Modeling Customer Engagement from Partial Observations

Jelena Stojanovic, Djordje Gligorijevic, Zoran Obradovic
Computer & Information Sciences Department, Temple University
1925 N 12th Street
Philadelphia, USA
{jelena.stojanovic, gligorijevic, zoran.obradovic}@temple.edu

ABSTRACT

It is of high interest for a company to identify customers expected to bring the largest profit in the upcoming period. Knowing as much as possible about each customer is crucial for such predictions. However, their demographic data, preferences, and other information that might be useful for building loyalty programs is often missing. Additionally, modeling relations among different customers as a network can be beneficial for predictions at an individual level, as similar customers tend to have similar purchasing patterns. We address this problem by proposing a robust framework for structured regression on deficient data in evolving networks with a supervised representation learning based on neural features embedding. The new method is compared to several unstructured and structured alternatives for predicting customer behavior (e.g. purchasing frequency and customer ticket) on user networks generated from customer databases of two companies from different industries. The obtained results show 4% to 130% improvement in accuracy over alternatives when all customer information is known. Additionally, the robustness of our method is demonstrated when up to 80% of demographic information was missing where it was up to several folds more accurate as compared to alternatives that are either ignoring cases with missing values or learn their feature representation in an unsupervised manner.

CCS Concepts

•**Information systems** → **Data mining;**

Keywords

Structured Learning, Feature Learning, User Networks, Loyalty Programs, Deficient Data

1. INTRODUCTION

Companies utilize loyalty programs to enforce personalized customer relationship management. These programs can be considered as a guiding force of marketing endeavors, as good loyalty program has the power to turn a business into a customer-oriented profit machine [34]. Users' behavior can greatly differ, and so rewards and promotions that do not care about each individual user's behavior can result in a severe revenue decline [27].

In order to wisely plan enhancement of future customer engagement, it could be of great use to predict future behavior of customers including their customer's ticket[1] and visit frequency, which are good indicators of purchasing habits, and are also important indicators of the company's success. Therefore, companies are interested to detect different types of customers and to model their purchasing habits in order to properly build customer-tailored loyalty programs [9] and deepen customers loyalty to the brand.

In order to model users' behavior, a large variety of data needs to be collected. Each user action generates plenty of useful information about the user's habits, rendering loyalty programs one of marketing's biggest data-generating mechanisms[2]. Collection of data on actions and purchase details of customers can be fairly easy, but collection of demographic and other preference data may be challenging. Even though users are usually willing to turn over their basic demographic data (such as gender or date of birth) in exchange for perceived value, they are often dissuaded from using loyalty services if required to provide more than basic information or answer questionnaires. Additionally, many users may completely skip providing any demographic information if it is not obligatory. As such, analyzing customer data often requires dealing with a large fraction of missing values, which can severely limit the representational power of predictive models. However, companies would still like to infer about their customers in order to identify where customer-engagement marketing efforts should take place. For instance, based on certain estimates, such as forthcoming amount spent or visit frequency, they can quickly react to the market demand using personal recommendations via both online and offline channels, and/or by setting special offers with rewards, and discounts in order to increase the rate of customer retention.

A valuable source of information can be found in latent relations of customers. As similar customers tend to have similar purchasing behavior, aforementioned predictive objectives can potentially be achieved by modeling those relations and observing customers as nodes in a network. Therefore, our focus is primarily set on the structured models that are capable of utilizing such information. Continuous (Gaussian) Conditional Random Fields are such a model developed for structured regression [24], that has been successfully applied to a large variety of domains including climate [25, 28], energy [8], social networks [30] and healthcare [10, 11].

CIKM'16 , October 24-28, 2016, Indianapolis, IN, USA

DOI: http://dx.doi.org/10.1145/2983323.2983854

[1]Customer ticket is a common term for total dollar amount of transactions that customer spend over a certain period of time

[2]http://data-informed.com/customers-view-loyalty-programs-caution/ accessed May 2016

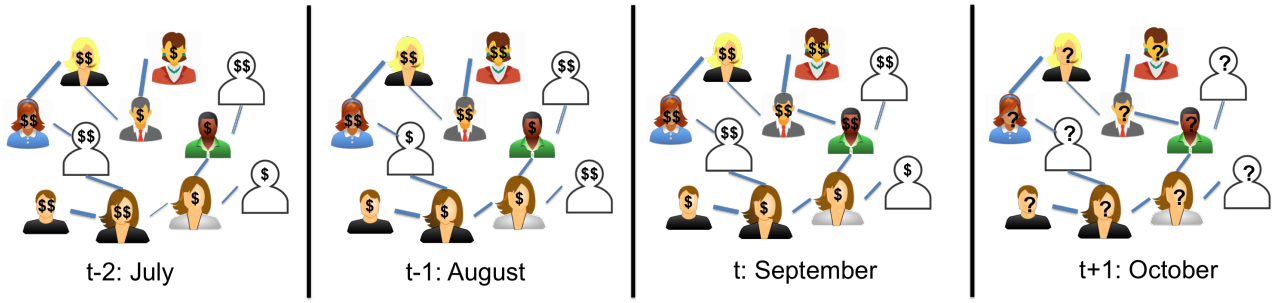


Figure 1: **Attributed** (purchase history and demographic data) **weighted** (different strength of customers similarity) **temporal** (months/quarters) **partially observed** (missing data in some of the nodes) **network** of customers in which explanatory variables (X) are partially observed (blank users: demographic data are missing) and the response variables (y) represent measurements of customer engagement (customer ticket or visit frequency). The goal is to learn parameters of the model on training data ($..., t-2, t-1, t$) and predict continuous responses on test examples ($t+1$).

This model is capable of structured regression for predicting customer tickets and visit frequencies, while modeling relationships among customers. However, it is limited to a given representation of the data, and is not robust to deficiency in explanatory variables. In order to improve representational and predictive power of this model, as well as to provide model robustness when a large fraction of explanatory variables are missing, we propose a supervised neural based feature embedding approach capable of determining a latent feature representation from partially observed explanatory variables within a structured regression framework.

The key contributions of this paper are summarized below:

- Modeling customer data is formulated as a structured regression problem, with emphasis on prediction of future customer's ticket and visit frequency, where a novel deep structured feature learning framework is proposed for joint learning of customers representation and their correlations in a supervised manner;
- The robustness of the approach is demonstrated while missing a large fraction of very useful demographic data in various patterns on several tasks (up to 80% of missing values);
- The model has shown experimental benefits compared to ten alternative models, including ones that are ignoring cases with scarce demographics as well as those that try to compensate for the deficiency of demographic data in an unsupervised fashion;
- The power and the generalization ability of the proposed approach are demonstrated on two challenging customer engagement applications on real-life data from different industries.

2. DATA

Customer engagement problems and proprietary datasets used to characterize effectiveness of the proposed method versus alternatives are described in this section.

2.1 Customer engagement data

Data from the business domain used in this study are based on electronically collected customer engagement information. Besides their purchase history (e.g. number of visits, items bought, discounts used, spending, etc.), we are partially familiar with their demographics, such as gender, age and similar information that a customer is asked to provide during an online registration/enrollment process. However, as previously mentioned, there is a number of reasons why customers would not provide their demographics. Furthermore, a company can decide to simplify enrollment process for the convenience of customers, thus choosing not to collect a valuable set of information. Even though some informative data about customers is missing, we would still like to accurately infer their future spending habits and the frequency of their visits.

For these two problems, in our experiments we drew datasets from two companies involved in different industry domains:

- The first company is from the entertainment industry and a large part of their loyalty programs are based on the monthly membership fee, thus it is important to estimate how often a customer will visit in the following month.
- The second company represents a global luxury lifestyle brand in the body and home products industry, which bases their members' rewards on quarterly spending. Estimating how much different customers will spend in the next quarter can be used to come up with new exclusive special offers and rewards to deepen customer engagement with the brand.

Therefore, we conducted experiments reported in Section 6 aimed to account for predictions of customers' future spending and visit frequency based on their recorded purchase history and their partially observed demographics. In the first application, data is collected over several months. However, the number of members for which demographics are known measures in the order of thousands. In the second application data is collected from the year 2012 and is aggregated on a quarterly level. The number of customers from the sample used in our experiments for which we know complete information (so that we can examine the influence of different processes and control experiments) measures in the order of hundreds.

Identification of the two companies is not shown in this study for privacy reasons. Also, showing sensitive information such as exact numbers of customers or exact customers' tickets and visit frequencies from the companies' databases is not being reported in this study.

2.2 Problem set-up

For each company we have a set of N customers $c_i \in \mathcal{C} = \{c_1, ..., c_N\}$. For each customer c_i we are familiar with the response variable y_i and a vector of m explanatory attributes: a P-dimensional vector of purchase data collected by the transaction system x_p and a D-dimensional vector of (partially) observed demographic data x_d, so that $m = P + D$ and $x_i = [x_p^{(i)}, x_d^{(i)}]$.

We observe a set of N customers $\mathcal{C}$ over T time steps and model them as a network as shown in Figure 2. Edges in the graph are weighted and represent similarity of response variables of the nodes. The goal is to predict values of the response variable y in each node of the graph in the following time step $t+1$.

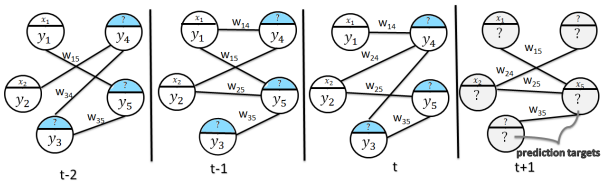
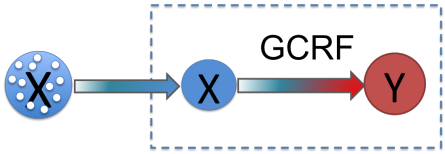


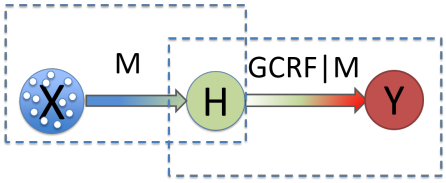
Figure 2: Attributed weighted network of customers observed over time in which explanatory variables (X) are partially missing (blue nodes) and dependent variables (y) represent the measure of customer engagement. The goal is to learn parameters of the model on data of the initial t steps and predict continuous response values y at time step $t+1$ (grey nodes).

3. RELATED WORK

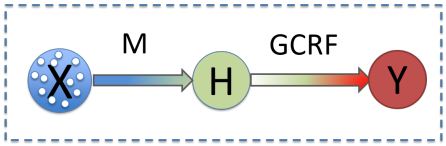
Given that we are dealing with a regression on data with underlying structures, we are using a structured regression model that has shown recent success in many applications, the GCRF model [24, 25]. Our proposed approach employs learning feature representations to improve predictive power of this method, as well as to handle existing data deficiency. Thus, in terms of robust modeling of explanatory feature mappings and desired predictive task, we could employ several strategies: a) Predictive modeling on a complete set (or subset) of *existing raw data*, ignoring partially observed nodes (schematics displayed in Figure 3a), b) *Unsupervised approach*: a common approach where features are learned in an unsupervised fashion prior to learning the predictive model (displayed in Figure 3b), c) *Supervised approach*: where features are learned simultaneously with the predictive model (displayed in Figure 3c). In this section we briefly discuss some of the state-of-the-art approaches related to unsupervised and supervised learning of input feature representations.



(a) Using GCRF model on the existing data



(b) Applying GCRF on a transformed representation H obtained by unsupervised approach



(c) Joint supervised learning of the input transformation $\mathcal{M}$ and the GCRF model.

Figure 3: Alternatives for applying GCRF on partially observed graphs

Unsupervised feature learning.

Many feature learning tasks are defined as unsupervised learning problems. For example, recent success of Deep Restricted Boltzman Machines [29] or Deep Autoencoders [13] has shown benefits of unsupervised feature learning. However, the main limitation of unsupervised feature learning is that it constrains to a parameter space of allowed solutions [1], and as such may not necessarily be optimized for the actual problem at hand. Typically, unsupervised feature learning methods determine the mapping $\mathcal{M}$ (Figure 3b) by optimizing an unsupervised objective (e.g., minimizing the reconstruction error), and afterwards a prediction algorithm can be applied to such transformed data.

Although this transformation $\mathcal{M}$ could capture the underlying structure, it does not necessarily capture the objective of the overall regression (e.g. maximizing the log likelihood of a regressor). Therefore the process involves two objectives: a) minimizing the reconstruction error as an unsupervised objective and b) maximizing the marginal likelihood as a supervised objective. Some studies have demonstrated benefits of training CRFs on feature representations learned by unsupervised learners for classification problems [7, 20]. In this study we will use this type of approach as one of baselines for structured regression and show that our proposed method outperforms such models in terms of accuracy.

Supervised feature learning.

In the area of supervised feature learning, several approaches were proposed for the CRF based classification [6, 7, 16, 17, 23, 25], where benefits were demonstrated on various applications. For example, learning of hidden states (or units) between explanatory variables X and response variable y is considered [16, 23], where this model is used for object detection and gesture recognition [23] and for optical character recognition, text classification, protein structure prediction, and part-of-speech tagging [16]. The approach is also applied to a phone classification task [17], and to ad targeting [6]. Success on a large variety of tasks has provided enough evidence that such methodology, if applied to continuous CRF's, could improve the model's representational power as well.

A different approach to modeling hidden units is to use neural networks architectures in the association potentials of the CRFs. This approach has shown benefits for both classification [7] and regression [25]. However, these models were either incapable of modeling complex relationships of response variables (only used a linear chain or a tree structure) or the interaction potential of the CRF's used predefined network structure as input, independently of other explanatory variables. In our approach, we propose using a neural architecture for the supervised mapping, on top of which both representation, as mapping $X \rightarrow y$, and a general graph structure are learned simultaneously.

In our experiments different related published approaches (described in Section 5.2) are used as baselines for comparison to the proposed supervised feature learning method. We provide evidence in several case studies for two prediction tasks that a supervised strategy is not only more accurate, but is more robust when applied to partially observed data.

4. THE MODEL

Here, we first describe the Gaussian Conditional Random Fields (GCRF) model and provide its interpretation in Section 4.1. Further, we specify the proposed Deep Feature Learning GCRF (DFL–GCRF) model in Section 4.2 and define feature embedding via neural mapping (Section 4.2.2), as the chosen mapping function for deep feature learning model.

4.1 Gaussian Conditional Random Fields

Gaussian Conditional Random Fields (GCRF) [24] is a discriminative structured regression model. The model captures both the

network structure of variables of interest (y) and relationship between attribute values of the nodes (X) and the target variable y. It is a model over a general graph structure (not only chains or trees), and can represent the relationships of the nodes as a function of time, space, or any user-defined structure. It models the structured regression problem by estimating a joint continuous distribution over all nodes. GCRF takes the following log-linear form:

$$P(y|X) = \frac{1}{Z(x,\alpha,\beta)} exp(\phi(y,X,\alpha,\beta)) = \frac{1}{Z(x,\alpha,\beta)} exp(\sum_{i=1}^{N} A(\alpha,y_i,X) + \sum_{i\sim j} I(\beta,y_i,y_j,X)) = \frac{1}{Z} exp(-\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k (y_i - R_k(X))^2 - \sum_{i\sim j}\sum_{l=1}^{L} \beta_l {S_{ij}}^{(l)} (y_i - y_j)^2) \quad (1)$$

The first part of the log-linear form $A(\alpha, y_i, X) = -\sum_{k=1}^{K} \alpha_k (y_i - R_k(X))^2$ is called *association potential* and it aims to model associations $X \rightarrow y_i$ using K different functions $R_k(X)$, which we will call unstructured predictors, as they are modeling these associations independently by learning from data or by using domain knowledge. Parameters of the association potential α_k are learned as degrees of belief towards each unstructured regressors. Given by the squared error $\sum_{i=1}^{N}(y_i - R(X))^2$, larger belief α is learned to correspond to the more accurate unstructured predictor.

The second part $I(\beta, y_i, y_j, X) = -\sum_{l=1}^{L} \beta_l {S_{ij}}^{(l)} (y_i - y_j)^2$ is called *interaction potential* and its goal is to utilize a graph structure S, that should be a weighted undirected network whose edges S_{ij} denote how similar two nodes are, or more precisely, how similar their response values y_i and y_j are. Parameters β are learned as degrees of belief towards similarity metrics and their values are governed by the product of similarity metric and squared distance $\sum_{i\sim j} S_{ij}(y_i - y_j)^2$. If this distance is small, relative value of β will be larger and the entire model will take the structure as an important source of information.

The normalization term

$$Z(x,\alpha,\beta) = \int_y \exp(\phi(y,X,\alpha,\beta))dy \quad (2)$$

and in general case, estimating this term is intractable. However, using quadratic feature functions, as demonstrated in Eq. 1, enables an elegant representation of the log-linear form as a multivariate Gaussian distribution [24]:

$$P(y|X) = \frac{1}{(2\pi)^{\frac{N}{2}} \mid \Sigma \mid^{\frac{1}{2}}} exp\left(-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right), \quad (3)$$

which allows efficient convex optimization. Here, Σ^{-1} represents the diagonally dominant inverse covariance matrix, and for this model takes the form:

$$\Sigma^{-1} = \begin{cases} 2\sum_{k=1}^{K} \alpha_k + 2\sum_g \sum_{l=1}^{L} \beta_l S_{ig}^{(l)}(x), i = j \\ -2\sum_{l=1}^{L} \beta_l S_{ij}^{(l)}(x), i \neq j \end{cases} \quad (4)$$

The posterior mean is given by $\mu = \Sigma b$, where b is defined as

$$b_i = 2\left(\sum_{k=1}^{K} \alpha_k R_k(X)\right). \quad (5)$$

4.1.1 Learning and inference

The learning task is to optimize parameters α and β by maximizing the conditional log–likelihood $\mathcal{L}$,

$$(\hat{\alpha}, \hat{\beta}) = \underbrace{argmax}_{\alpha,\beta} \mathcal{L} = \underbrace{argmax}_{\alpha,\beta} logP(y|X;\alpha,\beta). \quad (6)$$

Parameters α and β are learned by a gradient-based optimization. Gradients of the conditional log-likelihood are:

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = -\frac{1}{2}(y-\mu)^T \frac{\partial \Sigma^{-1}}{\partial \alpha_k}(y-\mu) + (\frac{\partial b^T}{\partial \alpha_k} - \mu^T \frac{\partial \Sigma^{-1}}{\partial \alpha_k})(y-\mu) + Tr(\Sigma \frac{\partial \Sigma^{-1}}{\partial \alpha_k}) \quad (7)$$

$$\frac{\partial \mathcal{L}}{\partial \beta_l} = -\frac{1}{2}(y+\mu)^T \frac{\partial \Sigma^{-1}}{\partial \beta_l}(y-\mu) + Tr(\Sigma \frac{\partial \Sigma^{-1}}{\partial \beta_l}) \quad (8)$$

Maximizing the conditional log–likelihood is a convex objective, and can be optimized using standard Quasi-Newton optimization techniques. Constraint of positive-semi definiteness of matrix Σ^{-1} ensures that the distribution is Gaussian. Therefore, to make the optimization unconstrained, the exponential transformation of parameters $\alpha_k = e^{u_k}$ and $\beta_l = e^{v_l}$ is used in GCRF [24].

In this model, prediction is governed by two parts, the association and interaction potentials. The association potential guides the main prediction power of the GCRF model and clearly, the more accurate the unstructured models are, the more GCRF will assimilate those predictions. On the other hand, as these unstructured predictors usually do not take into account the structure information, interaction potential will compensate that by introducing similarity matrix S, and bringing the predictions of the connected nodes closer together. A combination of the two potentials provides better accuracy than the unstructured predictors alone. However, as both unstructured predictor R and similarity matrix S are given (learned prior to GCRF model learning) they introduce a bias in the model. That is why in this study we propose a more complex, non-convex generalization of the GCRF model where R and S are learned within the GCRF framework. This extension will remove the bias of using pre-trained inputs. However, the bias will still be present in the form of chosen R and S functions. The trade off between model convexity and performance is a well studied topic and a number of studies have pointed out that convexity does not necessarily lead to the more powerful models [4, 14]. The new model will optimize the R and S for the overall regression goal and as such will improve its representational power.

4.2 Feature learning with the GCRF model

Most existing approaches often rely on a two-step process where a latent representation of explanatory variables is trained first, and its output is used to generate potentials for the structured predictor. This piece-wise training is, however, suboptimal, as the deep features are learned while ignoring the dependencies between the variables of interest. However, when learned jointly they can improve their predictive power by exploiting complementary information to build on the available data, and thus be beneficial for the overall regression task.

In order to implement this approach to the existing GCRF framework [24] and show its benefits, we have extended GCRF by:

- learning unstructured predictors $R(X,\theta)$ and similarity functions $S(x_i, x_j, \psi)$ together with learning α and β parameters of GCRF, rather than using them as pre-trained;

- defining a feature mapping function $\mathcal{M}(X, \xi)$ that takes available explanatory variables $x_i \in \mathbb{R}^{\mathrm{m}}$, for $i = 1, ..., N$ and maps them into $\mathbb{R}^{\mathrm{h}}$. [3] Both unstructured predictors and similarity metrics will be dependent on newly generated features, and we can formalize them as $R(\mathcal{M}(X, \xi), \theta)$ and $S(\mathcal{M}(X, \xi), \psi)$.

As our model performs feature learning together with learning input–output mapping and complex outputs' relations in a deep framework, we refer to this model as Deep Feature Learning GCRF model (DFL–GCRF). The diagram of the DFL–GCRF model is given in Figure 4.

This approach adds an additional three groups of parameters that are trained simultaneously with previously defined parameters α and β. In order for this extension to work, the unstructured predictor $R(x, \theta)$, similarity function $S(x_i, x_j, \psi)$ and feature mapping function $\mathcal{M}(x_i, \xi)$, need to be differentiable functions w.r.t. their parameters.

The final log-linear form of the DFL–GCRF:

$$P(y|X) = \frac{1}{Z} exp(-\sum_{i=1}^{N}\sum_{k=1}^{K} \alpha_k (y_i - R_k(\mathcal{M}(X, \xi), \theta_k))^2 - \sum_{l=1}^{L}\sum_{i\sim j} \beta_l {S_{ij}}^{(l)}(\mathcal{M}(X, \xi), \psi_l)(y_i - y_j)^2 \quad (9)$$

Then the inverse covariance (precision) matrix Σ^{-1} changes its form to:

$$\Sigma^{-1} = \begin{cases} 2\sum_{k=1}^{K} \alpha_k + 2\sum_g \sum_{l=1}^{L} \beta_l S_{ig}^{(l)}(\mathcal{M}(X, \xi), \psi_l), i = j \\ -2\sum_{l=1}^{L} \beta_l S_{ij}^{(l)}(\mathcal{M}(X, \xi), \psi_l), i \neq j \end{cases} \quad (10)$$

as well as b:

$$b_i = 2\left(\sum_{k=1}^{K} \alpha_k R_k(\mathcal{M}(X, \xi), \theta_k)\right) \quad (11)$$

The first moment of the multivariate Gaussian is obtained in the same way as before: $\mu = Q^{-1}b$.

This form of the model has the potential of using any linear or non-linear differentiable unstructured predictor, and any positive differentiable similarity function (the choice of these functions are presented in Section 5.1). However, joint optimization of the unstructured predictors and similarity metric with the GCRF doesn't allow for a convex optimization objective. An additional layer of complexity is introduced with the mapping function $\mathcal{M}(X, \xi)$. We describe solution for this complex optimization in the following section. With these additions we obtain a highly powerful and robust algorithm for modeling complex relationships.

4.2.1 Learning and Inference

The learning task is now to optimize parameters α, β, θ, ψ, ξ by maximizing the conditional log-likelihood,

$$(\hat{\alpha}, \hat{\beta}, \hat{\theta}, \hat{\psi}, \hat{\xi}) = \underbrace{argmax}_{\alpha,\beta,\theta,\psi,\xi} logP(y|X). \quad (12)$$

The modeled distribution is a multivariate Gaussian. Therefore, even though the objective function is no longer convex, it is a smooth function. As such, the parameters can still be learned by the gradient based methods with *warm start* techniques to avoid obvious local minimums [2]. The partial derivatives of the conditional log likelihood w.r.t. parameters α and β are given in in the Eq. 7 and Eq. 8, and derivatives w.r.t parameter θ_k will be:

$$\frac{\partial \log P}{\partial \theta_k} = \frac{\partial \log P}{\partial R_k}\frac{\partial R_k}{\partial \theta_k} \quad (13)$$

where $\frac{\partial \log P}{\partial R_k} = 2\alpha_k^T(y - \mu)$ and the second component depends on the chosen function R_k. The derivatives w.r.t parameters ψ_l are

$$\frac{\partial \log P}{\partial \psi_l} = -\frac{1}{2}(y + \mu)^T \frac{\partial \Sigma^{-1}}{\partial S_l}\frac{\partial S_l}{\partial \psi_l}(y - \mu) + \frac{1}{2}Tr(\Sigma\frac{\partial \Sigma^{-1}}{\partial S_l}\frac{\partial S_l}{\partial \psi_l}) \quad (14)$$

where derivatives depend on the chosen function S_l. Finally, derivatives w.r.t parameters ξ are

$$\frac{\partial \log P}{\partial \xi} = -\frac{1}{2}(y - \mu)^T \frac{\partial \Sigma^{-1}}{\partial \mathcal{M}}\frac{\partial \mathcal{M}}{\partial \xi}(y - \mu) + (\frac{\partial b}{\partial \mathcal{M}}\frac{\partial \mathcal{M}}{\partial \xi} - \mu^T\frac{\partial \Sigma^{-1}}{\partial \mathcal{M}}\frac{\partial \mathcal{M}}{\partial \xi})(y - \mu) + \frac{1}{2}Tr(\Sigma\frac{\partial \Sigma^{-1}}{\partial \mathcal{M}}\frac{\partial \mathcal{M}}{\partial \xi}), \quad (15)$$

where derivatives depend on the chosen input transformation $\mathcal{M}$, which will be discussed in detail in Section 4.2.2. The procedure for a gradient based optimization of the DFL–GCRF model is provided in the Algorithm 1.

Algorithm 1 DFL–GCRF optimization procedure

Input: Training data **X**, **y**
Initialize $\theta, \psi, \alpha, \beta, \xi$

(a) Estimate ξ by an unsupervised feature mapping strategy

(b) Estimate θ by learning unstructured predictor on mapped input space

(c) Estimate ψ by optimizing similarity for given nodes

(d) Estimate α, β by optimizing the GCRF model that uses unstructured predictor and similarity learned in steps 2(b) and 2(c) as inputs using Equations 7 and 8

repeat
Apply gradient based optimization to estimate all parameters using Equations 7, 8, 13, 14 and 15
until convergence

To avoid overfitting, which is a common problem for maximum likelihood optimization, we added regularization terms for α, β, θ, ψ, ξ to the log-likelihood to penalize large outputs of the parameters. The maximum posterior estimate of y is then obtained by computing the expected value μ: $\hat{y} = \underbrace{argmax}_{y} P(y|X) = \mu$.

In Section 5.1 a particular implementation of the architecture used in our experiments will be described in more details, including the choice of R (unstructured predictors), S (similarity) and $\mathcal{M}$ (mapping) functions.

4.2.2 Neural Mapping for GCRF

We consider the general setting where $\mathcal{M}(X, \xi)$ can be any arbitrary function of ξ and X. Options for $\mathcal{M}(X, \xi)$ reported in literature include different matrix factorization approaches on feature matrix X [15, 35] or various kernels [5, 22]. Matrix factorization approaches as well as different kernel approaches often fail to outperform neural feature mappings on tasks of feature learning [21].

[3] The dimension of latent features h is arbitrarily chosen by the user

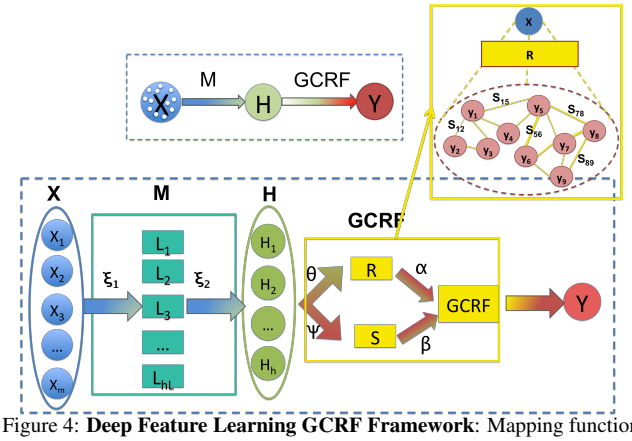


Figure 4: **Deep Feature Learning GCRF Framework**: Mapping function in our experiments is a neural network. This neural network will map explanatory attributes to attributes of a latent layer (H) and on such mapped data, GCRF is applied with linear regression as an unstructured predictor (R) and Gaussian kernel as a similarity function (S). Parameters of the mapping function (ξ), as well as parameters of the unstructured predictor (θ) and the similarity function (ψ) are learned together with GCRF objective function and its parameters (α and β).

Other approaches include dictionary learning [18], which often includes L_1 regularization to enforce sparsity, and as a consequence affects smoothness of the optimization function [33]. These methods require slow, dedicated methods for optimization, thus limiting their applicability to smaller datasets. In this study, a neural feature embedding architecture (Figure 4) is used, as recent advances in the field of feature learning show promising results when applying such embeddings to a variety of tasks in many domains [3, 7, 19]. Mapping function $\mathcal{M}(X, \xi) = \sigma(\xi X + b)$, with Sigmoid $\sigma(g) = \frac{1}{1+e^{-g}}$ and a matrix of mapping weights ξ, constitutes the first layer of deep architecture of DFL–GCRF framework (implementation details are given in Section 5.1).

Our hypothesis for such defined framework is that examples with partially missing inputs yield nearly equally good hidden representation as completely observed inputs. Previously, neural mapping approaches were successfully applied to the task of reconstructing corrupted input features, and theoretical analysis from several perspectives was provided on its validity [32]. The difference comparing to this approach is that we expect the neural mapping $\mathcal{M}(X, \xi)$ to learn response variable values y from deficient explanatory variables X supervisedly, instead of reconstructing the explanatory variables X, first, in an unsupervised manner. Experimental results showed that the proposed model successfully outperformed this baseline model and thus, backed up our hypothesis. Finally, our approach can be formalized as learning a stochastic mapper which transforms mixture of full and deficient data points to a manifold that allows for a linear mapping to the response variable $p(y|(H|(X_d, X_p)))$, with H being a representation of heterogeneous inputs that captures the main variations in the data w.r.t. response variable.

5. EXPERIMENTAL SETUP

In this section an experimental setup for the proposed and baseline methods are described.

5.1 The proposed method setup

In our experiments the DLR-GCRF uses a neural mapping as shown in Figure 4 with the following architecture: an input layer of dimensionality $\mathbb{R}^{\mathrm{m}}$, one hidden layer of dimension $\mathbb{R}^{\mathrm{h_L}}$, and $\mathbb{R}^{\mathrm{h}}$ dimensional output with distributed features. The number of neurons in a hidden layer of this neural mapping is $h_L \approx \frac{N}{\gamma*(m+h)}$ [12], where the number of outputs of this neural mapping is h and it is 3 in our experiments, and γ is chosen arbitrary in $[2-10]$ range.

The choice of unstructured predictor R is a linear model, which uses the first two learned distributed features, while the choice of S is a Gaussian kernel learned on the third distributed feature.

5.2 Baseline models setup

To evaluate the effectiveness of the DFL–GCRF model, we are comparing it to ten alternative methods from groups of models that are using the complete set or a subset of existing data as inputs and models that use latent features learned in an unsupervised manner.

First, we test the performance of the following baseline methods that learn their parameters only on the observed part of the data (ignoring the cases with missing inputs), as described in the Figure 3a.

- Linear Regression (iLR): We applied the unstructured linear predictor which captures the linear influence of explanatory variables X to a response variable y;
- Gaussian Processes Regression (iGP) [26]: We tested the GP model with a Gaussian Kernel $GK(x_i, x_j)$. Kernel optimized via GP objective function was further used as a network structure for structured models;
- Gaussian Conditional Random Fields (iGCRF): We also evaluate the GCRF model which utilizes the unstructured predictor and the available structure (in our experiments structure is node covariates learned with Gaussian Kernel).

Further, we compared the proposed model to several models from the group of unsupervised feature learning methods, as shown in Figure 3b: here, we apply one of the mapping functions and afterwards learn the GCRF regression model on such a mapped dataset (H). To isolate all other effects, we always used the same set-up for the structured GCRF model. This consists of an unstructured predictor of the GCRF model learned on mapped feature space using a linear model and interactions modeled via a Gaussian kernel function. Baseline mapping functions in this category that we applied are:

- Deep Autoencoders (DAE) [13, 32]: DAE aims to automatically learn features from unlabeled data by minimizing the input reconstruction error, namely, by learning a compressed, distributed representation (encoding) for a set of input data, typically for the purpose of dimensionality reduction;
- Principal Component Analysis (PCA) [31]: PCA aims to find a linear projection of high dimensional data into a lower dimensional subspace such that the variance is retained and the least square reconstruction error is maximized;
- Neural Mapping (NM) is learned in a supervised manner by optimizing a neural network (NN) for regression – mapping is defined as the last hidden layer of the neural network. The architecture of the NM is exactly the same as that of the neural mapping in the DFL–GCRF model. Note that this mapping is learned with the neural network optimization function and not with the GCRF optimization function;
- Zero imputation: In the situation when data are missing, a 0 value is imputed. As baselines in this category we used LR–0, GP–0 and GCRF on such 0–based imputed dataset.

The effectiveness of the proposed (DFL–GCRF) vs baseline methods (NM + GCRF, NN, DAE + GCRF, PCA + GCRF, GCRF, LR–0, GP–0, iGCRF, iLR and iGP) is evaluated on two applications described in Section 2 and the metric used for evaluation is the coefficient of determination defined as $R^2 = 1 - \frac{\sum_i (y_i - \mu_i)^2}{\sum_i (y_i - \hat{y})^2}$, where y_i and μ_i are true and predicted value for customer c_i, and $\hat{y}$ is the mean value for all customers in $\mathcal{C}$. We limit the values of R^2 to $[0, 1]$ scale, as we treat predictors with negative R^2 performance as useless, while predictors that obtain $R^2 = 1$ are considered to be a perfect fit to the data.

6. EXPERIMENTAL RESULTS

In this section the results are shown for: (1) predicting the customers' visit frequency in the following month, and (2) predicting individual customer's ticket in the upcoming quarter using their partially available demographic data, as well as their purchase history.

6.1 Prediction of visit frequency

The first company bases its membership on a monthly fee such that customers can use a certain number of provided services (depending on the program they signed up for). To provide "one-to-one" type messaging and added value that is unique to the customer, the aim is to estimate how often each customer uses purchased services in the following month. Even though these services are free of charge at the visit (for customers who paid the monthly fee), they often spend money on side products and services during the visit and bring additional revenue to the company. Therefore, with the knowledge of estimates of forthcoming visit frequencies, the company may build additional special offers to incentivize rare visitors or may reward the most loyal customers by, for example, providing instant benefits for a specific upcoming event. Additionally, the company may use this information to further adapt existing programs or educate and remind customers via targeted e-mail campaigns. To evaluate performance of the proposed model versus the alternatives, we conducted a variety of experiments corresponding to several real life situations that might occur with the loyalty program data.

6.1.1 *Predicting visits frequency on fully observed data*

In the first experiment, we assumed that all demographics prompted by the loyalty program about the company's customers are known. Such data is used to experimentally compare the proposed model with the baseline algorithms described in Section 5.2. Results of this experiment are shown in Table 1 in terms of R^2. Note that zero imputation and ignoring the missing cases are equivalent when there are no missing data.

Table 1: Accuracy comparison of DFL–GCRF vs 7 alternatives on complete data for prediction of a customer's visit frequency for the following month.

model	R^2
DFL–GCRF	**0.9147**
NM+GCRF	0.8793
GCRF	0.8652
GP	0.8582
NN	0.8525
LR	0.8502
PCA+GCRF	0.8350
DAE+GCRF	0.8063

From Table 1 we observe that the proposed DFL–GCRF model predicted frequency of visits with more accuracy than any of the alternatives considered. The improvements range from about 4% to 16.5%, where every percent of improvement can make a huge difference in terms of decision making when it comes to detecting the users from which the most revenue is generated. Additionally, we can see from the percentages of improvement that GCRF and GP models that account for correlations among the members were more accurate than unstructured predictors (NN and LR). Therefore, modeling relations between customers seems to be beneficial for this prediction task. Our experiments also suggest that unsupervised feature learning models tend to significantly lack accuracy since they are working with lower dimensional inputs. However, methods with neural feature mapping were more accurate. For example, supervised DFL–GCRF and unsupervised feature learning approach NM+GCRF were the best performing models (we also observe that structured model, NM+GCRF, brings improvements to the unstructured NN model). This, henceforth, confirms our assumption of superiority of neural mappings over alternatives as discussed in Section 4.2.2

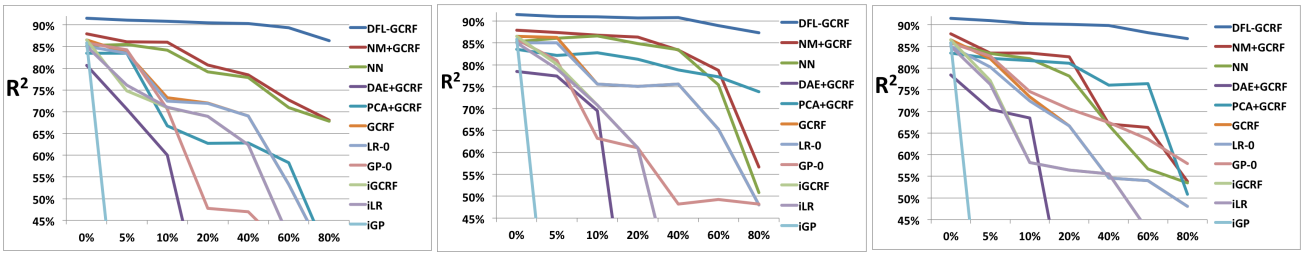
6.1.2 *Influence of data missingness mechanism on predicting visits frequency*

In the initial experiments we assumed that all customers were willing to share their demographic data whilst applying for a loyalty program. However, in practice, demographics are lacking in many cases. This is the why we conducted experiments where demographic data is reduced to a fraction of customers. Three types of missingness of demographic data were considered: a) removing demographics of random customers, b) removing data of customers who are least frequent visitors, and c) removing demographic data of the most frequent visitors. These three scenarios were considered for different fractions of missingness (5% – 80%) in order to characterize their robustness in various situations. The results for all missingness levels are shown in terms of R^2 in Figure 5.

The proposed DFL–GCRF model has outperformed the alternatives and has demonstrated the largest robustness in all three experiments. The overall accuracy improvement was about 5% to 55% vs. nontrivial alternatives for 10% of missing data and about 50% to 368% for 80% of missing data. Some of the baseline models (e.g. iGP) were not better than a mean predictor and are rendered as useless.

Demographic data missing at random (MAR).

In the experiment reported at Figure 5a we examined the situation in which random customers do not reveal their demographics. This way of inducing missingness does not mimic a real process. However, it is an unbiased way of examining the power of the models to handle missing demographic data. The most robust results were obtained by the proposed DFL–GCRF where rather stable accuracy is obtained up to 60% of missing data. NN and NM+GCRF were also somewhat robust but less accurate than DFL–GCRF. The accuracy in other baseline models considered dropped quite fast (after 10% of missing demographic data), with the exceptions of LR-0 and GCRF (which uses LR-0 as an unstructured predictor) that managed to maintain larger R^2 up to 60% of missing demographic data. The unsupervised feature learning models failed even after a few percentages of missing demographic data was induced. The unsupervised DAE approach of reconstructing inputs [32] under-performed on this task, as shown in Figure 5, and we see that the approach of supervised learning of the mapping function yielded vastly better results, and thus justified our original hypothesis. Also, we can see that imputation is a better strategy than ignoring data for each model where we employed these two strategies (LR, GP and GCRF).



(a) missing at random (b) missing for rare visitors (c) missing for frequent visitors

Figure 5: R^2 of predicting visits frequency by DFL–GCRF model vs ten alternatives for up to 80% of demographic information missing by 3 mechanisms.

Demographic data missing for the least frequent customers.

Experiments reported at Figure 5b examine accuracy when demographic data was missing for rare visitors, which is a common scenario in practice. The customers that are not well engaged with the brand may not be willing to spend their time and energy in filling out forms for registration purposes. The results of this experiment were similar to MAR results. This is due to the fact that majority of the customers we are modeling are actually customers with low frequencies of visits. The main difference between these two results is that when demographic data is missing for the least frequent users the accuracy of the models tend to drop more slowly than in MAR's case (accuracy remains relatively high up to 40% of missingness, rather than up to 10% in MAR). The top three models in these experiments were still neural feature learning models; with the addition of PCA as an unsupervised approach providing good results in the overall missingness induction process. We conclude that the most frequent users contribute the largest amount of variability in the data and thus are the ones from which PCA linear feature mapping can benefit the most.

Demographic data missing for the most frequent customers.

The results shown at Figure 5c are obtained for missing demographic data in a fraction of customers that are frequent visitors (the most loyal ones). We found that this hurts accuracy the most. The main difference versus results shown at Figures 5a and 5b is a drop in accuracy that occurred as soon as 5% of demographic data for the most frequent users was missing, which is about twice as large, compared to other missingness mechanisms. Additionally, missing values for the most frequent visitors reduced accuracy the most for all of the examined fractions.

6.2 Prediction of a customer's ticket

There are several ways in which this company can reward (and therefore incentivize) return customers, which include providing free platinum reward memberships to individuals that spend more than a certain amount over the course of a fiscal quarter, as well as sign–up, preferential, and birthday–related offers and rewards. However, while the company encourages customers to disclose birthday information by offering discounts during their birth month, many members still choose to keep this information private. This unwillingness for sharing makes determining future expenditures more difficult, but the company is still interested in selling more to each particular customer, either small or big spender, so it is important to be able to identify them. Our approach allows for accurate prediction of a customer's ticket even when a large fraction of customer demographic information is missing. To evaluate the power of the proposed method versus alternatives for the regression task of predicting customer's ticket for the following quarter, we conducted experiments based on several real life situations that might occur with the loyalty program:

- Experiment 1: both demographic data (queried at time of enrollment) and purchase history data is available for all customers
- Experiment 2: a random fraction of customers do not provide their demographic data, but their full purchase history is available
- Experiment 3: small spenders do not reveal their demographics, but purchasing details are available
- Experiment 4: big spenders do not reveal demographics, but purchasing details are available

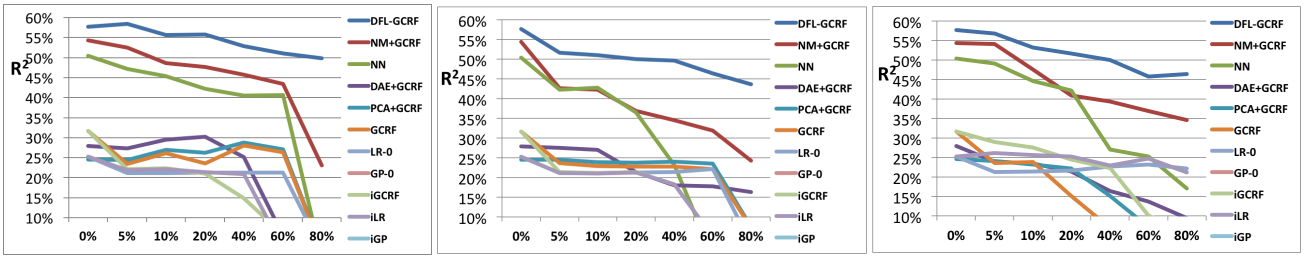
6.2.1 *Predicting customer's ticket on fully observed data*

The results for Experiment 1 are summarized in Table 2. We observe that the proposed DFL–GCRF model provides a wide range of improvements over alternatives (6% to 135%). In practice even a small percent of improvement vastly improves the quality of decision making in this application. We also see that NM+GCRF proves to be the best runner-up model, while GP, which fared well in previous experiments, underperformed on this predictive task (it was worse than a trivial mean predictor). Additionally, the rest of the baselines compare rather unfavorably to the proposed DFL–GCRF, which suggests that yet again, unsupervised feature learning is not an optimal strategy for predictive purposes.

Table 2: Accuracy comparison of DFL–GCRF vs 7 alternatives on complete data for prediction of a customer's ticket for the following quarter.

model	R^2
DFL–GCRF	**0.5771**
NM+GCRF	0.5436
NN	0.5041
GCRF	0.3165
DAE+GCRF	0.2789
LR	0.2527
PCA+GCRF	0.2454
GP	0

The GCRF model that is using LR as an unstructured predictor performs marginally better than unsupervised feature mappings, but it was less accurate than non-linear models.



(a) missing at random (b) missing for small spenders (c) missing for big spenders

Figure 6: R^2 of predicting customer's ticket by DFL–GCRF model vs ten alternatives for up to 80% of demographic information missing by 3 mechanisms.

6.2.2 *Influence of data missingness mechanisms on customer's ticket estimation*

In order to examine the robustness of DFL–GCRF for the customer's ticket prediction problem we induce up to 80% of missing values in demographic variables via three mechanisms. The R^2 results of ten baseline models and the DFL–GCRF are shown in Figure 6 for these missingess mechanisms.

We observe that, in all three experiments, the overall accuracy of the three neural-based models is the highest, and the gap between those models and the remaining baseline models is much larger for this dataset. From the non-neural based baselines in all three experiments we see that they are almost unaffected by the increasing missingness in the customer demographic data. We can thus conclude that for this application, these 8 alternative methods failed to utilize demographic information. For three neural-based models, we see that the accuracy drops much faster as compared to experiments reported in Section 6.1, even for DFL–GCRF (even though the drop of DFL–GCRF is the smallest compared to the alternative models). In the comparative test of R^2 and robustness, DFL–GCRF once again offers the best performance among the models used for different missingness mechanisms and different amounts of missing data, as shown in Figure 6.

The improvements of our model as compared to alternatives span an even larger range, starting from 11.85% and reaching into the thousands in some cases.

7. CONCLUSION

In this paper we introduced Deep Feature Learning GCRF, a powerful deep model for structured regression that learns hidden feature representation jointly with learning complex interactions of nodes in a graph. We have applied this method to two real-world customer engagement problems and provided evidence that the proposed method is capable of learning meaningful features for the purpose of regression, and outperforming other published alternatives developed with a similar aim. Additionally, we have tested the robustness of our method and other baselines when up to 80% of demographic data is missing by three mechanisms, and thus examined potential cases of missingness that might occur in the actual databases of companies. In future work we aim to further examine different feature learning approaches aimed to further improve both accuracy and robustness. We additionally aim to extend this approach to detect different groups of similar nodes in a network such that the model would work equally well in highly heterogeneous applications.

8. ACKNOWLEDGMENT

This research was supported in part by DARPA Grant FA9550–12–1–0406 negotiated by AFOSR, National Science Foundation through major research instrumentation grant number CNS–09–58854. The authors gratefully acknowledge use of the data, services and facilities of the Clutch Holdings LLC, Ambler, PA.

9. REFERENCES

[1] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[2] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. In *Neural Networks: Tricks of the Trade*, pages 437–478. Springer, 2012.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[4] Y. Bengio, Y. LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5), 2007.

[5] Y. Cho and L. K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, 2009.

[6] N. Djuric, V. Radosavljevic, M. Grbovic, and N. Bhamidipati. Hidden conditional random fields with distributed user embeddings for ad targeting. In *IEEE International Conference on Data Mining*, 2014.

[7] T.-M.-T. Do and T. Artieres. Neural conditional random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 177–184, 2010.

[8] T. Dokic, P. Dehghanian, P.-C. Chen, M. Kezunovic, Z. Medina-Cetina, J. Stojanovic, and Z. Obradovic. Risk assesment of a transmission line insulation breakdown due to lightning and sever weather. In *HICSS-49*, 2016.

[9] G. R. Dowling and M. Uncles. Do customer loyalty programs really work? *Research Brief*, 1, 1997.

[10] D. Gligorijevic, J. Stojanovic, and Z. Obradovic. Improving confidence while predicting trends in temporal disease networks. In *4th Workshop on DMMH, 2015 SIAM International Conference on Data Mining*, 2015.

[11] D. Gligorijevic, J. Stojanovic, and Z. Obradovic. Uncertainty propagation in long-term structured regression on evolving networks. In *Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, 2016.

[12] M. Hagan, H. B. Demuth, M. Beale, and O. De Jesus. *Neural network design*. Martin Hagan, 2014.

[13] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[14] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] O. Levy and Y. Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[16] L. Maaten, M. Welling, and L. K. Saul. Hidden-unit conditional random fields. In *International Conference on Artificial Intelligence and Statistics*, pages 479–488, 2011.

[17] M. Mahajan, A. Gunawardana, and A. Acero. Training algorithms for hidden conditional random fields. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2006.

[18] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *26th Annual International Conference on Machine Learning*. ACM, 2009.

[19] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[20] A.-r. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, M. Picheny, et al. Deep belief networks using discriminative features for phone recognition. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5060–5063. IEEE, 2011.

[21] B. Perozzi, R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In *20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 701–710. ACM, 2014.

[22] D. Qin, X. Chen, M. Guillaumin, and L. V. Gool. Quantized kernel learning for feature matching. In *Advances in Neural Information Processing Systems*, pages 172–180, 2014.

[23] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. on Pattern Analysis & Machine Intelligence*, 2007.

[24] V. Radosavljevic, S. Vucetic, and Z. Obradovic. Continuous conditional random fields for regression in remote sensing. In *19th European Conf. on Artificial Intelligence*, 2010.

[25] V. Radosavljevic, S. Vucetic, and Z. Obradovic. Neural gaussian conditional random fields. In *Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2014.

[26] C. E. Rasmussen. *Gaussian processes for machine learning*. Citeseer, 2006.

[27] B. Stauss, M. Schmidt, and A. Schoeler. Customer frustration in loyalty programs. *International Journal of Service Industry Management*, 16(3):229–252, 2005.

[28] J. Stojanovic, M. Jovanovic, D. Gligorijevic, and Z. Obradovic. Semi-supervised learning for structured regression on partially observed attributed graphs. In *SIAM International Conference on Data Mining*, 2015.

[29] T. Tieleman. Training restricted boltzmann machines using approximations to the likelihood gradient. In *25th International Conference on Machine learning*, 2008.

[30] A. Uversky, D. Ramljak, V. Radosavljević, K. Ristovski, and Z. Obradović. Panning for gold: using variograms to select useful connections in a temporal multigraph setting. *Social Network Analysis and Mining*, 4(1):1–13, 2014.

[31] L. J. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

[32] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[33] M. Wytock and Z. Kolter. Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting. In *Proc. of the 30th International Conference on Machine Learning (ICML-13)*, pages 1265–1273, 2013.

[34] Y. Yi and H. Jeon. Effects of loyalty programs on value perception, program loyalty, and brand loyalty. *Journal of the Academy of Marketing Science*, 31(3):229–240, 2003.

[35] J. Zhou, F. Wang, J. Hu, and J. Ye. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *20th ACM SIGKDD Internation Conference on Knowledge Discovery and Data Mining*, pages 135–144. ACM, 2014.