Under A Modern Optimization Lens:

Writing Quality Prediction using Writing Processes

Authors:
Lucas Goh
Kevin Sheng

Table of Contents

1 Purpose and Scope

The landscape of online shopping has been transformed significantly by the internet and AI. Customer expectations for purchase delivery have shortened from weeks to days, and even to just a few hours. The emergence of delivery drones introduces an entirely new set of challenges that demand data science solutions. In this project, we aim to develop an efficient and optimized solution for a simulation of managing and scheduling a fleet of delivery drones to fulfill customer orders as quickly as possible.

2 Data

Our data is sourced from a competition on Kaggle organized by Google. The data is provided as a plain text file containing exclusively ASCII characters with lines terminated with a single '\n' character at the end of each line (UNIXstyle line endings).

2.1 Data Description

In the dataset, we are provided with four categories of information:

1) Parameters of the simulation, which includes number of drones available, a drone's capacity, map's grid size, etc
2) *p* products. which includes product types and product weights
3) *w* warehouses, which includes locations of warehouses and availability of product types at each warehouse
4) *n* orders, which includes the delivery location [a, b] and the quantity of products in each order

2.2 Cost Variable

The simulation takes place in a two-dimensional grid. Each cell is identified by a pair of integer coordinates [a,b]. In the simulation, we define an cost variable measured as time t_i, which equals the time when the last order *i* is completed. All drones take one unit of t_i to travel one unit of Euclidean distance, and many drones can travel at the same time. Additionally, we also assume each drone takes an additional unit of t_i to load and unload all the products.

3 Optimization Objective

Our objective function is defined as:

$$min \sum_{i=1}^{n} \frac{t_i}{T},$$

where *T* is the total time used to complete all the orders.

This objective prioritize two things. First, it minimizes the completion times of individual orders. More importantly, it encourages a reduction in the variability in completion times. Since each order's completion time is divided by the total time T, the objective function implicitly pushes towards a balance, where no single order's completion time is disproportionately large.

4 Exploratory Data Analysis

5 Baseline Model

Our baseline model is a naive approach where each order is completed at one time. The objective value can be easily calculated for this approach with the following steps:

1. Assign each order to nearest warehouses (see 4.1)
2. Define d_{ij} as the distance between order i and warehouse j.
3. Define wa_i as the number of warehouses assigned to order i.
4. Define t_i as $t_i := t_{i-1} + 2 + max_j(d_{ij})\ for\ i = 1\ to\ n,\ and\ t_0 = 0$.
5. Define $T\ as\ max_i(t_i)$.
6. Now, we can compute $\frac{t_i}{T}$ for all orders i from 1 to n.

We note that this approach assumes that there are sufficient drones that can return to the warehouses for each order and they be readied before the next order starts. The baseline objective value using this approach is 626.22.

4 Solution Approach

As there is not a straighforward formulation to solve this problem, our approach is a 3-step optimization approach that aims to minimize the objective value:

1. Optimally Assign Orders to Nearest Warehouses
2. Capacitated Vehicle Routing Problem
3. Drone Assignment Problem

This project addresses a formidable large-scale logistics optimization problem, characterized by its complexity due to the interplay between numerous warehouses, orders, product types, and drone operations. The intricacies arise from the multi-dimensional nature of the problem, where multiple warehouses serve an extensive array of orders, each comprising a variety of product types fulfilled at different times.

The original formulation of this problem was a monolithic model that aimed to capture all variables and constraints simultaneously. However, given the size and complexity, this led to a non-linear and non-convex problem space, making it computationally intractable. The sheer number of decision variables, coupled with the diversity of constraints, presented a significant challenge for both formulating and solving the problem using conventional optimization methods.

To navigate this complexity, we adopted a problem decomposition approach. Inspired by methodologies used in the bus station assignment and vehicle routing problems taught in class, we shifted our focus to a three-step modular problem. This decomposition simplifies the original problem by breaking it down into more manageable sub-problems, each addressing different aspects of the logistics chain:

4.1 Assignment of Orders to Warehouses

The initial phase of our approach involves solving an assignment problem: specifically, allocating orders to warehouses. Our exploratory data analysis (EDA) indicates that this task is feasible. We reached this conclusion by examining the aggregate supply of each product across all warehouses and comparing it with the corresponding total demand from all orders. Our findings showed that for each product category, the aggregate supply exceeds the aggregate demand.

Ideally, each order would be fulfilled by the nearest warehouse to minimize logistics complexities. However, our EDA revealed that this is not always possible, due to disparities in product availability at individual warehouses. Consequently, we must sometimes allocate parts of an order to multiple warehouses to ensure complete fulfillment.

To address this, we've introduced a decision variable at the product level, denoted as x_{ijk}, representing the quantity of product i dispatched from warehouse j to fulfill part of order k.

4.2 Capacitated Vehicle Routing Problem

In this section, we analyze each warehouse independently. Take, for instance, Warehouse 1: we evaluate it along with all the orders allocated to it in the initial step. This scenario is treated as a distinct capacitated vehicle routing problem (CVRP). This process is then replicated for every other warehouse.

This particular CVRP differs from the formulations discussed in our classes. It incorporates unique elements such as demand constraints, capacity limitations, and the possibility of multiple rounds. Notably, drones are capable of executing the same route over several rounds.

Below is the detailed formulation of the problem:

Decision variable: x_{ijk}

4.3 Drone Assignment Problem

We will assess the second step by determining which warehouse completes all orders first. Subsequently, a drone from this efficient warehouse will be reallocated to the warehouse that takes the longest time. After this adjustment, we will re-evaluate the objective function to check for any improvement. This process will be iterated until the objective value stabilizes. The final outcome of this procedure will be the definitive assignment of drones.

5 Results

5.1 Model Results

5.2 Final Model

5.3 Model Insights

8 Limitations & Next step

9 Conclusion

10 Appendix

Figure 1 and 2: AUC and Accuracy for models predicting FYC Performance

Figure 3 and 4: AUC and Accuracy for models predicting BP with FYC features