6.S897/HST.956 Machine Learning for Healthcare

Lecture 18: Disease Progression and Subtyping Part 1

Instructors: David Sontag, Peter Szolovits

1 Disease Progression Modelling

There are three questions we hope to answer when modelling disease progression:

1. Where is the patient in the disease trajectory? How long is the patient likely to live?
2. When will the disease progress?
3. How will treatment affect disease progression?

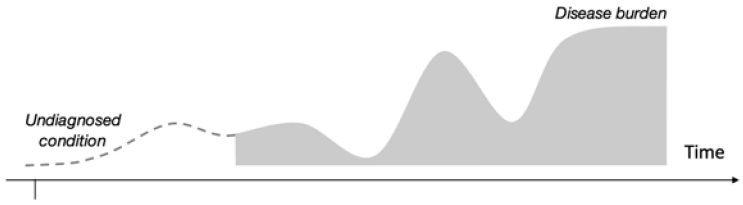

Figure 1: A graph showing disease progression for a patient. On the x-axis is time. On the y-axis is "disease burden", which can be measured in the amount of symptoms, the amount of pain medication, etc. With diseases such as cancer, disease burden may decrease initially, but grow with time. Second line treatments often result in a trough.

2 Staging

When a patient first checks into a hospital, the hospital will try to determine the stage of disease progress to determine treatment. Stages of a disease are characterized by various markers, which are detailed by a staging system. For example, in the Durie-Salmon Staging System, stage 1 of Multiple Myeloma is characterized by low M-component production rates, while stage 3 is be characterized by high M-component production rates (see Figure 2).

We might be interested in how to stage a specific patient. Or, we might be interested in describing the typical trajectory of a disease. This is especially important, because patients will use such decisions to inform their life decisions. Determining average trajectory is difficult for diseases which are rare, and for which little data is available.

Myeloma Staging Systems

Stage	Durie-Salmon Staging System	Revised International Staging System
I	All of the following: ○ Hemoglobin >10.5 g/dL ○ Serum calcium value normal or ≤12 mg/dL ○ X-ray studies of bone, normal bone structure (scale 0) or solitary bone plasmacytoma only ○ Low M-component production rate IgG value <5 g/dL; IgA value <3 g/dL ○ Urine light chains <4g/24 hours	○ Serum albumin >3.5 g/dL ○ Serum β_2-microglobulin <3.5 mg/L ○ No high-risk cytogenetics ○ Normal serum lactate dehydrogenase level
II	Neither stage I nor stage III ○ A—No renal failure (creatinine ≤2 mg/dL) ○ B—Renal failure (creatinine >2 mg/dL)	Neither stage I nor stage III
III	○ Hemoglobin value <8.5 g/dL ○ Serum calcium value >12 mg/dL ○ X-ray studies of bone, >3 lytic bone lesions ○ High M-component production rate IgG value >7 g/dL; IgA value >5 g/dL ○ Urine light chains >12 g/24 hours	○ Serum β_2-microglobulin >5.5 mg/L ○ High-risk cytogenetics t(4;14) t(14;16) del(17p) ○ Elevated serum lactate dehydrogenase level

Figure 2: Two different staging systems for Multiple Myeloma.

3 Subtyping

A different type of understanding involves subtyping disease diagnoses. That is, we want to understand the heterogeneity underlying different diagnoses. Perhaps, for a given disease, some subtypes might progress more quickly than others. Some subtypes may respond differently to a given treatment. Subtyping can even be thought of as re-defining the disease altogether.

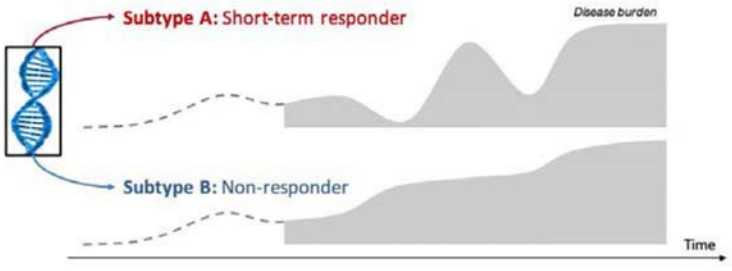

Figure 3: A graph showing how progression for the same disease may differ depending on subtype.

4 Multi-task Learning

Prognosis is a supervised machine learning problem. We would like to answer the question: given patient data $\vec{x}$ at time $t = 0$, what will the patient's status be at future points in time?

Let us consider a simple example, inspired by the work in [ZLNY12]. Suppose that we want to predict a binary variable at 6, 12, 24, 36, and 48 months given $\vec{x}$. This binary variable will represent the disease status of Alzheimer's. Disease status can be measured according to a variety of metrics including: the results of a simple cognition test, brain imaging, data collected on a mobile phone, etc.

The simplest approach to solve this problem is to build a model that simultaneously learns five different tasks (one task per time point) while ensuring that the parameters of the model make use of features that share common importance between tasks. This approach is useful in tackling the problem of data sparsity. Often, there is imbalance in the data, noise in the characterization of disease status, and censoring of data from later time points. Multi-task learning encourages the model to select a common set of bio-markers that are useful for prediction across all time points.

First, let us consider trying to predict a binary variable at 6 months and 12 months only. Then, our objective function will seek to minimize error:

$$\min_{\vec{w}_6, \vec{w}_{12}} \sum_i \left(y_i^6 - \vec{w}_6 \cdot \vec{x}\right) + \lambda_1 ||\vec{w}_6||^2 + \sum_i \left(y_i^{12} - \vec{w}_{12} \cdot \vec{x}\right) + \lambda_2 ||\vec{w}_{12}||^2$$

This objective function will perform two classification tasks: one at 6 months and one at 12 months. To tie the parameters $\vec{w}_6$ and $\vec{w}_{12}$ together, we can add a term to penalize the distance between $\vec{w}_6$ and $\vec{w}_{12}$. Then, our new objective function will be:

$$\min_{\vec{w}_6, \vec{w}_{12}} \sum_i \left(y_i^6 - \vec{w}_6 \cdot \vec{x}\right) + \lambda_1 ||\vec{w}_6||^2 + \sum_i \left(y_i^{12} - \vec{w}_{12} \cdot \vec{x}\right) + \lambda_2 ||\vec{w}_{12}||^2 + \lambda_2 ||\vec{w}_{12} - \vec{w}_6||_2 \tag{1}$$

This ensures that the parameters for the two tasks remain "close". Note that the $\lambda_2 ||\vec{w}_{12} - \vec{w}_6||_2$ term may be modified so that only a portion of the entries in $\vec{w}$ are constrained.

For multi-task with more than 2 tasks, we will need to generalize this notion of "closeness". Then, the closeness constraints can be represented by an undirected graph $G = (V, E)$, where V is the set of tasks. For an edge $(s, t) \in E$, the parameters for task s and t will be constrained to be close. Then, for more than 2 tasks, we replace (1) with the following:

$$\lambda_3 \sum_{(s,t) \in E} ||\vec{w}_s - \vec{w}_t||_2$$

There are many ways we can configure the edges. We could constrain all the parameters to be close to each other by specifying a complete graph. Or, we could arrange the tasks in a line: $\vec{w}_6 \leftrightarrow \vec{w}_{12} \leftrightarrow \vec{w}_{24} \leftrightarrow \vec{w}_{36} \leftrightarrow \vec{w}_{48}$. This is consistent with an assumption about the smoothness of disease progression. That is, we might assume that the difference between the disease progress at times t and $(t + 1)$ is relatively small.

In [ZLNY12], a multi-task model for predicting Alzheimer's progression is learned by optimizing an objective function similar to the one given in (1). The features in the input they used included demographic data, genetic data, the results of cognition tests, and lab test results. The parameters for prediction at time t are constrained to be close to the parameters at time $t + 1$. That is, a line graph is used. They found that increasing the penalty for parameter distance, λ_3, produced better results. Professor Sontag noted that this improvement is often seen in regimes where data is sparse.

5 Unsupervised Learning

Two primary goals of unsupervised learning in this context:

1. Discovery- getting new insights about disease, such as subtypes, to improve the design of clinical trials, or mechanism
2. Avoid overfitting- since data is scarce, one could save the label creation, a common source of overfitting, for the last step

5.1 K-means Clustering Algorithm

1. Hypothesize k random points as cluster centers
2. Assign data points to closet cluster center
3. Change the cluster center to the average of its assigned points
4. Repeat steps 2-3 until convergence (the cluster centers no longer change)

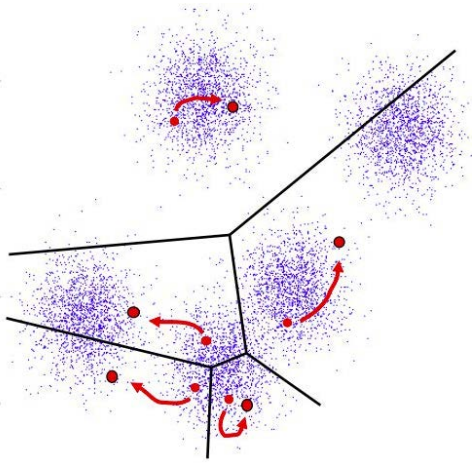

Figure 4: Here we have an example of 5 clusters being modelled using K-means. The plot is 2D, so humans can easily cluster the data visually, but K-means can also cluster high dimensional data.

5.2 K-means to Understand Asthma

How do we come up with better therapies for asthma? Some of the big questions include:

1. What are the different environmental or genetic factors that underlie certain subtypes of asthma?
2. Why do some people respond to therapy while others do not?
3. What are biomarkers and ways to predict who will respond to therapy?

In Halder et al.[HPS+08], they used a clustering algorithm to find five different subtypes of asthma. They used three different datasets:

1. Primary-care practices in the UK (less severe asthma), refractory
2. Asthma clinic in the UK (more severe asthma), and a 12-month

3. Longitudinal RCT clinical study (two different types of treatment).

The data preprocessing included normalized continuous variables and one-hot encoding for categorical variables. Some continuous variables were transformed by taking the logarithm, which allowed K-means to capture more of the range. There was some heterogeneity in the populations of each of the datasets.

5.3 Asthma Subtypes

The clusters include:

1. Early-onset atopic asthma- young population (average age of 14.6) with severe asthma exacerbations
2. Obese noneosinophilic- female (81%) and high BMI (36.2)
3. Benign asthma- not very severe asthma
4. Early symptom predominant
5. Inflammation predominant

The first dataset (primary-care) had clusters 1, 2, and 3. The second dataset had clusters 1, 2, 4, and 5. It excluded 3 because this population had more severe asthma. The patients participating in the clinical trial were classified to cluster 1, 2, or 4 based on their baseline data. The patients in each cluster were randomly assigned to either the clinical arm (standard clinical care) or the sputum arm (steroid titration to maintain normal eosinophil). They found that the number of patients that commenced oral corticosteroids was not different when all the patients were considered in aggregate. However, when they looked at the individual clusters, they found that the people in each cluster responded differently to treatment!

5.4 Limitations

1. It is challenging to differentiate between stage and subtype when doing unsupervised learning.
2. It assumes a single cluster explains all of the variation.
3. We have looked at using either supervised or unsupervised learning, but can we combine them?

References

[BSOM+14] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*, 33(7):1123–1131, 2014.

[HPS+08] Pranab Haldar, Ian D Pavord, Dominic E Shaw, Michael A Berry, Michael Thomas, Christopher E Brightling, Andrew J Wardlaw, and Ruth H Green. Cluster analysis and clinical asthma phenotypes. *American journal of respiratory and critical care medicine*, 178(3):218–224, 2008.

[WSW14] Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94. ACM, 2014.

[ZLNY12] Jiayu Zhou, Jun Liu, Vaibhav A Narayan, and Jieping Ye. Modeling disease progression via fused sparse group lasso. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1095–1103. ACM, 2012.

6.S897 / HST.956 Machine Learning for Healthcare

Spring 2019