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

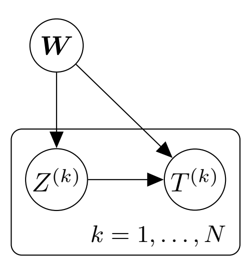
Example Structure:

- 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. Our primary interest is to perform inference on $\boldsymbol{W}$, so as to learn the levels of congestion associated with each vertex in G.

Figure 1. Representation of model as a directed graph

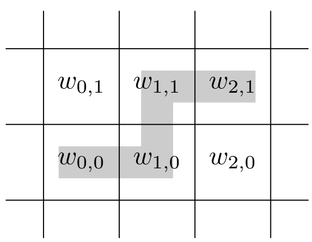


4.1. Parameterization

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 path-finding 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 shows a reasonable path from $(0, 0)$ to $(2, 1)$). Such a parameterization is more general than many in the literature; Zhan et al. (2013), for instance, uses the K-shortest path algorithm and considers the shortest 20 paths as a set of reasonable paths.

Figure 2. An example of a reasonable path



We parameterize the conditional distribution of $T^{(k)}$ as Normal, in the following reformulation of the directed graphical model:

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

where $p(Z^{(k)}|\boldsymbol{W})$ is a distribution over $\boldsymbol{Z}^{(k)}$. We consider two different ways to parameterize $p(Z^{(k)}|\boldsymbol{W})$: softmax regression and uniform. In the *softmax regression* model, a type of generalized linear model for discrete choice problems (McFadden et al., 1973), we parameterize the route choice such that

$$p(Z^{(k)}|\boldsymbol{W}) \propto \exp\left(- \sum_{i \in Z^{(k)}} w_i \right),$$

in order to encode the fact that drivers avoid routes that take a long period of time. In the *uniform* model, we simply assume that route choice is independent and uniform on the set of reasonable paths:

$$p(Z^{(k)}|\boldsymbol{W}) \propto 1.$$

The uniform model trades off realism in modeling for improvement in computation and training, as we see in Section 5.

In our application to the Manhattan dataset, we perform MLE inference, or, equivalently, MAP inference with $p(\boldsymbol{W}) \propto 1$. In principle, it is not difficult to parameterize the prior of $\boldsymbol{W}$ as an undirected graphical model, since we need only to supply edge and unary potentials. For instance, to impose a correlated prior on $\boldsymbol{W}$, as suggested by some (Hunter et al., 2009), we simply penalize large differences in neighboring weights in the edge potential, effectively assuming a prior model that is similar to a continuous version of the Ising model.

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.

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

  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)

We perform maximum likelihood inference, maximizing

$$\max_{\boldsymbol{W}} \log p(\{T^{(k)}\}_{k=1}^{N} | \boldsymbol{W}) = \max_{\boldsymbol{W}} \sum_{k=1}^{N} \log p(T^{(k)} | \boldsymbol{W}).$$

[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.

The log-likelihood is

$$\begin{aligned}&\log p(T^{(k)}|\boldsymbol{W})\\&=\log\left(\sum_{Z^{(k)}\in\boldsymbol{Z}^{(k)}}p(T^{(k)}|Z^{(k)},\boldsymbol{W})p(Z^{(k)}|\boldsymbol{W})\right)\\&=\log\left(\mathbb{E}_{Z^{(k)}}\left[p(T^{(k)}|Z^{(k)},\boldsymbol{W})\right]\right).\end{aligned}$$

The expectation is a sum of the size $|\boldsymbol{Z}^{(k)}|$, which, for an $m\times n$ trip[2], is of $\binom{m+n}{n}\approx\frac{(n+m)^n}{n^n}e^n$ terms. Computing this expectation is the main inference challenge of our project.

5.1. Inference in the uniform model

By assuming the uniform model $p(Z|\boldsymbol{W})\propto 1$, we gain the ability to work with an expectation over $\boldsymbol{W}$ instead of an expectation over $\boldsymbol{Z}$, since the probability that a particular weight is included in a path is readily computable from elementary combinatorics. We maximize an approximate lower bound of the log-likelihood by applications of Jensen's inequality:

$$\begin{aligned}\ell(\boldsymbol{W})&=\log\left(\mathbb{E}_{Z^{(k)}}\left[p(T^{(k)}|Z^{(k)},\boldsymbol{W})\right]\right)\\&=\log\left(\mathbb{E}_{Z^{(k)}}\left[\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}(T^{(k)}-\sum w_i)^2}\right]\right)\end{aligned}$$

Note that the Normal density is concave for $|T^{(k)}-\sum w_i|<\sigma$ and convex otherwise. Applying Jensen's inequality locally, we obtain that

$$B(\boldsymbol{W})=\frac{1}{2\sigma^2}\left(T^{(k)}-\sum_i w_i\pi_i\right)^2,$$

where π_i is the marginal probability of node i being included in a uniform route,[3] is a local lower bound for the negative log-likelihood for trips that we predict poorly and is a local upper bound for trips that we predict well, up to a constant. Thus $B(\boldsymbol{W})$ is a good approximation of the objective function that is easily computable, involving only mn as opposed to $\binom{m+n}{n}$ terms.

5.2. Inference in the softmax regression model

We now consider a more realistic but more complex model, where we parameterize $Z|\boldsymbol{W}$ as a GLM, namely as a softmax regression where $p(Z|\boldsymbol{W})\propto\exp\left(-\sum_{i\in Z}w_i\right)$, encoding drivers' preferences for shorter trips. The log-likelihood in this model is

$$\begin{aligned}&\ell(\boldsymbol{W})\\&=\log\left(\mathbb{E}_{Z^{(k)}}\left[\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(T^{(k)}-\sum_i w_i)^2}\right]\right)\\&=\log\left(\sum_{Z^{(k)}\in\boldsymbol{Z}^{(k)}}\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(T^{(k)}-\sum_i w_i)^2}p(Z^{(k)}|\boldsymbol{W})\right)\end{aligned}$$

Before we discuss our inference strategy, we first discuss a few difficulties of the model. At first glance, the latent variable structure seems lend well to the Expectation-Maximization algorithm (Dempster et al., 1977). However, the EM algorithm requires computing the term

$$Q(\boldsymbol{W}|\boldsymbol{W}^{(t)})=\mathbb{E}\left(\log(T^{(k)},Z^{(k)}|\boldsymbol{W})|T^{(k)},\boldsymbol{W}^{(t)}\right),$$

and computing the expectation requires the conditional distribution $p(Z|T,\boldsymbol{W})=\frac{p(T|Z,\boldsymbol{W})p(Z|\boldsymbol{W})}{p(T|\boldsymbol{W})}$, where the denominator is intractable to compute. We might also consider techniques in variational inference (Blei et al., 2017). However, the latent variable $Z^{(k)}$ is a random set of indices with the property that the indices form a path on the grid, and thus cannot be reasonably decomposed into independent components, ruling out the mean-field algorithm. Another promising option is stochastic gradient variational Bayes (SGVB) and variational autoencoders (Kingma & Welling, 2013), which is designed for large datasets for which the EM algorithm fails. However, SGVB requires a reparameterization of the latent variable drawn from an approximate distribution, in order for the gradients to be properly computed. In SGVB, one independently draws some ϵ from a distribution and obtains z via $z=g(\epsilon)$ for some continuous function g.[4] This is difficult to do in our context, since Z is not continuous nor scalar-valued. Thus it is difficult to find an ϵ and a continuous transformation to approximate samples from the distribution of Z.

Our inference strategy is based on a sampling-based approximating to the expectation over $\boldsymbol{Z}$. The key trick we use is that

$$\mathbb{E}_Z\left[f(Z^{(k)})\right]=|\boldsymbol{Z}^{(k)}|\mathbb{E}_{\tilde{Z}}\left[f(\tilde{Z}^{(k)})p_Z(\tilde{Z}^{(k)}|\boldsymbol{W})\right],\quad(2)$$

where $\tilde{Z}$ is drawn uniformly from $\boldsymbol{Z}$. This is the same technique as in importance sampling, except here the objective is to derive computationally tractable approximations, rather than maximizing the efficiency of the approximations. Exchanging log and expectation operator and applying (2) yields the following lower bound for negative

[2] By an $m\times n$ trip, we mean a trip with east-west distance n and north-south distance m

[3] π_i can be computed analytically. Suppose the source and destination of the trip are (n,m) apart and vertex i is (a,b) away from the source. Then, by elementary combinatorics,

$$\pi_i=\frac{\binom{a+b}{a}\binom{n+m-a-b}{n-a}}{\binom{n+m}{n}}$$

[4] Here we are using the notation in (Kingma & Welling, 2013).

log-likelihood, up to a constant,

$$\frac{1}{2L\sigma^2}\sum_{j=1}^{L}\left[T^{(k)}+\sum_{i\in\tilde{Z}_j^{(k)}} w_i\right]^2-\frac{1}{L}\sum_{i=1}^{L}\sum_{i\in\tilde{Z}_j^{(k)}} w_i$$
$$+\operatorname*{logsumexp}_{i=1,\dots,L}\left(-\sum_{\tilde{Z}_i} w_i\right),$$

where we sample uniformly and independently $\{\tilde{Z}_j^{(k)}\}_{j=1}^{L}$.

- 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?
- 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?

References

Blei, David M, Kucukelbir, Alp, and McAuliffe, Jon D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted), 2017.

Dempster, Arthur P, Laird, Nan M, and Rubin, Donald B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

Hunter, Timothy, Herring, Ryan, Abbeel, Pieter, and Bayen, Alexandre. Path and travel time inference from gps probe vehicle data. *NIPS Analyzing Networks and Learning with Graphs*, 12(1), 2009.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

McFadden, Daniel et al. Conditional logit analysis of qualitative choice behavior. 1973.

Zhan, Xianyuan, Hasan, Samiul, Ukkusuri, Satish V, and Kamga, Camille. Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, 33: 37–49, 2013.

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