MACHINE LEARNING FOR HEALTHCARE
6.S897, HST.S53

Lecture 7: Physiological and laboratory time-series

Prof. David Sontag
MIT EECS, CSAIL, IMES
Outline of today’s class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion
Labs and physiological time-series

• Typical use cases:
  1. Risk stratification, e.g. predict clinical deterioration, or diagnosis
  2. Infer patient’s past, current, or future health state from noisy observations, e.g. heart rate or glucose levels

• Approach taken varies depending on:
  • Is labeled data available?
  • Do we have a good mechanistic/statistical model?
  • How much training data is there?
Physiological time-series

Fig. 4. Probes used to collect vital signs data from an infant in intensive care. 1) Three-lead ECG, 2) arterial line (connected to blood pressure transducer), 3) pulse oximeter, 4) core temperature probe (underneath shoulder blades), 5) peripheral temperature probe, 6) transcutaneous probe.

(Quinn et al., TPAMI 2008)
Heart rate dynamics

Looking at examples of the heart rate model being applied as a Kalman filter to heart rate sequences are shown in Figure 5. The top panels show sequences of noisy heart rate observations, and the lower panel shows estimates of the high frequency and low frequency dynamics for a baby. Much of the time, infants in intensive care are in a stable condition. Because infants with a low gestational age are usually asleep and motionless, there tends to be low variability in their vital signs when in a stable condition. The central nervous system. Instead, the approach adopted here is to model the other observed channels. Our resulting joint model is univariate in each observation channel, so that interactions between a number of different sub-systems including the heart rate observations, which are generally the least stable and most difficult to model of the observed channels. We then go on to show how this approach is adapted to model the other observed channels. The dynamics can be formulated using autoregressive (AR) processes, such that an AR(2) signal varying around an ARIMA(1,1,0) baseline. An ARIMA model is a compelling choice for the baseline, because it has a block diagonal structure. The measurements are therefore generally taken to be made up of a baseline with low frequency components and with residual between the original sequences and the moving-averaged resulting signal is taken to be the low frequency baseline. The ARIMA(1,1,0) model has the form described in (Quinn et al., TPAMI 2008).
Confounded by interventions & measurement errors

(Blood pressure) Blood sample
Oxygen uptake

TcPO$_2$ (kPa)
TcPCO$_2$ (kPa)
Sys. BP
Dia. BP
BS
TR

Time (s)
0 1000 2000 3000 4000 5000
0 10 20 30 40 50

Drop out
Recalibration
Blood sample

(Quinn et al., TPAMI 2008)
Can we identify the artifactual processes?

• Once identified, can remove for use in downstream predictive tasks (must deal with missing data)

• Can help mitigate alarm fatigue by not alerting the clinicians when unnecessary

• More broadly, can we maintain beliefs about the true physiological values of a patient?
(Switching) linear dynamical systems

- Conditioned on $s_t$, linear Gaussian state-space models (Kalman filters):

$$x_t \sim \mathcal{N}\left( A^{(s_t)} x_{t-1} + d^{(s_t)}, Q^{(s_t)} \right)$$

$$y_t \sim \mathcal{N}\left( C^{(s_t)} x_t, R^{(s_t)} \right)$$
(Switching) linear dynamical systems

- Full model:
Learning SLDS models

- Assume some labeled training data \( \{s, y\} \)
- *True state* \( x \) assumed to never be observed
- Parameterization for \( x \) depends on states \( s \)
- Learn using expectation maximization
Outline of today’s class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion
Example of risk stratification: predicting morbidity in preterm newborns

Saria et al.,
Science Translational Medicine 2010
Measuring heart rate variability

(Saria et al., Science Translational Medicine 2010)
Learning algorithm / model

- Logistic regression used to predict whether baby will be “high morbidity” (HM):
  \[ P(HM|v_1,v_2,...,v_n) = \left(1 + \exp\left(b + w_0 * c + \sum_{i=1}^{n} w_i * f(v_i)\right)\right)^{-1} \]

- Features computed using 3 hours of data and nonlinear Bayesian model:
  - Estimated \( P(v_i | C) \) for each class of patient \( C={\text{HM or LM}} \) using parametric models such as exponential, Weibull, lognormal, gamma
  - Use log odds ratio of observed value as feature if observed, 0 otherwise:
    \[ \log \frac{P(v_i|HM)}{P(v_i|LM)} \]
  - Assumes data missing at random
Feature importance

- Mean heart rate
- Short-term variability of heart rate
- Long-term variability of heart rate
- Mean respiratory rate
- Short-term variability of respiratory rate
- Long-term variability of respiratory rate
- Mean oxygen saturation
- % of time spent below 85% oxygen saturation

Fig. 3. The significance of different physiological parameters in predicting high morbidity.

(A) The learned weight ($w_i$ in Eq. 1) for each physiological parameter incorporated in PhysiScore; error bars indicate variation in the weight over the different folds of the cross-validation.

(B) The nonlinear function associating the parameter with the risk of high versus low morbidity.
Outline of today’s class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. **RNNs with missing values (on MIMIC)**
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion
Modeling sequential data with neural networks

- Let \( x_t \in \mathbb{R}^d \) denote the patient’s data at time \( t \)
- By the chain rule, any distribution can be factorized as:
  \[
p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})
\]
- Train a neural net that composes history to predict next time step:
  \[
g(x_1, \ldots, x_{t-1}) \in \mathbb{R}^k : \text{composition function}
\]
  \[
w_i \in \mathbb{R}^k, \quad b^i \in \mathbb{R} : \text{parameters} i = 1, \ldots, d
\]
  \[
p(x_t^i = 1 \mid x_1, \ldots, x_{t-1}) = \frac{e^{w^i \cdot g(x_1, \ldots, x_{t-1}) + b^i}}{1 + e^{w^i \cdot g(x_1, \ldots, x_{t-1}) + b^i}}
  = \text{logistic}(w^i \cdot g(x_1, \ldots, x_{t-1}) + b^i)
\]
Modeling sequential data with neural networks

\[ p(x_t^1 | x_1, \ldots, x_{t-1}) = \text{logistic}(w^1 \cdot h_{t-1} + b^1) \]

\[ h_{t-1} = g \left( \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{bmatrix} \right) \]
Modeling sequential data with neural networks

\[
p(x_t^2 | x_1, \ldots, x_{t-1}) = \text{logistic}(w^2 \cdot h_{t-1} + b^2)
\]

\[
h_{t-1} = g \left( \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{bmatrix} \right)
\]
Modeling sequential data with neural networks

\[ p(x_t^d | x_1, \ldots, x_{t-1}) = \text{logistic}(w^d \cdot h_{t-1} + b^d) \]

\[ h_{t-1} = g \left( \begin{bmatrix} x_{t-1} \\ x_{t-2} \\ x_{t-3} \end{bmatrix} \right) \]
Recurrent neural networks (RNNs)

Maintain a hidden state vector $h_t$ that is recursively calculated

$$p(x_t^i | x_1, \ldots, x_{t-1}) = \text{logistic}(w^i \cdot h_{t-1} + b^i)$$

RNN language models widely used in natural language processing: state-of-the-art performance for speech recognition and machine translation
Recurrent neural networks (RNNs)

Maintain a hidden state vector $h_t$ that is recursively calculated

$$p(x_t^i | x_1, \ldots, x_{t-1}) = \text{logistic}(w^i \cdot h_{t-1} + b^i)$$

Significant interest in using RNNs for disease progression modeling:

RNNs versus HMMs

Equivalently, viewing the RNN as a Markov model:

\[ h_3 = g \left([h_2 \mid x_3]\right) \]

- **Advantage:** Very powerful
- **Disadvantages:**
  - Not easy to deal with missing data in \( x \)
  - No ability to discover structure in \( x \) – can overfit if \( d \) is large
  - All randomness due to exogenous factors must be captured in \( x \) observations
  - Difficult (not impossible) to incorporate prior knowledge and to combine as part of a more complex model
RNNs versus HMMs

Equivalently, viewing the RNN as a Markov model:

Can’t remove the edges from $x$ to $h$ in this model, because it becomes useless (due to $h$ transitions being deterministic):

$$p(x_t \mid x_1, \ldots, x_{t-1}) = p(x_t)$$
Timing matters! Measurement motifs

(Pivovarov et al., JBI 2014)
How can we exploit missingness in RNNs?

Missingness comes from various reasons.

Represent and Utilize Missing Values

Two representations of missingness:

- **Masking $M$:**
  
  Whether a variable is missing or not.

- **Time Interval $\Delta$:**
  
  How long a variable has been missing.

**Decay Term $\gamma$:** A flexible transformation on $\Delta$ jointly learned with deep model.

$$
\gamma_t = \exp\{-ReLU(W\gamma \delta_t + b_\gamma)\}
$$

$$
\tilde{x}_t^d \leftarrow m_t^d x_t^d + (1 - m_t^d)\gamma x_t^d x_t^d + (1 - m_t^d)(1 - \gamma x_t^d)\tilde{x}^d
$$

GRU-D model

- Decay on the last observations.
- Decay on the hidden states.

Quantitative Evaluation

Evaluations on synthetic dataset with different missing rates

- GRU-mean
- GRU-forward
- GRU-simple
- GRU-D

Evaluations for mortality early prediction

- GRU-mean
- GRU-forward
- GRU-simple
- GRU-D

AUC score on mortality prediction

<table>
<thead>
<tr>
<th>Models</th>
<th>MIMIC-III</th>
<th>PhysioNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR-forward</td>
<td>0.7589</td>
<td>0.7423</td>
</tr>
<tr>
<td>SVM-forward</td>
<td>0.7908</td>
<td>0.8131</td>
</tr>
<tr>
<td>RF-forward</td>
<td>0.8293</td>
<td>0.8183</td>
</tr>
<tr>
<td>LR-simple</td>
<td>0.7715</td>
<td>0.7625</td>
</tr>
<tr>
<td>SVM-simple</td>
<td>0.8146</td>
<td>0.8277</td>
</tr>
<tr>
<td>RF-simple</td>
<td>0.8294</td>
<td>0.8157</td>
</tr>
<tr>
<td>LSTM-mean</td>
<td>0.8142</td>
<td>0.8025</td>
</tr>
<tr>
<td>GRU-mean</td>
<td>0.8192</td>
<td>0.8195</td>
</tr>
<tr>
<td>GRU-forward</td>
<td>0.8252</td>
<td>0.8162</td>
</tr>
<tr>
<td>GRU-simple</td>
<td>0.8380</td>
<td>0.8155</td>
</tr>
<tr>
<td>Ours</td>
<td>GRU-D</td>
<td>0.8527</td>
</tr>
</tbody>
</table>
Qualitative Evaluation

Input decay plots of all 33 variables for mortality prediction on PhysioNet dataset

- Get a few important variables, e.g., weight, arterial pH, temperature, and respiration rate, etc.

Histograms of hidden state decay for mortality prediction on PhysioNet dataset

- Parameters related to variables with smaller missing rate are more spread out.
Outline of today’s class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. **CNNs for predicting disease onsets from longitudinal lab tests**
5. Project discussion
Multi-task prediction of disease onsets from longitudinal lab tests

Goal:
- Early diagnosis of diseases for people who do not already have the disease.
- Going toward raw biological signals (i.e. lab measurements) and learning rich representations directly from the raw input

Framework: Multi-task Supervised Prediction

Input Biomarkers over time

Output Disease onsets over time

(Razavian et al., “Multi-task Prediction of Disease Onsets from Longitudinal Laboratory Tests”. 1st Conference on Machine Learning and Health Care, 2016)
Cohort

298,000 individuals, with at least once a year lab measurement for 3 consecutive years included

- **Input**: Comprehensive lab panel + cholesterol (18 labs)
- **Output**: 133 conditions.

- **Exclusion Per Disease**: Anyone with even 1 measurement prior to start of prediction window
  - Done via masking the gradients in SGD process for excluded patients per task.

- Randomly Split to train(100K), validate(100K) and test(98k) set
CNN-1: Convolution over Labs then Time

Input labs

Vertical Convolution (+ReLU+batchnorm) (Kernel sizes: |Labs| x 1)

Vertical Convolution (+ReLU+batchnorm) (Kernel sizes: |previous layer filters| x 1)

Lab Combination Subnetwork: Vertical convolution to combine labs

Temporal Subnetwork: Temporal pooling and temporal convolution

2 Layers of Dropout + Fully connected + ReLU + batchnorm + Log Softmax

P(Y_1=1|input)
P(Y_3=1|input)
P(Y_M=1|input)
CNN-2: Multi-resolution Convolution over Time

Temporal convolution in 3 resolutions.

2 Layers of Dropout + Fully connected + ReLU

batchnorm + Log Softmax

P(Y_M=1|input)

P(Y_3=1|input)

P(Y_1=1|input)
LSTM for Sequence Embedding

<table>
<thead>
<tr>
<th>Lab name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creatinine</td>
</tr>
<tr>
<td>Urea nitrogen</td>
</tr>
<tr>
<td>Potassium</td>
</tr>
<tr>
<td>Glucose</td>
</tr>
<tr>
<td>Alanine aminotransferase</td>
</tr>
<tr>
<td>Aspartate aminotransferase</td>
</tr>
<tr>
<td>Protein</td>
</tr>
<tr>
<td>Albumin</td>
</tr>
<tr>
<td>Cholesterol</td>
</tr>
<tr>
<td>Triglyceride</td>
</tr>
<tr>
<td>Cholesterol.in LDL</td>
</tr>
<tr>
<td>Calcium</td>
</tr>
<tr>
<td>Sodium</td>
</tr>
<tr>
<td>Chloride</td>
</tr>
<tr>
<td>Carbon dioxide</td>
</tr>
<tr>
<td>Urea nitrogen/Creatinine</td>
</tr>
<tr>
<td>Bilirubin</td>
</tr>
<tr>
<td>Albumin/Globulin</td>
</tr>
</tbody>
</table>

Long Short Term Memory Recurrent Units

batchnorm + Log Softmax

- $P(Y_1=1|\text{input})$
- $P(Y_3=1|\text{input})$
- $P(Y_M=1|\text{input})$

2 Layers of Dropout + Fully connected + ReLU

Connected to the last LSTM memory unit
## Results

**Goal:** predict new onset of diseases 3 months in advance

<table>
<thead>
<tr>
<th>ICD9 Code and disease description</th>
<th>LR</th>
<th>LSTM</th>
<th>CNN1</th>
<th>CNN2</th>
<th>Ens</th>
<th>Pos</th>
</tr>
</thead>
<tbody>
<tr>
<td>585.6 End stage renal disease</td>
<td>0.886</td>
<td>0.917</td>
<td>0.910</td>
<td>0.916</td>
<td>0.920</td>
<td>837</td>
</tr>
<tr>
<td>285.21 Anemia in chr kidney dis</td>
<td>0.849</td>
<td>0.866</td>
<td>0.868</td>
<td>0.880</td>
<td>0.879</td>
<td>1598</td>
</tr>
<tr>
<td>585.3 Chr kidney dis stage III</td>
<td>0.846</td>
<td>0.851</td>
<td>0.857</td>
<td>0.858</td>
<td>0.864</td>
<td>2685</td>
</tr>
<tr>
<td>584.9 Acute kidney failure NOS</td>
<td>0.805</td>
<td>0.820</td>
<td>0.828</td>
<td>0.831</td>
<td>0.835</td>
<td>3039</td>
</tr>
<tr>
<td>250.01 DMI wo cmp nt st uncntrl</td>
<td>0.822</td>
<td>0.813</td>
<td>0.819</td>
<td>0.825</td>
<td>0.829</td>
<td>1522</td>
</tr>
<tr>
<td>250.02 DMII wo cmp uncntrld</td>
<td>0.814</td>
<td>0.819</td>
<td>0.814</td>
<td>0.821</td>
<td>0.828</td>
<td>3519</td>
</tr>
<tr>
<td>593.9 Renal and ureteral dis NOS</td>
<td>0.757</td>
<td>0.794</td>
<td>0.784</td>
<td>0.792</td>
<td>0.798</td>
<td>2111</td>
</tr>
<tr>
<td>428.0 CHF NOS</td>
<td>0.739</td>
<td>0.784</td>
<td>0.786</td>
<td>0.783</td>
<td>0.792</td>
<td>3479</td>
</tr>
<tr>
<td>V053 Need prphyl vc vrl hepat</td>
<td>0.731</td>
<td>0.762</td>
<td>0.752</td>
<td>0.780</td>
<td>0.777</td>
<td>862</td>
</tr>
<tr>
<td>790.93 Elvtd prstate spcf antgn</td>
<td>0.666</td>
<td>0.758</td>
<td>0.761</td>
<td>0.768</td>
<td>0.772</td>
<td>1477</td>
</tr>
<tr>
<td>185 Malign neopl prostate</td>
<td>0.627</td>
<td>0.757</td>
<td>0.751</td>
<td>0.761</td>
<td>0.768</td>
<td>761</td>
</tr>
<tr>
<td>274.9 Gout NOS</td>
<td>0.746</td>
<td>0.761</td>
<td>0.764</td>
<td>0.757</td>
<td>0.767</td>
<td>1529</td>
</tr>
<tr>
<td>362.52 Exudative macular degen</td>
<td>0.687</td>
<td>0.752</td>
<td>0.750</td>
<td>0.757</td>
<td>0.765</td>
<td>538</td>
</tr>
</tbody>
</table>

**AUC** sorted by maximum AUC achieved by any of the models
Observations

- Rich representation learning improves prediction quality of weaker tasks in the multi-task settings

- Most gains are on tasks where the predictive features are NOT directly included in the input already
  - Confirmed by the case study of chronic kidney disease progression, and our most-improved outcomes

- Different representation learning methods (CNN1, CNN2, LSTM) show similar improvements.

- Ensemble of best models always further improves results

https://github.com/clinicalml/deepDiagnosis
A Model for Imputation On Correlated Biomarkers

• Imputation model based on structured multivariate kernel regression/smoothing\(^1\)

• Formulated as unsupervised learning method

\[ E_{x \sim P(x|t=t_{new})}[x] = \frac{(K * \bar{X}_{\text{train}})(t_{new})}{(K * I(\bar{X}_{\text{train} : \text{obs}}))(t_{new})} \]

Data: 30K Individuals from the original training set.
Dataset split equally between train, test and validate set.
Loss: MSE. Train and evaluate only on (lab, person) with more than 1 observation.
Outline of today’s class

1. State space models for physiological condition modeling
2. Physiological assessment score for preterm infants
3. RNNs with missing values (on MIMIC)
4. CNNs for predicting disease onsets from longitudinal lab tests
5. Project discussion
#5: Disease progression in multiple myeloma

Clinical mentor: Nikhil Munshi, MD
Dana-Farber Cancer Institute
Professor of Medicine, Harvard Medical School

- Blood cancer, affecting 0.8% of US population at some point in their lifetime. 5 year survival rate is 49%
- Major advances in treatment, with 10+ new drugs on the market, more in clinical trials
- Project goal: predict patient survival and time to disease progression
- Data for ~1000 individuals:
  - Cytogenetics, mutations, gene expression
  - Biomarker levels across time (eg immunoglobulin levels)
  - Clinical outcomes including disease status, time to response, treatment response
  - Adverse events (eg anemia, bone pain, renal failure…)
  - Quality of life measures (e.g. appetite loss, fatigue) and symptoms
  - Treatment therapies including combination treatments
#5: Disease progression in multiple myeloma

Clinical mentor: Nikhil Munshi, MD
Dana-Farber Cancer Institute
Professor of Medicine, Harvard Medical School

![Graph showing disease progression and treatments over time with lab values and responses marked.](image-url)
Data for multiple myeloma project

Simple form –
Takes just a few minutes to request the data, and no training needed
#6: Machine learning on medical images

**Clinical mentor:** Quanzheng Li, Ph.D.
Massachusetts General Hospital, Department of Radiology
Center for Clinical Data Science
Associate Professor, Harvard Medical School

- Led one of the top teams in Camelyon 2016 competition on cancer metastasis detection (pathology)
- Could use publicly available data and propose your own project in consultation with him

- One project he proposed:
  Study transfer learning using chest CT images from patients in two cohorts, emphysema and lung cancer