Project Title

Jiafeng Chen Yufeng Ling Francisco Rivera

Abstract

- This document describes the expected style, structure, and rough proportions for your final project write-up.
- While you are free to break from this structure, consider it a strong prior for our expectations of the final report.
- Length is a hard constraint. You are only allowed max **8 pages** in this format. While you can include supplementary material, it will not be factored into the grading process. It is your responsibility to convey the main contributions of the work in the length given.

1. Introduction

Example Structure:

- What is the problem of interest and what (high-level) are the current best methods for solving it?
- How do you plan to improve/understand/modify this or related methods?
- Preview your research process, list the contributions you made, and summarize your experimental findings.

2. Background

Example Structure:

- What information does a non-expert need to know about the problem domain?
- What data exists for this problem?
- What are the challenges/opportunities inherent to the data? (High dimensional, sparse, missing data, noise, structure, discrete/continuous, etc?)

3. Related Work

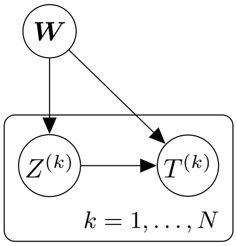
Example Structure:

- What 3-5 papers have been published in this space?
- How do these differ from your approach?
- What data or methodologies do each of these works use?
- How do you plan to compare to these methods?

4. Model

We represent a city's road network with a connected graph $G = (V, E)$. Assume that each vertex $i \in V$ is associated with a weight w_i, representing the cost of traversing vertex i. A trip is represented by a path in G, and the distribution of the trip's duration depends on the weights w_i of vertices included in the path. Note that the choice of the path can in general depend on the collection of weights $\boldsymbol{W}$. In full generality, the model is represented by Figure 1, where trips in the data are indexed by (k), $T^{(k)}$ is the observed trip duration, and $Z^{(k)}$ is the path taken by trip k, a latent variable.

Figure 1. Representation of model as a directed graph

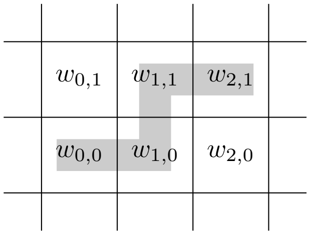


4.1. Parameterization

In principle, in our application to the New York City taxi data, we may take G to be a graph representing the exact road network in New York City, where each vertex is a *road segment* and the directed edge $(i, j) \in E$ if one can drive directly onto road segment j from road segment i; such a parameterization allows w_i to be directly interpretable as a measure of congestion on road segment i. However, such a detailed construction presents serious computational challenges when training on a large dataset, since solving pathfinding problems and computing minimal paths are non-trivially expensive.[1]

To avoid these challenges, we parameterize G as an undirected rectangular grid. Despite not being able to pinpoint weights w_i to congestion of specific road segments, we are nonetheless able to interpret the weights w_i as representative of congestion on a small patch of land. We may now represent a path $Z^{(k)}$ as a set of indices i of grid points traversed by the path. In full generality, there are an infinite number of paths connecting any two points i, j on the grid, but the vast majority of these paths are not sensible. Thus we restrict the set of possible paths for trip k to a *set of reasonable paths* $\boldsymbol{Z}^{(k)}$, where each path in $\boldsymbol{Z}^{(k)}$ travels strictly in the direction of the destination. For instance, if the destination of j is to the northeast of the starting location i, then the set of reasonable paths $\boldsymbol{Z}$ are the set of paths that only involve northward or eastward movements (e.g. Figure 2).

Figure 2. An example of a reasonable path



We have the following model:

$$\boldsymbol{W} \sim p(\boldsymbol{W})$$
$$Z^{(k)} \sim p(Z^{(k)}|\boldsymbol{W})$$
$$T^{(k)}|\boldsymbol{W}, Z^{(k)} \sim \mathcal{N}\left(\sum_{i \in Z^{(k)}} w_i, \sigma^2\right),$$

Example Structure:

- What is the formal definition of your problem?
- What is the precise mathematical model you are using to represent it? In almost all cases this will use the probabilistic language from class, e.g.

  $$z \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

  But it may also be a neural network, or a non-probabilistic loss,

  $$h_t \leftarrow \mathrm{RNN}(x_t, h_{t-1})$$

  This is also a good place to reference a diagram such as Figure **??**.
- What are the parameters or latent variables of this model that you plan on estimating or inferring? Be explicit. How many are there? Which are you assuming are given? How do these relate to the original problem description?

5. Inference (or Training)

- How do you plan on training your parameters / inferring the states of your latent variables (MLE / MAP / Backprop / VI / EM / BP / ...)
- What are the assumptions implicit in this technique? Is it an approximation or exact? If it is an approximation what bound does it optimize?
- What is the explicit method / algorithm that you derive for learning these parameters?

Algorithm 1 Your Pseudocode

6. Methods

- What are the exact details of the dataset that you used? (Number of data points / standard or non-standard / synthetic or real / exact form of the data)
- What are the exact details of the features you computed?
- How did you train or run inference? (Optimization method / hyperparameter settings / amount of time ran / what did you implement versus borrow / how were baselines computed).
- What are the exact details of the metric used?

7. Results

- What were the results comparing previous work / baseline systems / your systems on the main task?
- What were the secondary results comparing the variants of your system?
- This section should be fact based and relatively dry. What happened, what was significant?

8. Discussion

- What conclusions can you draw from the results section?

[1] Manhattan has on the order of 10^4 road segments, and the dataset contains the order of 10^7 trips for January 2009 alone.

Table 1. This is usually a table. Tables with numbers are generally easier to read than graphs, so prefer when possible.

- Is there further analysis you can do into the results of the system? Here is a good place to include visualizations, graphs, qualitative analysis of your results.
- What questions remain open? What did you think might work, but did not?

9. Conclusion

- What happened?
- What next?