Unsupervised Learning for Computational Phenotyping

Chris Hodapp (chodapp3@gatech.edu)

***Abstract*—With large volumes of health care data comes the research area of computational phenotyping, making use of techniques such as machine learning to describe illnesses and other clinical concepts from the data itself. The "traditional" approach of using supervised learning relies on a domain expert, and has two main limitations: requiring skilled humans to supply correct labels limits its scalability and accuracy, and relying on existing clinical descriptions limits the sorts of patterns that can be found. For instance, it may fail to acknowledge that a disease treated as a single condition may really have several subtypes with different phenotypes, as seems to be the case with asthma and heart disease. Some recent papers cite successes instead using unsupervised learning. This shows great potential for finding patterns in Electronic Health Records that would otherwise be hidden and that can lead to greater understanding of conditions and treatments. This work implements a method derived strongly from Lasko *et al.*, but implements it in Apache Spark and Python and generalizes it to laboratory time-series data in MIMIC-III. It is released as an open-source tool for exploration, analysis, and visualization, available at: https://github.com/Hodapp87/mimic3_phenotyping.**

***Index Terms*—Big data, Health analytics, Data mining, Machine learning, Unsupervised learning, Computational phenotyping**

I. Introduction & Background

The field of *computational phenotyping*[1] has emerged recently as a way of learning more from the increasing volumes of Electronic Health Records available, and the volume of this data ties it in naturally with fields like machine learning and data mining. The "traditional" approach of supervised learning over classifications has two noted problems:

- It requires the time and attention of that domain expert in order to provide classification information over which to train a model, and this requirement on human attention limits the amount of data available (and, to an extent, its accuracy.)
- It tends to limit the patterns that can be found to what existing classifications acknowledge. If a disease treated as a single condition really has multiple subtypes with different phenotypes, the model will not reflect this - for instance, asthma and heart disease[9].

Some recent papers[13], [9], [5] cite successes with approaches instead using unsupervised learning on time-series data. In Lasko *et al.*[9], such an approach applied to serum uric acid measurements was able to distinguish gout and acute leukemia with no prior classifications given in training. Marlin *et al.*[13] examines 13 physiological measures from a pediatric ICU (such as pulse oximetric saturation, heart rate, and respiratory rate). Che *et al.*[1] likewise uses ICU data, but focuses on certain ICD-9 codes rather than mortality.

This approach still has technical barriers. Time-series in healthcare data frequently are noisy, spare, heterogeneous, or irregularly sampled, and commonly Gaussian processes are employed here in order to condition the data into a more regular form as a pre-processing step. In [9], Gaussian process regression produces a model which generates a continuous, interpolated time-series providing both predicted mean and variance, then applies a two-layer stacked sparse autoencoder (compared with a five-layer stacked denoising autoencoder in [1], without Gaussian process regression).

The goal undertaken here was to reimplement a combination of some earlier results (focusing mainly on that of Lasko *et al.*[9]) using Apache Spark and the MIMIC-III critical care database[6], and able to run primarily on a "standard" Spark setup such as Amazon Elastic MapReduce. However, presently it relies on Python in order to use Keras and scikit-learn for feature learning and t-SNE.

The software behind this work is also released as an open source tool for accomodating exploration, analysis, and visualization using the techniques described herein. It is available at: https://github.com/Hodapp87/mimic3_phenotyping.

II. Approach & Implementation

The main problems that the implementation tries to address within these parameters are:

- Loading MIMIC-III data into a form usable from Spark
- Identifying relevant laboratory tests, admissions, and ICD-9 codes on which to focus
- Preprocessing the time-series data with Gaussian process regression
- Using a two-layer stacked sparse autoencoder to perform feature learning
- Visualizing the new feature space and identifying potential clusters

This section describes the general approach, and the Experimental Evaluation section on page 4 gives specific examples that were tested.

A. Loading & Selecting Data

The MIMIC-III database is supplied as a collection of `.csv.gz` files (that is, comma-separated values, compressed with gzip). By way of `spack-csv`, Apache Spark 2.x is able to load these files natively as tabular data, i.e. a `DataFrame`. All work described here used the following tables[6]:

- `LABEVENTS`: Timestamped laboratory measurements for patients

- `DIAGNOSES_ICD`: ICD-9 diagnoses for patients (per-admission)
- `D_ICD_DIAGNOSES`: Information on each ICD-9 diagnosis
- `D_LABEVENTS`: Information on each type of laboratory event

The process requires two ICD-9 categories, and one LOINC code for a lab test. Admissions are filtered to those which contain a lab time-series of the given LOINC code and containing at least 3 samples, containing either ICD-9 category mutually exclusively (that is, at least one diagnosis of the first ICD-9 category, but none of the second, or at least one of the second ICD-9 category, and none of the first).

As an aid to this process, the tool can produce a matrix in which each row represents an ICD-9 category (starting with the most occurrences and limited at some number), each column represents a likewise ordered LOINC code (http://loinc.org/), and each intersection contains the number of admissions with an ICD-9 diagnosis of that category and a laboratory time-series of that LOINC code. It does not tell whether a pair of ICD-9 categories, mutually excluding each other, may produce enough data, but it may still give a meaningful estimate. The below shows an example of this matrix, limited to the top 4 LOINC codes (incidentally, all blood measurements) and top 12 ICD-9 categories for space reasons:

ICD-9	LOINC code			
category	11555-0	11556-8	11557-6	11558-4
427	12456	12454	12458	12779
276	11392	11393	11393	11962
428	11198	11195	11196	11515
401	11186	11187	11188	11525
518	11386	11387	11386	11545
250	8238	8238	8240	8574
414	9243	9242	9242	9412
272	7733	7733	7736	7919
285	6423	6421	6422	6720
584	6541	6541	6542	6834
V45	4862	4860	4861	5068
599	3815	3816	3815	3983

All processing at this stage was done via Spark's DataFrame operations, aside from the final conversion to an RDD containing individual time-series.

B. Preprocessing

1) Time Warping: The covariance function that is used in Gaussian process regression (and explained after this section) contains a time-scale parameter τ which embeds assumptions on how closely correlated nearby samples are in time. This value is assumed not to change within the time-series - that is, it is assumed to be *stationary*[9]. This assumption is often incorrect, but under the assumption that more rapidly-varying things (that is, shorter time-scale) are measured more frequently, an approximation can be applied to try to make the time-series more stationary - in the form of changing the distance in time between every pair of adjacent samples in order to shorten longer distances, but lengthen shorter ones[9]. For original distance d, the warped distance is $d' = d^{1/a} + b$, using $a = 3, b = 0$ (these values were taken directly from equation 5 of [9] and not tuned further).

Thomas Lasko also related in an email that this assumption (that measurement frequency was proportional to how volatile the thing being measured is) is not true for all medical tests. He referred to another paper of his[8] for a more robust approach, however, this is not used here.

2) Gaussian Process Regression: In order to condition the irregular and sparse time-series data from the prior step, Gaussian process regression was used. The method used here is what Lasko *et al.*[9] described, which in more depth is the method described in algorithm 2.1 of Rasmussen & Williams[15].

In brief, Gaussian process regression (GPR) is a Bayesian non-parametric, or less parametric, method of supervised learning over noisy observations[3], [9], [15]. It is not completely free-form, but it infers a function constrained only by the mean function (which here is assumed to be 0 and can be ignored) and the covariance function $k(t, t')$ of an underlying infinite-dimensional Gaussian distribution. It is not exclusive to time-series data, but t is used here as in this work GPR is done only on time-series data.

That covariance function k defines how dependent observations are on each other, and so a common choice is the squared exponential[3]:

$$k(t, t') = \sigma_n^2 \exp\left[\frac{-(t-t')^2}{2l^2}\right]$$

Note that k approaches a maximum of σ_n^2 as t and t' are further, and k approaches a minimum of 0 as t and t' are closer. Intuitively, this makes sense for many "natural" functions: we expect closer t values to have more strongly correlated function values, and l defines the time scale of that correlation.

The rational quadratic function is used here instead as it better models things that may occur on many time scales[9]:

$$k(t, t') = \sigma_n^2 \left[1 + \frac{(t-t')^2}{2\alpha\tau^2}\right]^{-\alpha}$$

Gaussian process regression was implemented from algorithm 2.1 of Rasmussen & Williams[15], copied below:

$$L = \text{cholesky}(K + \sigma_n^2 I) \tag{1}$$
$$\boldsymbol{\alpha} = L^\top \backslash (L \backslash \mathbf{y}) \tag{2}$$
$$\bar{f}_* = \mathbf{k}_*^\top \boldsymbol{\alpha} \tag{3}$$
$$\bar{v} = L \backslash \mathbf{k}_* \tag{4}$$
$$\mathrm{V}\left[f_*\right] = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^\top \mathbf{v} \tag{5}$$
$$\log p(\mathbf{y}|X) = -\frac{1}{2}\mathbf{y}^\top \boldsymbol{\alpha} - \sum_i \log L_{ii} - \frac{n}{2}\log 2\pi \tag{6}$$

X is the training inputs (a $n \times 1$ matrix for a time-series of m samples), y is a $n \times 1$ matrix with the corresponding values to X, and $\mathbf{x}_*$ is a test input (a scalar), for which $\bar{f}_*$ are $\mathrm{V}\left[f_*\right]$ the predictions (respectively, predictive mean and variance).

K is a $n \times n$ matrix for which $K_{ij} = k(X_i, X_j)$, $\mathbf{k}_*$ is a $m \times 1$ matrix for which $(\mathbf{k}_*)_i = k(X_i, \mathbf{x}_*)$, and $A \backslash B$ (both A and B matrices) is the matrix x such that $Ax = B$.

Matrices L and $\boldsymbol{\alpha}$ in effect represent the Gaussian process itself (alongside the covariance function and its hyperparameters), and can be reused for any test inputs $\mathbf{x}_*$. This implementation also exploits the fact that the algorithm trivially generalizes to multiple $\mathbf{x}_*$ in matrix form and produces multiple $\bar{f}_*$ and $\mathrm{V}[f_*]$.

Line 6 provides the log marginal likelihood of the target values $\mathbf{y}$ given the inputs X, and this was the basis for optimizing the hyperparameters of the covariance function. In specific, in order to optimize the hyperparameters σ_n, τ, and α, every individual time-series was transformed with the time-warping described in the prior section, standardized to mean of 0 and standard deviation of 1 (note that the mean and standard deviation must be saved in order to undo this transformation when interpolating), and then $\sum \log p(\mathbf{y}|X)$ over the entire training set was maximized using a grid search.

Hyperparameter optimization is likely the most time-consuming step of processing, and Gaussian process regression is very sensitive to their values. However, this step also is a highly parallelizable one, and so it was amenable to the distributed nature of Apache Spark (and likely to more efficient methods such as gradient descent).

The values (i.e. y) in each individual time-series were standardized to a mean of 0 and standard deviation of 1, and this standardization was then reversed on the interpolated values.

3) Interpolation: The remainder of algorithm 2.1 is not reproduced here, but the code directly implemented this method using the rational quadratic function (and hyperparameters given above) as the covariance function. This inferred for each individual time-series a continuous function producing predictive mean and variance for any input t (for which they use the notation $\mathbf{x}^*$ for "test input").

As in [9], all of these inferred functions were then evaluated at a regular sampling of time values (i.e. via the test input $\mathbf{x}^*$) with padding added before and after each time series. The sampling frequency and the amount of padding depends on the dataset, and so can be specified when running the tool.

In effect, this mapped each individual time-series first to a continuous function, and then to a new "interpolated" time-series containing predicted mean and variance at the sampled times described above. The interpolated time-series first had the reverse transformation applied from their standardization (i.e. the stored mean was added back in), and this was then written to CSV files and used as input to later steps.

C. Feature Learning with Autoencoder

A stacked sparse 2-layer autoencoder was then used to perform feature learning. The implementation used here was a combination of what was described in [9] (which closely follows the UFLDL Tutorial from Stanford[14]), and François Chollet's guide[2] on using the Python library Keras to implement autoencoders. Specifically, Keras was used with Theano (GPU-enabled) as the backend; all data was loaded from the prior step with the pandas library.

These autoencoders have a fixed-size input and output, and in order to accomodate this, fixed-size contiguous patches

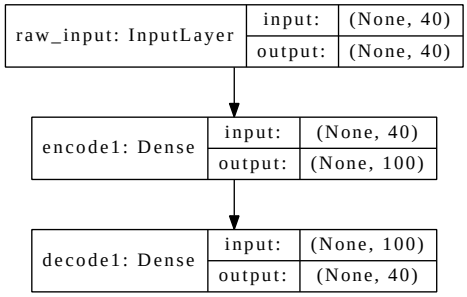

Fig. 1. First stage of Keras autoencoder

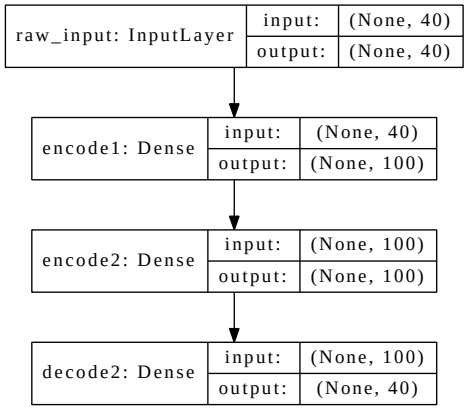

Fig. 2. Second stage of Keras autoencoder

were sampled from the interpolated time-series. As in [9], the patch size was set to the total number of padded samples, and patches were sampled uniformly randomly from all contiguous patches. Note that since Gaussian process regression produces both mean and variance predictions, the input and output size of the network are twice the patch size.

As in [9], the encoder and decoder layers used sigmoid encoders and linear decoders, and all hidden layers had 100 units. Both layers contained a sparsity constraint (in the form of L1 activity regularization) and L2 weight regularization, and performance appeared very sensitive to their weights.

This network was built up in stages in order to greedy layerwise train[14]. The first stage was the neural network in figure 1; this was trained with the raw input data (at both the input and the output), thus learning "primary" hidden features at the layer `encode1`. The first decoder layer `decode1` was then discarded and the model extended with another encoder and decoder, as in figure 2.

This model then was similarly trained on raw input data, but while keeping the weights in layer `encode1` constant. That is, only layers `encode1` and `decode2` were trained, and in effect they were trained on "primary" hidden features (i.e. `encode1`'s activations on raw input). Following this was "fine-tuning"[14] which optimized all layers, i.e. the stacked autoencoder as a whole, again using the raw input data.

The final model then discarded layer `decode2`, and used the activations of layer `encode2` as the learned sparse features. (Elsewhere in the paper, "second-layer learned features" refers to these activations on a given patch of time-series

input. “First layer learned features” refers to the activations of `encode1`.)

D. Visualization & Classification

The final processing of the tool consists of using the learned sparse features of the prior step (from both the first and second layers) as input into two separate steps: a visualization by way of t-SNE (t-distributed Stochastic Neighbor Embedding), and training a logistic regression classifier.

Both of these steps used the Python library scikit-learn, and respectively sklearn.manifold.TSNE and sklearn.linear_model.LogisticRegression.

III. Experimental Evaluation

As a test of the tool described in this paper, experiments were run on a selection of the data. Particularly, the LOINC code `1742-6`, corresponding to MIMIC-III `ITEMID` of 50861, Alanine Aminotransferase (ALT), was used, and the ICD-9 categories 428 (heart failure) and 571 (chronic liver diseases), corresponding to ALT’s use as a biomarker for liver health and to suggest congestive heart failure. All time-series for this were in international units/liter (IU/L).

This were selected to a total of 3,553 unique admissions (1,782 for ICD-9 category 428, and 1,771 for 571), 3,320 patients (1,397 females, 1,923 males), and 34,047 time-stamped samples. 70% of these admissions were randomly selected for the training set, and the remaining 30% for the testing set (2,473 and 1,080 respectively). The interpolated series from Gaussian process regression were padded by 10 samples at the beginning and end and sampled at 0.25 days (thus 2.5 days of padding), producing a total of 198,830 samples.

These interpolated time-series were randomly resampled to 7,419 patches of 20 samples long (thus, the neural network used inputs and outputs of 40 nodes). 20% of these were set aside for cross-validation.

Figure 3 is an example of a time-series from this data. The solid black line is the “original” time-series, the red line is the warped version, and the blue line is the version after interpolation with Gaussian process regression (with one standard deviation plotted on the surrounding dotted line). Figure 4 is several other randomly-chosen time-series from the data as examples.

When training the autoencoder on this data, manual tuning led to an L1 activity regularization (i.e. sparsity constraint) of 10^{-4} and L2 weight regularization of 10^{-3}.

An interesting detail which figure 2 of [9] shows is that the effects of the first layer can be visualized directly in the form of its weights (not its activations; this is the result of its training, not of any input). As they form a 40x100 array, each 40-element vector corresponds to a sort of time-series signature which that unit is detecting. Figure 5 shows the corresponding plot of first-layer weights (i.e. `encode1`) after it is trained on the subset described here.

This shows similar structure as in [9] (including considerable redundancy), but with different sorts of signatures. Particularly, it seems to single out edges and certain kinds of ramps.

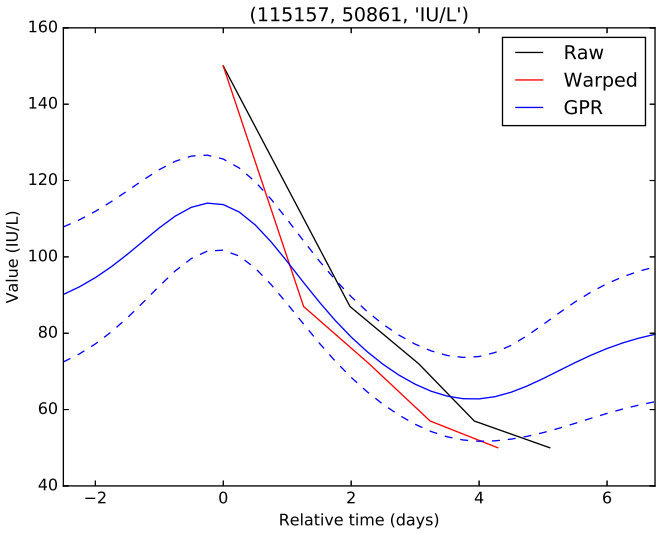

Fig. 3. Time-series: Original, warped, and GPR interpolated

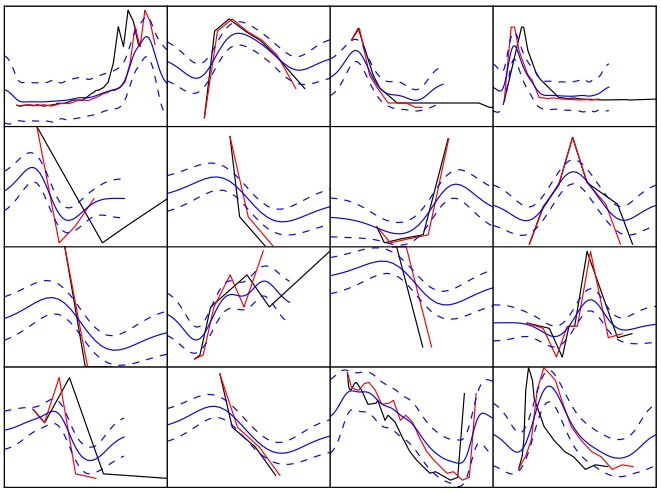

Fig. 4. Example time-series

The t-SNE results, shown below in figures 6 and 7, are inconclusive here. It appears to have extracted some structure (which in the second layer is better-refined), but this structure does not seem to relate clearly with the labels.

The classifier here is not producing useful results; particularly, it is producing an AUC of 0.5 on both the 1st and 2nd layer features. The reason for the poor performance is not known, but due to this, it was not compared against any “baseline” classifier or expert-feature classifier as in [9]. Overall, further work is needed here.

IV. Conclusions & Further Work

The tool is released as open source built on openly available libraries and (mostly) open data sources. It was sufficient to produce all diagrams, plots, and analysis in this paper. However, it still needs further experimentation to produce meaningful results, and the intention is that it can be a starting point for this.

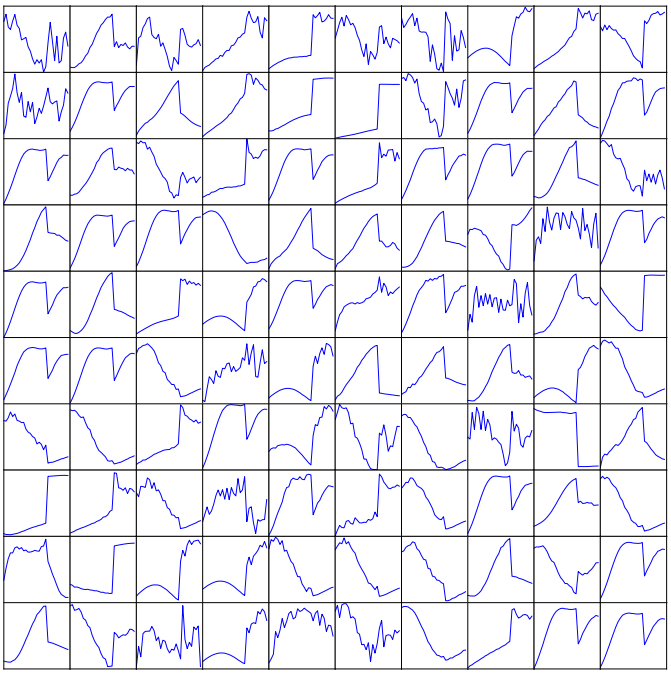

Fig. 5. Autoencoder first-layer weights, shown as 100 time-series

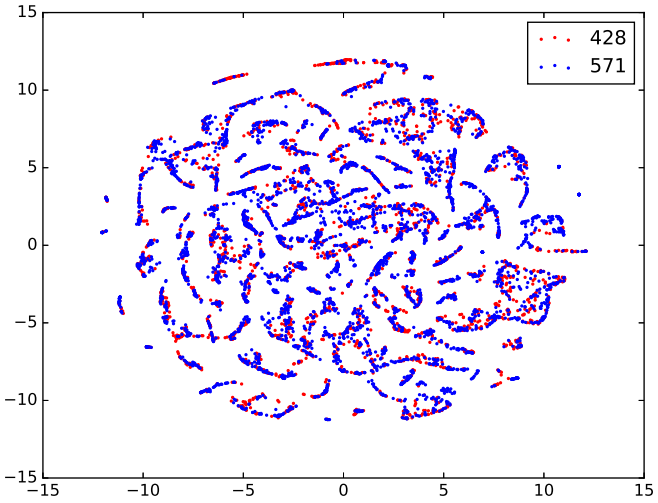

Fig. 6. t-SNE on first-layer learned features

The examples were restricted by the tool's use of ICD-9 categories, which may have been too broad to produce meaningful clustering. Generalizing this would be useful, as would other diagnostics to give clues into the feature learning process (such as plots of second-layer learned features).

The original goal of running as much as possible within Apache Spark on "standard" infrastructure such as Amazon EMR or Databricks was not fully met. Further integration with Apache Spark still is possible; the autoencoders perhaps could be implemented in DL4J (a native Java library supporting Apache Spark) or Spark's built-in `pyspark` support may allow the Keras and scikit-learn code to run directly on that infrastructure via `spark-submit`. The R language also has many relevant libraries, and SparkR may at some point permit their more seamless use.

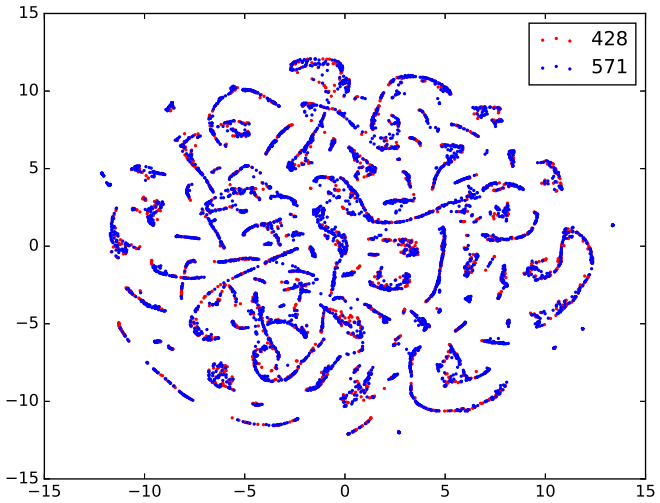

Fig. 7. t-SNE on second-layer learned features

Several optimizations also may help. hyperparameter optimization is currently done with a grid search, but would more sensibly done with a more intelligent optimization algorithm (such as SGD). The time warping function has parameters that could be tuned, or more extensive changes[8] could be made to try to make the time-series more stationary. Other covariance functions may be more appropriate as well.

Some other areas should perhaps be explored further too. One incremental change is in the use of multiple-task Gaussian processes (MTGPs); the work done here handles only individual time-series, while MIMIC-III is rich in parallel time-series that correlate with each other. Ghassemi *et al.*[4] explored the use of MTGPs to find a latent representation of multiple correlated time-series, but did not use this representation for subsequent feature learning. Another incremental change is in the use of Variational Autoencoders (VAEs) to learn a feature space that is sufficiently low-dimensional that techniques such as t-SNE are not required for effective visualization.

A more extensive change could involve using recurrent neural networks (RNNs). Deep networks such as RNNs such as Long Short-Term Memories (LSTMs) have shown promise in their ability to more directly handle sequences[16] and clinical time-series data, including handling missing data[10], [12], [11]. However, they are primarily used for supervised learning, but could potentially be treated similarly as autoencoders (as in [7]), that is, trained with the same input and output data in order to learn a reduced representation of the input. This approach would avoid some of the need to perform Gaussian Process Regression, however, it still may not cope well with time-series data that is very irregular.

Acknowledgment

Thank you to Dr. Jimeng Sun, Sungtae An, and the other TAs for their time and advice in this project.

References

[1] Z. Che, D. Kale, W. Li, M. Taha Bahadori, and Y. Liu. Deep Computational Phenotyping. *Proceedings of the 21st ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, pages 507–516, 2015.

[2] F. Chollet. Building Autoencoders in Keras. https://via.hypothes.is/https://blog.keras.io/building-autoencoders-in-keras.html.

[3] M. Ebden. Gaussian Processes: A Quick Introduction. (August), 2015.

[4] M. Ghassemi, T. Naumann, T. Brennan, D. a. Clifton, and P. Szolovits. A Multivariate Timeseries Modeling Approach to Severity of Illness Assessment and Forecasting in ICU with Sparse , Heterogeneous Clinical Data. *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 446–453, 2015.

[5] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. Clifton, and G. D. Clifford. Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, 104(2):444–466, 2016.

[6] A. E. W. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[7] M. Klapper-Rybicka, N. N. Schraudolph, and J. Schmidhuber. Unsupervised Learning in LSTM Recurrent Neural Networks. *Icann*, pages 674—-681, 2001.

[8] T. A. Lasko. Nonstationary Gaussian Process Regression for Evaluating Clinical Laboratory Test Sampling Strategies. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015(8):1777–1783, jan 2015.

[9] T. A. Lasko, J. C. Denny, and M. A. Levy. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLoS ONE*, 8(6), 2013.

[10] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell. Learning to Diagnose with LSTM Recurrent Neural Networks. *Iclr*, pages 1–18, 2015.

[11] Z. C. Lipton, D. C. Kale, and R. Wetzel. Directly Modeling Missing Data in Sequences with RNNs: Improved Classification of Clinical Time Series. 56(2016):1–17, 2016.

[12] Z. C. Lipton, D. C. Kale, and R. C. Wetzell. Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 705–714, 2015.

[13] B. M. Marlin, D. C. Kale, R. G. Khemani, and R. C. Wetzel. Unsupervised Pattern Discovery in Electronic Health Care Data Using Probabilistic Clustering Models.

[14] A. Ng, J. Ngiam, C. Y. Foo, Y. Mai, and C. Suen. UFLDL Tutorial. http://deeplearning.stanford.edu/wiki/index.php/UFLDL_Tutorial.

[15] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning.*, volume 14. 2004.

[16] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems (NIPS)*, pages 3104–3112, 2014.