Machine Learning for Healthcare
HST.956, 6.S897

Lecture 5: Risk stratification (continued)

David Sontag
Course announcements

• Recitation Friday at 2pm (4-153) – optional
• PS1 due tonight; PS2 out Tuesday
Outline for today’s class

1. Risk stratification (continued)
   - Deriving labels
   - Evaluation
   - Subtleties with ML-based risk stratification

2. Survival modeling
Where do the labels come from?

Typical pipeline:
1. Manually label several patients’ data by “chart review”
2. A) Come up with a simple rule to automatically derive label for all patients, or
   B) Use machine learning to get the labels themselves
Step 1: Visualization of individual patient data is an important part of chart review.

Demographic information

Patient events list

Events, as they occur for the first time in patient history

https://github.com/nyuvis/patient-viz
Step 2: Example of a rule-based phenotype

Source: https://phekb.org/sites/phenotype/files/T2DM-algorithm.pdf
Step 2: Example of a rule-based phenotype
Outline for today’s class

1. Risk stratification (continued)
   - Deriving labels
   - Evaluation
   - Subtleties with ML-based risk stratification

2. Survival modeling
Receiver-operator characteristic curve

Want to be here

Obtained by varying prediction threshold

Full model
Traditional risk factors

Diabetes 1-year gap

True positive rate
False positive rate
Receiver-operator characteristic curve

Area under the ROC curve (AUC)

AUC = Probability that algorithm ranks a positive patient over a negative patient

Invariant to amount of class imbalance

False positive rate

Diabetes 1-year gap
Receiver-operator characteristic curve

Risk stratification usually focuses on just this region (because of the cost of interventions)

Diabetes 1-year gap

True positive rate

False positive rate

- Full model $\text{AUC}=0.78$
- Traditional risk factors $\text{AUC} = 0.74$
- Random $\text{AUC} = 0.5$
Calibration (note: different dataset)

Model Predicting infection in the ER

Actual Probability

Predicted Probability

fraction of patients the model predicts to have this probability of infection
Outline for today’s class

1. Risk stratification (continued)
   - Deriving labels
   - Evaluation
   - Subtleties with ML-based risk stratification

2. Survival modeling
Non-stationarity: 
*Diabetes Onset After 2009*

→ Automatically derived labels may change meaning

Non-stationarity:  
*Top 100 lab measurements over time*

![Heatmap showing non-stationarity in lab measurements over time.](image)

Time (in months, from 1/2005 up to 1/2014)

→ Significance of features may change over time

[Figure credit: Narges Razavian]
Non-stationarity: 
*ICD-9 to ICD-10 shift*

Significance of features may change over time

[Figure credit: Mike Oberst]
Re-thinking evaluation in the face of non-stationarity

- How was our diabetes model evaluation flawed?
- Good practice: use test data from a future year:

![Figure 5-1: Partitioning of data into training/development and test sets, based on an 80-20 split of patient IDs and time intervals of 2007-2013 and 2014-2016.](Figure credit: Helen Zhou)
Intervention-tainted outcomes

- Example from today’s readings:
  - Patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia
  - Thus, we learn: \( \text{HasAsthma}(x) \implies \text{LowerRisk}(x) \)

- What’s wrong with the learned model?
  - Risk stratification drives interventions
  - If low risk, might not admit to ICU. But this was precisely what prevented patients from dying!

[Caruana et al., Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. KDD 2015.]
Intervention-tainted outcomes

• Formally, this is what’s happening:

“A long survival time may be because of treatment!

• How do we address this problem?
• First and foremost, must recognize it is happening – interpretable models help with this
Intervention-tainted outcomes

- Hacks:
  1. Modify model, e.g. by removing the `HasAsthma(x) => LowerRisk(x)` rule
     I do not expect this to work with high-dimensional data
  2. Re-define outcome by finding a pre-treatment surrogate (e.g., lactate levels)
  3. Consider treated patients as *right-censored* by treatment

**Example:**
Intervention-tainted outcomes

- The rigorous way to address this problem is through the language of **causality**:

  \[
  \text{Patient, } X \quad \rightarrow \text{Intervention, } T \quad \rightarrow \text{Outcome, } Y \text{ (death)}
  \]

  (admit to the ICU?)

  (everything we know at triage)

  Will admission to ICU lower likelihood of death for patient?

- We return to this in Lecture 14
No big wins from deep models on structured data/text

Rajkomar et al., Scalable and accurate deep learning with electronic health records. *Nature Digital Medicine*, 2018

Recurrent neural network & attention-based models trained on 200K hospitalized patients
Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

<table>
<thead>
<tr>
<th>Inpatient Mortality, AUROC(^1) (95% CI)</th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning 24 hours after admission</td>
<td>0.95 (0.94-0.96)</td>
<td>0.93 (0.92-0.94)</td>
</tr>
<tr>
<td>Full feature enhanced baseline at 24 hours after admission</td>
<td>0.93 (0.92-0.95)</td>
<td>0.91 (0.89-0.92)</td>
</tr>
<tr>
<td>Full feature simple baseline at 24 hours after admission</td>
<td>0.93 (0.91-0.94)</td>
<td>0.90 (0.88-0.92)</td>
</tr>
<tr>
<td>Baseline (aEWS(^2)) at 24 hours after admission</td>
<td>0.85 (0.81-0.89)</td>
<td>0.86 (0.83-0.88)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>30-day Readmission, AUROC (95% CI)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning at discharge</td>
<td>0.77 (0.75-0.78)</td>
<td>0.76 (0.75-0.77)</td>
</tr>
<tr>
<td>Full feature enhanced baseline at discharge</td>
<td>0.75 (0.73-0.76)</td>
<td>0.75 (0.74-0.76)</td>
</tr>
<tr>
<td>Full feature simple baseline at discharge</td>
<td>0.74 (0.73-0.76)</td>
<td>0.73 (0.72-0.74)</td>
</tr>
<tr>
<td>Baseline (mHOSPITAL(^3)) at discharge</td>
<td>0.70 (0.68-0.72)</td>
<td>0.68 (0.67-0.69)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Length of Stay at least 7 days AUROC (95% CI)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning 24 hours after admission</td>
<td>0.86 (0.86-0.87)</td>
<td>0.85 (0.85-0.86)</td>
</tr>
<tr>
<td>Full feature enhanced baseline at 24 hours after admission</td>
<td>0.85 (0.84-0.85)</td>
<td>0.83 (0.83-0.84)</td>
</tr>
<tr>
<td>Full feature simple baseline at 24 hours after admission</td>
<td>0.83 (0.82-0.84)</td>
<td>0.81 (0.80-0.82)</td>
</tr>
<tr>
<td>Baseline (mLiu(^4)) at 24 hours after admission</td>
<td>0.76 (0.75-0.77)</td>
<td>0.74 (0.73-0.75)</td>
</tr>
</tbody>
</table>

Comparison to Razavian et al. ‘15

\[\text{Rajkomar et al. ‘18 electronic supplementary material:} \]
[https://static-content.springer.com/esm/art\%3A10.1038\%2Fs41746-018-0029-1/MediaObjects/41746_2018_29_MOESM1_ESM.pdf]
No big wins from deep models on structured data/text

Supplemental Table 1: Prediction accuracy of each task of deep learning model compared to baselines

<table>
<thead>
<tr>
<th>Inpatient Mortality, AUROC(^1) (95% CI)</th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning 24 hours after admission</td>
<td>0.95 (0.94-0.96)</td>
<td>0.93 (0.92-0.94)</td>
</tr>
<tr>
<td>Full feature enhanced baseline at 24 hours after admission</td>
<td>0.93 (0.92-0.95)</td>
<td>0.91 (0.89-0.92)</td>
</tr>
<tr>
<td>Full feature simple baseline at 24 hours after admission</td>
<td>0.93 (0.91-0.94)</td>
<td>0.90 (0.88-0.92)</td>
</tr>
<tr>
<td>Baseline (aEWS(^2)) at 24 hours after admission</td>
<td>0.85 (0.81-0.89)</td>
<td>0.86 (0.83-0.88)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>30-day Readmission, AUROC (95% CI)</th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning at discharge</td>
<td>0.77 (0.75-0.78)</td>
<td>0.76 (0.75-0.77)</td>
</tr>
<tr>
<td>Full feature enhanced baseline at discharge</td>
<td>0.75 (0.73-0.76)</td>
<td>0.75 (0.74-0.76)</td>
</tr>
<tr>
<td>Full feature simple baseline at discharge</td>
<td>0.74 (0.73-0.76)</td>
<td>0.73 (0.72-0.74)</td>
</tr>
<tr>
<td>Baseline (mHOSPITAL(^3)) at discharge</td>
<td>0.70 (0.68-0.72)</td>
<td>0.68 (0.67-0.69)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Length of Stay at least 7 days, AUROC (95% CI)</th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep learning 24 hours after admission</td>
<td>0.86 (0.86-0.87)</td>
<td>0.85 (0.85-0.86)</td>
</tr>
<tr>
<td>Full feature enhanced baseline at 24 hours after admission</td>
<td>0.85 (0.84-0.85)</td>
<td>0.83 (0.83-0.84)</td>
</tr>
<tr>
<td>Full feature simple baseline at 24 hours after admission</td>
<td>0.83 (0.82-0.84)</td>
<td>0.81 (0.80-0.82)</td>
</tr>
<tr>
<td>Baseline (mLiu(^4)) at 24 hours after admission</td>
<td>0.76 (0.75-0.77)</td>
<td>0.74 (0.73-0.75)</td>
</tr>
</tbody>
</table>

Comparison to Razavian et al. ‘15

[Keep in mind: Small wins with deep models may disappear altogether with dataset shift or non-stationarity (Jung & Shah, JBI ‘15)]

No big wins from deep models on structured data/text – why?

• Sequential data in medicine is very different from language modeling
  – Many time scales, significant missing data, and multi-variate observations
  – Likely *do exist* predictive nonlinear interactions, but subtle
  – Not enough data to naively deal with the above two

• Medical community has already come up with some very good features
1. Risk stratification (continued)
   - Deriving labels
   - Evaluation
   - Subtleties with ML-based risk stratification

2. Survival modeling
Survival modeling

- We focus on right-censored data:

![Diagram showing survival analysis](image)

Event occurrence e.g., death, divorce, college graduation

Censoring

Survival modeling

• Why not use classification, as before?
  – Less data for training (due to exclusions)
  – Pessimistic estimates due to choice of window

• What about regression, e.g. minimizing mean-squared error?
  – T is non-negative, may want long tails
  – If we just naively removed censored events, we would be introducing bias
Notation and formalization

• Data are \((x, T, b)\) = (features, time, censoring), where \(b = 0, 1\) denotes whether time is of censoring or event occurrence.

• Let \(f(t) = P(t)\) be the probability of death at time \(t\).

• Survival function: the probability of an individual surviving beyond time \(t\),

\[
S(t) = P(T > t) = \int_{t}^{\infty} f(x)dx.
\]

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]
The survival function is represented by $S(t)$, which is given as follows:

$$S(t) = \Pr(T \leq t).$$

The survival function monotonically decreases with $t$, and the initial value is 1 when $t = 0$, which represents the fact that, in the beginning of the observation, 100% of the observed subjects survive; in other words, none of the events of interest have occurred.

On the contrary, the cumulative death distribution function $F(t)$, which represents the probability that the event of interest occurs earlier than $t$, is defined as

$$F(t) = 1 - S(t),$$

and death density function can be obtained as $f(t) = \frac{d}{dt}F(t)$ for continuous cases, and $f(t) = \left\{F(t + t) - F(t)\right\}/t$, where $t$ denotes a small time interval, for discrete cases.

Figure 2 shows the relationship among these functions.

In survival analysis, another commonly used function is the hazard function $h(t)$, which is also called the force of mortality, the instantaneous death rate or the conditional failure rate [Dunn and Clark 2009]. The hazard function does not indicate the chance or probability of the event of interest, but instead it is the rate of event at time $t$ given that no event occurred before time $t$. Mathematically, the hazard function is defined as:

$$h(t) = \lim_{t \to 0} \frac{\Pr(t \leq T < t + t | T \geq t)}{t} = \lim_{t \to 0} \frac{F(t + t) - F(t)}{t} \cdot \frac{1}{S(t)} = f(t) \cdot \frac{1}{S(t)}.$$

Thus, the survival function defined in Eq. (2) can be rewritten as

$$S(t) = \exp(-H(t)),$$

where $H(t)$ is the cumulative hazard function.

---

Fig. 2: Relationship among different entities $f(t), F(t)$ and $S(t)$.

Kaplan-Meier estimator

- Example of a non-parametric method; good for unconditional density estimation

\[
\hat{S}_{K-M}(t) = \prod_{k:y(k) \leq t} \left\{ 1 - \frac{d(k)}{n(k)} \right\}
\]

Observed event times
\[
y(1) < y(2) < \cdots < y(D)
\]

\(d(k) = \#\) events at this time
\(n(k) = \#\) of individuals alive and uncensored

Survival probability, \(S(t)\)

[Figure credit: Rebecca Peyser]
Maximum likelihood estimation

- Commonly parametric densities for \( f(t) \):

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Survival function ( S(t) )</th>
<th>Density function ( f(t) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential ((\lambda &gt; 0))</td>
<td>( \exp(-\lambda t) )</td>
<td>( \lambda \exp(-\lambda t) )</td>
</tr>
<tr>
<td>Weibull ((\lambda, \phi &gt; 0))</td>
<td>( \exp(-\lambda t^\phi) )</td>
<td>( \lambda t^{\phi-1} \exp(-\lambda t^\phi) )</td>
</tr>
<tr>
<td>Log-normal ((\sigma &gt; 0, \mu \in \mathbb{R}))</td>
<td>( 1 - \Phi((\ln t - \mu)/\sigma) )</td>
<td>( \varphi((\ln t - \mu)/\sigma)(\sigma)^{-1} )</td>
</tr>
<tr>
<td>Log-logistic ((\lambda &gt; 0, \phi &gt; 0))</td>
<td>( 1/(1 + \lambda t^\phi) )</td>
<td>( (\lambda t^{\phi-1})/(1 + \lambda t^\phi)^2 )</td>
</tr>
<tr>
<td>Gamma ((\lambda, \phi &gt; 0))</td>
<td>( 1 - I(\lambda t, \phi) )</td>
<td>( {\lambda^\phi / \Gamma(\phi)}t^{\phi-1} \exp(-\lambda t) )</td>
</tr>
<tr>
<td>Gompertz ((\lambda, \phi &gt; 0))</td>
<td>( \exp\left{\frac{\lambda}{\phi} (1 - e^{\phi t})\right} )</td>
<td>( \lambda e^{\phi t} \exp\left{\frac{\lambda}{\phi} (1 - e^{\phi t})\right} )</td>
</tr>
</tbody>
</table>

[Ha, Jeong, Lee. Statistical Modeling of Survival Data with Random Effects. Springer 2017]
Maximum likelihood estimation

• Two kinds of observations: censored and uncensored

  Uncensored likelihood

  \[ p_\theta(T = t \mid x) = f(t) \]

  Censored likelihood

  \[ p_\theta^{\text{censored}}(t \mid x) = p_\theta(T > t \mid x) = S(t) \]

• Putting the two together, we get:

  \[
  \sum_{i=1}^{n} b_i \log p_\theta^{\text{censored}}(t \mid x) + (1 - b_i) \log p_\theta(t \mid x)
  \]

  Optimize via gradient or stochastic gradient ascent!
Evaluation for survival modeling

- Concordance-index (also called C-statistic): look at model’s ability to predict relative survival times:

\[
\hat{c} = \frac{1}{\text{num}} \sum_{i:b_i = 0} \sum_{j:y_i < y_j} I[S(\hat{y}_j | X_j) > S(\hat{y}_i | X_i)]
\]

- Illustration – blue lines denote pairwise comparisons:

Black = uncensored
Red = censored

- Equivalent to AUC for binary variables and no censoring

Final thoughts on survival modeling

• Could also evaluate:
  – Mean-squared error for uncensored individuals
  – Held-out (censored) likelihood
  – Derive binary classifier from learned model and check calibration

• Partial likelihood estimators (e.g. for cox-proportional hazards models) can be much more data efficient