Robust Spatio-Temporal Purchase Prediction via Deep Meta Learning

Huiling Qin[1,2,3], **Songyu Ke**[2,3,4], **Xiaodu Yang**[2,3,5], **Haoran Xu**[1,2,3], **Xianyuan Zhan**[2,3*], **Yu Zheng**[1,2,3*]
[1] School of Computer Science and Technology, Xidian University, Xi'an, China
[2] JD Intelligent Cities Research, Beijing, China
[3] JD iCity, JD Technology, Beijing, China
[4] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[5] Artificial Intelligence Institute, Southwest Jiaotong University, China
{orekinana,ryanxhr}@gmail.com,{songyu-ke,msyuzheng}@outlook.com,zhanxianyuan@jd.com,xiaodu.yang@foxmail.com

Abstract

Purchase prediction is an essential task in both online and offline retail industry, especially during major shopping festivals, when strong promotion boosts consumption dramatically. It is important for merchants to forecast such surge of sales and have better preparation. This is a challenging problem, as the purchase patterns during shopping festivals are significantly different from usual cases and also rare in historical data. Most existing methods fail at this problem due to the extremely scarce data samples as well as the inability to capture the complex macroscopic spatio-temporal dependencies in a city. To address this problem, we propose the Spatio-Temporal Meta-learning Prediction (STMP) model for purchase prediction during shopping festivals. STMP is a meta-learning based spatio-temporal multi-task deep generative model. It adopts a meta-learning framework with few-shot learning capability to capture both spatial and temporal data representations. A generative component then uses the extracted spatio-temporal representation and input data to infer the prediction results. Extensive experiments demonstrate the meta-learning generalization ability of STMP. STMP outperforms baselines in all cases, which shows the effectiveness of our model.

Introduction

Reliable purchase prediction is essential for both the online and offline retail industry to optimize the supply chain, reduce operational costs, and improve revenue. This becomes even vital during major shopping "festivals" (e.g., Black Friday in the US, 11.11 shopping carnivals in China), the sudden burst of sales boosted by big promotions poses a great challenge for retailers, causing problems such as stockouts or even system crash, leading to a bad shopping experience for consumers. However, accurate purchase forecasts on different categories of products can enable better preparation for their inventories and design appropriate promotion strategies for products. More importantly, knowing the spatial distribution of product purchases in different categories also helps to design effective temporary shipping strategies to ensure more efficient dispatching of products among warehouses across different regions during such a special period of time. In this work, we aim to predict the purchase of multiple categories of products in different regions, while improving the prediction accuracy of the burst purchase behavior during major shopping festivals. This is a particularly challenging problem, the difficulties mainly arise from two aspects:

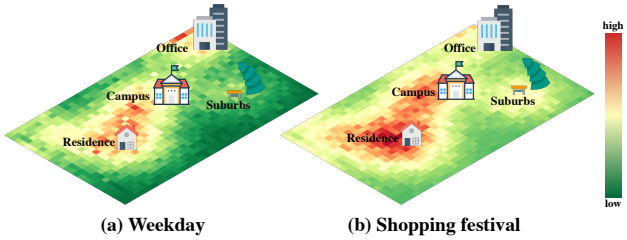

Figure 1: Purchase distribution in different space and time.

Limited reference samples. The spatial distribution of shopping malls and stores, demographic properties of regions as well as day types (e.g. weekdays, weekends, or shopping carnivals) often leads to diverse shopping behavior, as shown in Figure 1. Conventional solution for purchase prediction is to training an exclusive model for each region or day type (Yi et al. 2018). However, the data from a single region or a specific day type are too sparse to support training a good model. The becomes even worse for prediction on shopping festivals, as these festivals only occur few times a year, leading to extremely limited data samples. How to transfer information from different regions and day types to train a model while remaining their own specific characteristics is a challenge.

Complex spatio-temporal purchase patterns. The population density and demographic characteristics are highly heterogeneous in different regions of a city, which could lead to distinct consumption behaviors. Moreover, the economic development, population growth and movement across regions will also shape-changing consumption behaviors in the long-term. Modeling such complex spatio-temporal purchase patterns can be very difficult. To accurately capture the spatial properties of regions, it requires considering a comprehensive set of spatial features, such as point-of-interest (POI) distribution, demographic features of regions, etc. Furthermore, the temporal information gathered from historical data may be biased towards normal purchase patterns, which can not provide sufficient information to support predicting the burst of sales during shopping festivals. How to capture the

*Yu Zheng and Xianyuan Zhan is the corresponding author.

complete spatio-temporal representation of purchase patterns for the modeling process is another challenging task.

In this paper, we propose the Spatio-Temporal Meta-learning Prediction (STMP) model to address aforementioned challenges. The contributions of our work are two-fold:

- **Task-specific spatio-temporal representation learning.** As the purchase pattern is diverse across different regions and day types, the prediction task in each region or day type can be seen as an individual task. We develop a flexible amortization network to learn task-specific spatial-temporal representations. It jointly models the dynamic purchase time-series features and static spatial features to learn a representative spatio-temporal embedding. Moreover, it integrates instances in the same task by an instance pooling operation to capture a meta-representation of the task, which significantly improves the expressive power of the model in the few-shot scenario.
- **Knowledge transfer between tasks.** To tackle the data scarcity issue, we design a generative model that combines the task-specific spatio-temporal representations and the embedding of the current purchase data to enable knowledge transferring among different tasks. This is achieved using shared parameters in the generative model to learn the meta-knowledge from different tasks, which plays a key role in transforming information across different tasks to perform target prediction.

We use a large high-quality online purchase dataset from JD.com for evaluation. Extensive experiments demonstrate the meta-learning generalization ability of our model. In all cases, STMP outperforms the competing baselines, which demonstrate its effectiveness.

Related Work

There is limited literature related to purchasing prediction problem for shopping festivals. For this consideration, we extend the survey to the general time-series prediction problem, especially some works on deep learning and meta-learning based methods used in the probabilistic prediction problems.

Deep learning-based time-series prediction. The only work in literature for purchase prediction in "shopping festivals" is by Zeng et al. (Zeng et al. 2019). It applies collaborative filtering to recommend items for different consumers, and predict whether a purchase will happen. In another purchase prediction related study, Wang et al. (Wang et al. 2019a) propose a hybrid model that incorporates the advantage of both classical time-series model and deep learning method, which can capture complex patterns in the data and capable of solving large-scale problems. In addition, some deep learning-based methods (Lai et al. 2018; Qin et al. 2017; Ma et al. 2017) for time-series prediction can also be used in normal purchase prediction scenarios. But these prediction models either require large amounts of task-relevant data or focus only on capture ordinary time-series patterns, which are incapable to handle the burstiness in a temporal sequence.

Meta-learning generative model Many applications require predictions to be made on myriad small datasets. In such cases, it is natural to desire learners can rapidly adapt to new datasets at test time. These applications have given rise to a vast interest in few-shot learning (Fei-Fei, Fergus, and Perona 2006; Lake et al. 2011), which emphasizes data efficiency via information sharing across related tasks. Since the shopping festival is rare in the purchase data, our task can be framed as a few-shot prediction problem, thus also relevant to studies in the area of meta-learning. Moreover, as uncertainty is rife in few-shot problems, to enhance the robustness of the prediction model, some meta-learning algorithms extend to perform probabilistic inference on prediction problems (Grant et al. 2018). Related approaches include models with amortized Bayesian (Ravi and Beatson 2019), generative meta-learning (Rezende et al. 2016; Reed et al. 2018). While these models can perform probabilistic inference on prediction, their application is limited to notably less challenging tasks. How to extend few-shot learning to complex spatio-temporal prediction scenarios is still a challenging task.

In this work, we utilize the merits of both the above approaches and combine the meta-learning probabilistic inference with spatio-temporal modeling for purchase prediction in a few unusual but significant shopping "festivals".

Problem Statement

As purchase behavior differs greatly across space and time, we consider a prediction problem with I regions $S = \{s_i | i = 1 \cdots I\}$ (divided by the administrative boundaries) and J day types $DT = \{d_j | j = 1 \cdots J\}$ (including weekdays, weekends, and shopping festivals). In each region and day types, our goal is to accurately predict the number of future purchase orders, using historical purchase time-series data, times of products added to the shopping cart, and other external factors as features. Given a time window of length T, the time-series purchase orders are denoted as $Y^r = [y_1^r, \cdots, y_T^r]$, where $y_{it}^r = [y_{t,1}^r, \cdots, y_{t,C}^r]$, $r = (s_i, d_j) \in R$ ($R = S \times DT$) represents a specific region and day type, and C is the total number of product categories. The form of the shopping cart data is the same as the purchase order, which is denoted as $(X_d)^r = [(x_d)_1^r, \cdots, (x_d)_T^r]$ and $(x_d)_t^r = [(x_d)_{t,1}^r, \cdots, (x_d)_{t,C}^r]$. The external factors include region related static features (POIs and demographic profiles) and the festival interval (number of days between current day to the most recent weekend and shopping festival) are denoted as $(X_e)^{s_i}$. For brevity, we denote $X^r = [(X_d)^r, (X_e)]$.

For different regions and day types, we can split the prediction task into multiple sub-tasks, where each sub-task (for a specific region and day type) is perceived as an **SD-task**. Given previous notations, for each SD-task $r = (s_i, d_j)$, the features of the task can be unified as (X^r, Y^r). The training data $D^r = (x_t^r, y_t^r)_{t=1}^{T_n}$ and testing data $\tilde{D}^r = (\tilde{x}_t^r, \tilde{y}_t^r)_{t=1}^{T_m}$ are separated for each SD-task r. The overall purchase prediction problem can be naturally formulate as a multi-task prediction problem, which predict future purchase orders Y_{T+1} for all SD-tasks given $(X_d)_{\leq T}$, (X_e) and $Y_{\leq T} = \{y_1, \cdots, y_T\}$ from past T time steps. In the following section, we use D^r and y^r for each SD-task r as the input data and target.

Methodology

In this section, we introduce the STMP framework for the multi-task purchase prediction problem. STMP includes three

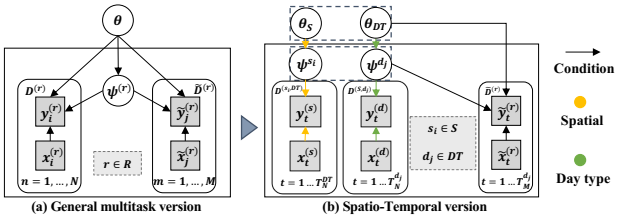

Figure 2: Graphical models for meta-learning framework.

key elements. First, we utilize meta-learning probabilistic inference (Finn, Abbeel, and Levine 2017) in our model to enhance the robustness of few-shot prediction during rare and uncertain shopping festivals. Second, we leverage shared hidden statistical structure between tasks to perform multi-task learning (Heskes 2000). This treatment allows sharing information between tasks about how to learn and perform inference using meta-learning (Thrun and Pratt 2012). Third, we enable fast learning of different SD-tasks via amortized variational inference (Shu et al. 2018), which maps the purchase data instances to a parameterized approximate posterior distribution. In the following, we first introduce the meta-learning probabilistic model for general multi-task learning problem, and further generalize it to the spatio-temporal prediction settings that relevant to our purchase prediction problem.

Meta-learning Probabilistic Model

To improve the robustness and enhance the model performance under limited samples, probabilistic modeling combined with meta-learning has draws lots of attention in solving the multi-task prediction problem (Gordon et al. 2019). Such a model framework is typically constructed upon two ingredients: (i) discriminative models are used to maximize predictive performance on the prediction tasks, which can be modeled as a regression problem; and (ii) the shared statistical structure across tasks in the probabilistic model is exploited to learn the meta-knowledge of different tasks. This model construction can be represented as a multi-task directed graphical model illustrated in Figure 2, in which x^r, y^r are model inputs and outputs for each task r, the shared parameters θ are common to all tasks and $\{\psi^r\}$ are the task-specific parameters. The shared parameter θ plays a key role in transferring meta-knowledge across tasks that greatly alleviate the data scarcity issue. On the other hand, the different patterns within each task are learned and parameterized by task-specific parameters $\{\psi^r\}$, which extract the targeted information for different tasks to further enhance the accuracy. The training D^r and testing $\tilde{D}^r$ data are distinguished for each task, which is a key treatment for few-shot learning.

Let X^r and Y^r be all the inputs and outputs for task r (for both training and testing). The joint probability distribution of the outputs Y^r and task-specific parameters $\{\psi^r\}$ for all tasks given the inputs X^r and shared parameters θ is:

$$p(\{Y^r,\psi^r\}_{r=1}^R|\{X^r\}_{r=1}^R,\theta) = \prod_{r=1}^R p(\psi^r|\theta)\prod_{n=1}^{N_r} p(y_n^r|x_n^r,\psi^r,\theta)\prod_{m=1}^{M_r} p(\tilde{y}_m^r|\tilde{x}_m^r,\psi^r,\theta)$$

where $p(\psi^r|\theta)$ is the probability distribution of task-specific parameter ψ^r given θ, and $p(y^r|x_n^r,\psi^r,\theta)$, $p(\tilde{y}^r|\tilde{x}_m^r,\psi^r,\theta)$ are the probability distributions of the training and testing outputs conditioned on inputs X^r and parameters. N^r, M^r are the number of training and testing instances of task $r \in R$.

In the following, we provide a spatio-temporal generalization for the meta-learning probabilistic model, and the goal is to meta-learn fast and obtain a accurate approximated posterior distribution of purchase orders in different SD-tasks.

Spatio-Temporal Meta-learning Probabilistic Inference

In the spatial-temporal purchase prediction scenario, we exploit commonalities and differences across SD-tasks and perform parameter learning jointly to improve learning efficiency and prediction accuracy for purchase orders in each region and day type, which is illustrated in Figure 2(b).

To learn the meta-knowledge across tasks, we employ point estimates for the shared parameters θ. As data from all SD-tasks provide sufficient information to jointly determine a common high-level statistical structure. Meanwhile, distributional estimates are used for the task-specific parameters $\{\psi^{(s_i)}\}_{i=1}^I$ and $\{\psi^{(d_j)}\}_{j=1}^J$ (correspond to the purchase behavior representation of specific regions and day types), since there might be few instances (shots) during shopping festival or in remote regions with limited data, which could result in a highly uncertain and less constrained statistical pattern. Considering that different regions and day types have different spatial and temporal characteristics, we separately learn the spatial and temporal shared parameters θ_s and θ_d instead of θ, and then integrate them by a specially designed training process (ST-training, will be introduced in later section).

Once the shared parameters θ_s, θ_d are learned, the probabilistic inference to the above multi-task few-shot learning model comprises two steps. First, form the posterior predictive distribution $p(\psi^r|D^r,\theta_{s/d})$ over the task-specific parameters ψ^r given the purchase observations. Second, compute the purchase posterior predictive $p(\tilde{y}^r|\tilde{x}^r,\psi^r,\theta_{s/d})$ based on input features $\tilde{x}^r$, parameters ψ^r, $\theta_{s/d}$. With slight abuse of notations, we write $p(\psi^r|D^r,\theta_{s/d})$ and $p(\tilde{y}^r|\tilde{x}^r,\psi^r,\theta_{s/d})$ as $p(\psi^r|D^r)$ and $p(\tilde{y}^r|\tilde{D}^r,\psi^r)$ for simplicity.

Approximate posterior predictive distribution. As directly model the posterior predictive distributions $p(\psi^r|D^r)$ is intractable, we instead approximate them using purchase order data and amortized variation inference (Bengio and LeCun 2014) implemented by a neural network, denoted as $q_\phi(\psi^r|D^r)$. The use of amortized variational inference and neural networks enable fast predictions at test time. The network of $q_\phi(\psi^r|D^r)$ accepts purchase observations as input, and outputs the mean and variance to form the predictive distribution of the task-specific spatio-temporal representation ψ^r associated with that observation. We can then optimize the parameters of the neural network instead of maintaining $I+J$ different sets of task-specific parameter distributions. The approximate posterior predictive distribution over the test output $\tilde{y}^r$ can thus be evaluated as:

$$q_\phi(\tilde{y}^r|\tilde{D}^r) = \int p(\tilde{y}^r|\tilde{D}^r,\psi^r)q_\phi(\psi^r|D^r)d\psi^r$$

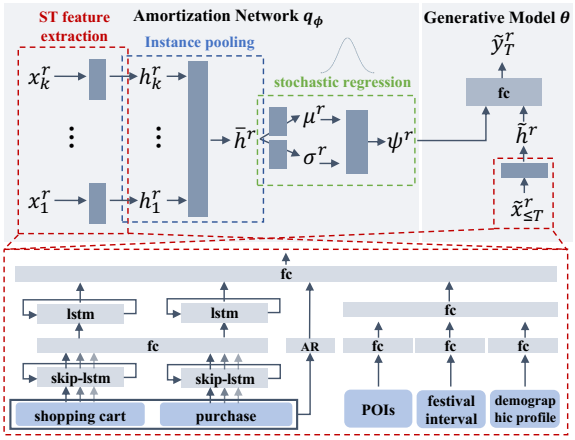

Figure 3: The construction of the proposed STMP model.

Meta-learning the approximate posterior predictive distribution. The quality of the approximate posterior predictive distribution for a specific SD-task can be evaluated by the KL-divergence between the true purchase distribution and the approximate posterior predictive distribution $\mathrm{KL}[p(y^r|D^r)||q_\phi(y^r|D^r)]$. In order to meta-learn fast and obtain an accurate approximation of the posterior predictive distribution for unseen shopping festivals or region with limited data, the goal of learning is set as to minimize the expected value of KL-divergence over different SD-tasks.

$$\phi^* = \underset{\phi}{\operatorname{argmin}}\, \mathbb{E}_R[\mathbb{E}_{p(D^r)}[\mathrm{KL}[p(y^r|D^r)||q_\phi(y^r|D^r)]]]$$

$$= \underset{\phi}{\operatorname{argmax}}\, \mathbb{E}_R[\mathbb{E}_{p(y^r,D^r)}[\log \int p(y^r|D^r,\psi^r)q_\phi(\psi^r|D^r)d\psi^r]]$$

As mentioned previously, the approximated posterior distribution of task-specific parameter $q_\phi(\psi^r|D^r)$ is modeled using a neural network, which we referred to as the **amortization network** with parameters ϕ. With task-specific parameters ψ^r sampled from the amortization network, a generative model $p(y^r|D^r,\psi^r)$ with shared global parameters $\theta_{s/d}$ is used to generate the predicted purchase y^r. The training process will therefore return the parameters ϕ and $\theta_{s/d}$ that best approximate the posterior distribution $p(y^r|D^r)$ in an average KL sense. Consequently, the representative capability of the amortization network (approximated posterior distribution $q_\phi(\psi^r|D^r)$) is essential to retrieve realistic purchase patterns for each SD-task. This plays a key role in recovering the true posterior distribution $p(\psi^r|D^r)$ through global optimization, and further leads to accurate purchase prediction.

To sum up, the proposed spatio-temporal meta-learning prediction (STMP) framework consists of an amortization network and a generative model. The details of the model construction and a new is ST-training procedure for the spatio-temporal prediction problem are described in later sections.

Model Construction

Amortization Network

We develop an amortization network (illustrated in Figure 3) to model the approximated posterior distribution $q_\phi(\psi^r|D^r)$ over spatio-temporal representation ψ^r of a target region and day type, which jointly considers the impacts of purchase features and related spatial feature. The overall construction of the amortization network is shown in Figure 3. To fully capture the temporal characteristics of purchase data of each SD-task, one needs to consider both the linear and non-linear temporal dependency among previous steps of historical purchase data, as well as the hidden periodic pattern of the purchase order and shopping cart records. To model such complex and dynamic time-series patterns, we use a linear transformation component to obtain a base estimate of purchase volume, and a special LSTM layer to capture the non-linear temporal pattern. The extracted temporal features are combined with static features through a feature fusion component to further support building a more informative spatio-temporal representation of purchase data.

Feature extraction. We first introduce a linear transformation component, modeled as a classical ***linear autoregressive*** (AR) model (Yates and Goodman 1999) that captures the linear temporal dependency among previous steps of purchase data, and obtains a rough but stable estimate of the current purchase orders from the historical purchase and shopping cart data. Denote the W as the parameters of the AR model and h_t are the purchase orders and shopping cart records of time step t. The AR model is formulated as:

$$h_t = W[h_1 \cdots h_{t-1}] + b$$

In the purchase prediction scenario, multiple shopping patterns with different periodicity may hide in the same time-series data. Conventional time-series models pay more attention to the neighboring time period and incapable of modeling multiple periodic patterns. In STMP, we use the ***skip-LSTM*** (Wang et al. 2019b) (see Figure 4(a)) which models the purchase time-series data with different skip intervals to learn multiple periodic patterns and outputs the temporal purchase embeddings for different intervals . The skip-LSTM is formulated as follows:

$$h_t = \mathrm{LSTM}_{\mathrm{skip}}(x_t, h_{t-p})$$

where $p = 1$ for mining recent purchase preference, 7 for the weekly pattern, and 30 for the monthly pattern.

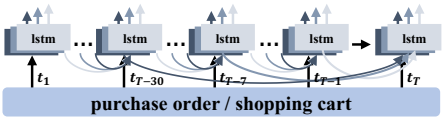

Figure 4: Skip-LSTM unit

Feature Fusion. To make full use of different features to give a more expressive representation, as shown in Figure 3, two ***fully connected components*** are used to integrate the dynamic and static features. One receives the output of the skip-LSTM to merge the temporal dependencies as well as long-term periodic characteristics of data. Another learns the embeddings of POIs, festival intervals and demographic profiles, and then aggregates them to output the final embedding of static features. This treatment balances the relative weights

of different groups of features, which helps to generate better feature representations from the data.

After obtaining the final feature embeddings, we can sample the task-specific spatio-temporal representation of an instance from $q_\phi(\psi^r|D^r)$ on the fly using means and variances provided by the amortization network. The biggest challenge of purchase prediction during shopping festivals is how to combat the data scarcity issue. To obtain a reliable spatio-temporal representation of an SD-task over the limited reference sample, we use the ***average pooling operation*** similar to (Qi et al. 2017) (shown as the "instance pooling" in the blue dotted box in Figure 3) to give the overall representation of training instances in a batch that comes from the same SD-task. The pooling operation can produce a more representative spatio-temporal embedding for instances, which is essential for the few-shot learning problem.

Generative Model for Purchase Prediction

We introduce a generative model $p(y^r|D^r, \psi^r)$ utilizing the shared statistical structure to learn the meta-knowledge of the SD-tasks. It greatly improves learning efficiency and prediction accuracy for task-specific prediction. The proposed generative model (illustrated in Figure 3) can be perceived similarly to the decoder of a VAE. It uses two different inputs to generate purchase predictions. The first input focuses on mining the current purchase embedding $\tilde{h}^r$ by applying the feature extraction and fusion techniques used in the amortization network on recent inputs $\tilde{x}^r_{\leq T}$, which describes the most recent variation of purchase patterns. The second input is the task-specific spatio-temporal representation ψ^r sampled from the amortization network ($q_\phi(\psi^r|D)$). It captures the macroscopic spatio-temporal pattern of the target task. Finally, the generative model combines the two inputs using a fully connected network to generate the prediction of purchase orders at a time step T.

ST-Training

As an SD-task only associated with limited data from a specific region and day type, which cannot fully describe the target purchase behavior. Inspired by multi-view learning (Xu, Tao, and Xu 2013), we build two views (spatial and temporal views) to describe different perspectives of data. These two views complement each other to enhance the learning information and obtain more accurate spatio-temporal data representation. The spatial view tends to learn the purchase pattern across different regions, while temporal view tends to mine the temporal changing patterns over different day types.

We proposed a new ST-training strategy to train STMP. In ST-training, both the amortization network $q_\phi(\tilde{y}^r|D^r)$ and the generative model $p(y^r|D^r, \psi^r)$ are jointly trained. To perform end-to-end training, we directly minimize $\mathrm{KL}[p(\tilde{y}^r|D^r)||q_\phi(\tilde{y}^r|D^r)]$ rather than $\mathrm{KL}[p(\psi^r|D^r)||q_\phi(\psi^r|D^r)]$, which gives following objective:

$$L(\phi, \theta_{s/d}) = -\mathbb{E}_R[\mathbb{E}_{p(D^r, \tilde{y}^r, \tilde{x}^r)}[\log \int p(\tilde{y}^r|\tilde{D}^r, \psi^r, \theta_{s/d}) q_\phi(\psi^r|D^r, \theta_{s/d}) d\psi^r]]$$

The training operation feeds a spatial or day type meta-representation ψ^r to the generator, which generates the predicted purchase order $\tilde{y}^r$ of task r. In ST-training, we alternatively select a spatial or temporal view by grouping the instances by corresponding region s_i or day type d_j, and then sample training data D^r. The sampled data are used to form the posterior predictive distribution $q_\phi(\psi^r|D^r)$, and further to compute $p_\phi(\tilde{y}|D^r, \psi^r)$. Alternating between different views during training allows transferring meta-knowledge across different regions or day types even when some regions or day types have a limited amount of data. Moreover, it also integrates the spatial and temporal shared statistical structure θ_s/d during training and improves the prediction accuracy. The complete training process is shown in Algorithm 1.

Algorithm 1 ST-Training

Input: Objective function $L(\phi, \theta_{s/d})$, spatial and day type dataset D, spatial region set S, and day type set DT.

Randomly initialize ϕ and θ; view $Q = S$;
repeat
  repeat
    Select a region or day type r from Q at random.
    Sample training data D^r from task r.
    Form the posterior predictive $q_\phi(\psi^r|D^r, \theta_{s/d})$.
    Compute the $\log q_\phi(y^r|D^r, \psi^r, \theta_{s/d})$.
    Update ϕ and $\theta_{s/d}$ by minimize $L(\phi, \theta_{s/d})$.
  until $\triangle L(\phi, \theta_{s/d}) < \epsilon$
  If $Q = S$, change view to DT, else to S .
until $\triangle L(\phi, \theta) < \epsilon$

Once the model is properly trained, the spatio-temporal representation ψ^r of different regions can be evaluated in advance by the amortization network using historical data. The predicted purchase orders can then be obtained by the generative model using task-specific spatio-temporal representation and embeddings of the current data features as input.

Experiments

In this section, we present the details of experiments, including dataset, model comparison and corresponding analysis.

Experimental Settings

Dataset We use a large-high-quality online purchase dataset from JD.com to evaluate our model. The JD dataset contains purchase order, shopping cart data as dynamic time-series features, and regional features such as POIs and demographic profile data as static features.

- **Purchase & Shopping cart data**. The study area of this work contains 18 regions of Beijing from 2015 to 2019. It contains 30 major product categories (e.g. apparel, electronics, food, etc.). The instances of purchase order and shopping cart are both region and time-dependent.
- **POIs data**. The POIs data contain the distribution of point location types (such as shops, hospitals, schools, etc.) in 18 regions of Beijing, which indirectly reflect purchase demand and the functional property of each region.
- **Spatial demographic data**. The spatial demographic data including population distributions of age, gender and individual buying power in 18 regions of Beijing, which reflect the regional purchase behavior of the population.

Methods	Overall		Weekend		Chinese national day		Double 11		Double 12		Mid-day promotion	
	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE
AR	0.0009	0.0294	0.0005	0.0236	0.0003	0.0181	0.0072	0.0848	0.0014	0.0372	0.0021	0.0459
DeepAR	0.0013	0.0363	0.0010	0.0312	0.0004	0.0203	0.0073	0.0856	0.0018	0.0428	0.0033	0.0566
LSTNet	0.0008	0.0295	0.0005	0.0236	0.0003	0.0181	0.0072	0.0848	0.0014	0.0371	0.0021	0.0459
Meta-GRU	0.0011	0.0339	0.0009	0.0296	0.0004	0.0196	0.0051	0.0717	0.0016	0.0404	0.0035	0.0591
STMP-AR	0.0010	0.0316	0.0009	0.0948	0.0008	0.0282	0.0081	0.090	0.0022	0.0469	0.0026	0.0509
STMP-VI	0.0003	0.0173	0.0002	0.0141	0.0002	0.0141	0.0019	0.0435	0.0005	0.0223	0.0011	0.0332
STMP-META	0.0004	0.0200	0.0002	0.0141	0.0002	0.0141	0.0019	0.0435	0.0005	0.0223	0.0011	0.0332
STMP-SKIP	0.0005	0.0223	0.0004	0.0200	0.0003	0.0173	0.0032	0.0565	0.0007	0.0264	0.0014	0.0374
STMP	**0.0004**	**0.0200**	**0.0002**	**0.0141**	**0.0001**	**0.0100**	**0.0016**	**0.0400**	**0.0004**	**0.0200**	**0.0011**	**0.0331**

Table 1: Evaluation results of STMP and the baseline methods for daily, weekend and shopping festivals scenarios

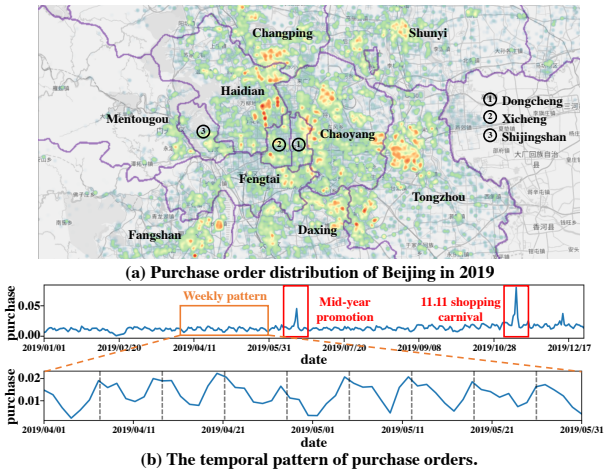

Figure 5: Spatio-temporal pattern of purchase in JD dataset.

Baselines We consider several state-of-the-art deep learning approaches and a few widely used time-series prediction methods as baselines:

- **AutoRegressive (AR)**. AR (Hamilton 1994) is a widely used time-series prediction model. Its prediction linearly depends on previous values as well as a stochastic term.
- **DeepAR**. DeepAR (Salinas et al. 2019) trains an autoregressive recurrent network on a large number of related time-series, which produces probabilistic forecasts.
- **LSTNet**. LSTNet (Lai et al. 2018) employs an autoregressive mechanism to produce the linear prediction based on the historical data, and then correct the prediction by a neural network, which consists of a temporal convolution, an LSTM, and multiple skip-LSTMs layers.
- **MetaGRU**. Inspired by ST-MetaNet (Pan et al. 2019), we design a network with MetaGRU units that consider the spatial meta-knowledge to enhance the prediction capability. It learns spatial meta-knowledge from the geographical information and generates the weights for the GRU units.
- **Variant of STMP**. We introduce multiple variants of STMP to fully evaluate its performance, including: 1) *STMP-AR*: we drop the AR component to evaluate the STMP model without linear transformation; 2) *STMP-SKIP*: there is no skip-LSTM layer to capture the periodic patterns in the time-series data; 3) *STMP-META*: the pooling layer is removed from the amortization network; 4) *STMP-VI*: to validate the performance of amortized variational inference, we use point estimation instead of distributional estimation for the ST-representation ψ^r.

Evaluation

We conduct experiments from four aspects: the purchase prediction during shopping festivals, the prediction accuracy in different SD-tasks, the generative performance under different number of shots, and results under different training settings. Mean squared error (MSE) and root mean squared error (RMSE) are used for the evaluation.

Prediction accuracy We first compare our model with the baselines on the JD dataset. To make a fair comparison, we present the best performance of each method under fine-tuned parameter settings in Table 1. It can be shown that STMP achieves superior performance over all the baseline methods, in both general purchase setting and bursty purchase prediction settings (e.g. shopping festivals on Double 11 and Mid-year promotion in China). In general purchase prediction scenarios, the shopping patterns are more regular and have the most sufficient data (about 300 days of a year). STMP outperforms the best baseline by at least 30% improvements on both MSE and RMSE. While in bursty purchase prediction scenarios, STMP still maintains about 30∼60% lower RMSE than the baselines. These results suggest that our method is effective and robust, especially in bursty purchase scenarios, when the actual purchase pattern is distinct than usual.

Several observations can be drawn from Table 1. DeepAR performs badly on almost all day types. A possible reason might be that it only considers non-linear temporal dependencies in time-series without accounting for more complex temporal characteristics (e.g. periodicity) and spatial attributes in the purchase data. LSTNet combines a linear transformation component with a recurrent neural network to add the linear dependency between historical and current temporal data. It improves the stability of prediction in non-special time periods but still can not adapt to the special shopping festivals with highly bursty purchase patterns. Meta-GRU has low prediction errors in the biggest shopping festival (11.11) due to the use of the meta-learning technique. However, only considering spatial meta-knowledge is insufficient to tackle such a complex prediction problem. STMP learns

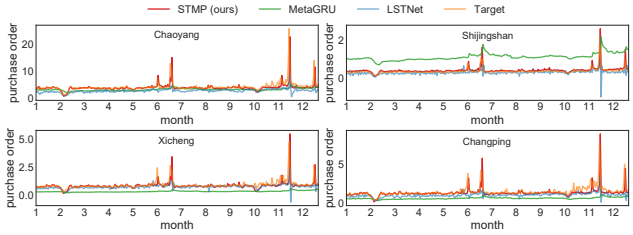

Figure 6: The predicted purchase for baselines and STMP in four representative regions

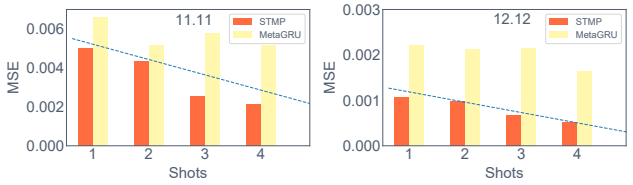

Figure 7: Results for different few-short settings

the task-specific spatial-temporal representation and uses the ST-training strategy to combine both the spatial and temporal meta-knowledge, which has superior expressive power.

To further investigate the impact of each model component of STMP, we also compare the performance of STMP with its variants in Table 1. It is observed that STMP-VI achieved good performance on overall and weekend scenarios that have relatively stable and certain purchase patterns, but perform worse than STMP on shopping festivals. This is because STMP-VI uses point estimates instead of distributional description of task-specific spatio-temporal representations, which is problematic when the data is limited or has highly uncertain patterns. Furthermore, the accuracy of STMP-SKIP is inferior compared with STMP or its other variants, but still outperforms other baselines. This testifies the need to account for multiple periodicity patterns in the purchase time-series. This is also reflected in Figure 5(b), we can see clear periodic patterns in the purchase time-series data. These demonstrate that modeling multiple periodicity property in data plays a significant role in improving model performance.

Prediction result in different SD-tasks Figure 6 shows the purchase prediction trends in 2019 of some typical baselines and STMP. We select four representative regions in Beijing that cover different numbers of purchase orders (see Figure 5(a) for detailed geographic locations). In most regions, the baselines are unable to predict the burstiness in purchase time-series well on shopping festivals, like the peaks in the purchase on Mid-year promotion and Double 11. They tend to give conservative results and only focus on improving overall accuracy. However, STMP utilizes the spatial-temporal representation of different regions and day types to support multi-task few-shot learning to facilitate better prediction for a specific region and day type. It can be observed that only STMP can predict the bursty purchase pattern well (fits to peaks in all figures).

Results of different shots of prediction Figure 7 provides a quantitative comparison between STMP and MetaGRU with different numbers of shots. We choose two major online shopping festival in the JD Dataset to demonstrate the effectiveness of STMP. As these shopping festivals only held once a year, for both STMP and MetaGRU, we conduct 1-shot to 4-shot experiments using different numbers of years' data (2015∼2018). From Figure 7, we can observe that the MSE of STMP decreases along with the increase of the number of shots (decrease from 0.005 to 0.002) in these shopping festivals. MetaGRU only considers learning meta-knowledge by transferring information across regions using static spatial features, but lacks consideration of transferring temporal meta-knowledge across different day types, which leads to less robust results for purchase prediction.

Methods	Double 11		Double 12		Mid-year promotion	
	MSE	RMSE	MSE	RMSE	MSE	RMSE
T-training	0.0039	0.0626	0.0008	0.0292	0.0015	0.0393
S-traning	0.0041	0.0643	0.0010	0.0321	0.0018	0.0422
ST-traning	**0.0016**	**0.0400**	**0.0004**	**0.0227**	**0.0011**	**0.0332**

Table 2: Experiment results under different training settings

Results under different training settings We further conduct experiments under different training modes, including T-training, S-training (only train in the temporal or spatial view), and combined ST-training to evaluate the strength of the ST-training. Table 2 shows the MSE and RMSE under different training settings for Double 11, Double 12, and Mid-year promotion shopping festivals. In the T-training, the pooling operation integrates instances in a batch from the same day type without considering the spatial knowledge. And the S-training integrates instances in a batch from the same region without considering the temporal information. From Table 2 we can see that even the incomplete T-training and S-training provide higher accuracy over the baselines listed in Table 1. However, lacking the joint meta-knowledge in space and time still leads to lower accuracy compared with the complete ST-training. The different training modes learn different transformations across regions and day types. The complementary training from spatial and temporal aspects can effectively improve the accuracy of the prediction.

Conclusion

We propose the Spatio-Temporal Meta-learning Prediction (STMP) model for purchase prediction during shopping festivals. Unlike other widely used approaches in time-series prediction problems, STMP jointly considers the short-term patterns and macroscopic spatio-temporal dependencies, which leads to superior performance in bursty purchase prediction tasks. In this model, we adopt a meta-learning framework with few-shot learning capability to capture task-specific spatio-temporal representations of data. The generative component of STMP uses the extracted spatio-temporal representation and input data to perform prediction inference. Extensive experiments on a large high-quality online purchase dataset from JD.com are used to evaluate the accuracy and meta-learning generalization ability of STMP. The proposed STMP outperforms baselines in all tasks, which demonstrate the effectiveness of the proposed model.

Acknowledgments

This work was supported by the National Key R&D Program of China (2019YFB2101801), the National Natural Science Foundation of China (No. 62076191), and the Beijing Nova Program (Z201100006820053).

References

Bengio, Y.; and LeCun, Y., eds. 2014. *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28(4): 594–611.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, 1126–1135.

Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; and Turner, R. E. 2019. Meta-Learning Probabilistic Inference for Prediction. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Grant, E.; Finn, C.; Levine, S.; Darrell, T.; and Griffiths, T. L. 2018. Recasting Gradient-Based Meta-Learning as Hierarchical Bayes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Hamilton, J. D. 1994. *Time series analysis*, volume 2. Princeton New Jersey.

Heskes, T. 2000. Empirical Bayes for Learning to Learn. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, 367–374.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 95–104.

Lake, B.; Salakhutdinov, R.; Gross, J.; and Tenenbaum, J. 2011. One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society*, volume 33.

Ma, F.; Chitta, R.; Zhou, J.; You, Q.; Sun, T.; and Gao, J. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 1903–1911.

Pan, Z.; Liang, Y.; Wang, W.; Yu, Y.; Zheng, Y.; and Zhang, J. 2019. Urban traffic prediction from spatio-temporal data using deep meta learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1720–1730.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, 77–85.

Qin, Y.; Song, D.; Chen, H.; Cheng, W.; Jiang, G.; and Cottrell, G. W. 2017. A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction. *CoRR* abs/1704.02971.

Ravi, S.; and Beatson, A. 2019. Amortized Bayesian Meta-Learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Reed, S. E.; Chen, Y.; Paine, T.; van den Oord, A.; Eslami, S. M. A.; Rezende, D. J.; Vinyals, O.; and de Freitas, N. 2018. Few-shot Autoregressive Density Estimation: Towards Learning to Learn Distributions. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Rezende, D. J.; Mohamed, S.; Danihelka, I.; Gregor, K.; and Wierstra, D. 2016. One-shot generalization in deep generative models. *arXiv preprint arXiv:1603.05106* .

Salinas, D.; Flunkert, V.; Gasthaus, J.; and Januschowski, T. 2019. DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting* .

Shu, R.; Bui, H. H.; Zhao, S.; Kochenderfer, M. J.; and Ermon, S. 2018. Amortized inference regularization. In *Advances in Neural Information Processing Systems*, 4393–4402.

Thrun, S.; and Pratt, L. 2012. *Learning to learn*. Springer Science & Business Media.

Wang, Y.; Smola, A.; Maddix, D. C.; Gasthaus, J.; Foster, D.; and Januschowski, T. 2019a. Deep factors for forecasting. *arXiv preprint arXiv:1905.12417* .

Wang, Y.; Yin, H.; Chen, H.; Wo, T.; Xu, J.; and Zheng, K. 2019b. Origin-destination matrix prediction via graph convolution: a new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1227–1235.

Xu, C.; Tao, D.; and Xu, C. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634* .

Yates, R. D.; and Goodman, D. J. 1999. Probability and stochastic processes. *John Willey & Sons* .

Yi, X.; Zhang, J.; Wang, Z.; Li, T.; and Zheng, Y. 2018. Deep distributed fusion network for air quality prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 965–973.

Zeng, M.; Cao, H.; Chen, M.; and Li, Y. 2019. User behaviour modeling, recommendations, and purchase prediction during shopping festivals. *Electronic Markets* 29(2): 263–274.