Mandev S. Gill, Postdoctoral Research Scientist, Department of Statistics, Columbia University

Research Objectives and Aims: Development of statistical methods for predicting health outcomes from high dimensional longitudinal health histories. Development of computationally efficient algorithms for estimation.

Proposed Approach:

The emergence of large-scale patient-level databases of electronic health records and administrative claims presents great opportunities for patient-level predictive modeling. Several important open questions remain. For example, what types of predictive modeling approaches and algorithms are most effective for the different types of highly irregular longitudinal patient data that comprise medical record databases? How can we most effectively leverage medical knowledge to improve predictive models? To address these questions, we will empirically evaluate the performance of predictive models on accurately predicting outcomes at the patient level. We plan four different paths to predictive modeling: a baseline prediction model; automatic, ontology-driven phenotyping without sparse coding to drive traditional predictive models; sparse coding without automatic phenotyping; and automatic phenotyping combined with sparse coding.

Much of the machine learning and predictive modeling literature assume that the data are in "regression format." However, raw observational patient data are often irregular and consist of events that occur over time periods (for example, in-patient visits or collections of successive monthly drug prescriptions), events that occur at particular moments in time (such as procedures or certain acute conditions), and measurements at particular times. There is also non-temporal data such as gender or date-of-birth. We believe that improvements in predictive modeling are possible by exploiting a more tailored and optimized patient representation. This motivates adaptation of sparse coding to patient predictive modeling.

"Sparse coding" represents a medical record as a sparse linear combination of a data-derived "dictionary" or basis (Olshausen and Field, 1996; Raina et al., 2007). In more detail, given real-valued vectors $x_1, \ldots, x_m$ as inputs, sparse coding aims to learn a set of real basis vectors $b_1, \ldots, b_n$ such that each x_i can be represented as $x_i = \sum_{j=1}^{n} a_j^i b_j$. Here, a_j^i is the *activation* of basis b_j for input x_i. Importantly, L_1 regularization is employed to ensure that the activation vectors $a^i = (a_1^i, \ldots, a_n^i)$ are extremely sparse, with most elements equal to zero. This enables the reconstruction of any input x_i from only a few basis vectors. To conceptualize this framework in the context of observational health data, we can think of the basis vectors as exemplar healthcare patterns, and real patients are represented as sparse linear combinations of the exemplars. For example, suppose patient i has only three nonzero coefficients:

$a^i_{22} = 0.3$, $a^i_{667} = 0.8$, and $a^i_{1883} = 0.1$. Then patient i may be accurately represented as 0.3 times exemplar pattern 22 plus 0.8 times exemplar pattern 667 plus 0.1 times exemplar pattern 1883. Thus sparse coding is an elegant way of summarizing patient data while preserving individual patient characteristics.

A major advantage of sparse coding is that it can scale to massive data via online optimization algorithms based on stochastic approximations (Mairal et al., 2009). Furthermore, sparse coding is more flexible than alternative approaches based on principal component analysis in that it does not require basis vectors to be orthogonal.

"Automatic phenotyping" combines ontological knowledge with novel data-driven algorithms to identify predictively-important phenotypes. Consider a patient with a particular disease, such as chronic kidney disease, and several comorbidities (such as diabetes, anemia, hypertension). Such a patient will have a very large set of events (such as laboratory measurements, medication orders, procedures, and clinical documentation) stored in their record. However, such low-level features can be explained by their underlying concepts, or phenotypes (in this case, the disease and comorbidities). Phenotyping transforms raw patient data into clinically relevant, high level concepts (Hripcsak and Albers, 2013). To identify patient phenotypes, we envision leveraging existing ontological knowledge to identify a target set of clinical concepts that form meaningful phenotypes. We aim to combine this approach with probabilistic graphical models that, given a patient record with heterogenous, low-level features, infer a distribution over target phenotypes.

As a starting point for large-scale probabilistic phenotyping, we describe in detail the UPhenome model of Pivovarov et al. (2015). Assume we have a large data set of patient records, where each record comprises four data types: medication orders, diagnosis codes, free-text notes, and laboratory tests. As a motivating example, consider the condition coronary artery disease. Coronary artery disease can be thought of as a collection of the four data types. There may be information contained in free-text patient notes about a patient's social history (e.g., smoking, diet, and exercise) as well as symptoms and signs (e.g., chest pain, shortness of breath, and dizziness). Diagnosis codes may contain references to comorbidities (e.g., hyperlipidemia, diabetes, hypertension). There may be related medication orders (e.g., baby aspirin, nitroglycerine, anti-coagulations). Finally, there may be relevant laboratory tests (e.g., LDL, HDL, total cholesterol, stress test, cardiac catherization, coronary artery bypass graft). In a clinical database, in part because of inaccuracy and sparsity, but also because of the nature of medicine, a patient may exhibit many of the aforementioned elements and yet have no explicit diagnosis of coronary artery disease. A phenotyping model such as the UPhenome model is designed to identify the presence of the coronary artery disease phenotype based on its evidence, rather than an explicit mention.

From the input of patient records, the UPhenome model outputs learned phenotypes as well as an inference mechanism to identify a specific phenotype distribution for an

unseen patient record. The UPhenome model has the following variables:

P - Number of phenotypes
R - Number of patient records
β_r - Phenotype distribution for patient record r

l_p - Medications distribution for phenotype p
O_r - Number of medication orders in record r
$\chi_{o,r}$ - Medication instance o in record r
$\epsilon_{o,r}$ - Phenotype assignment for medication o in record r

η_p - Diagnosis code distribution for phenotype p
I_r - Number of Diagnosis codes in record r
$\nu_{i,r}$ - Diagnosis code instance i in record r
$\gamma_{i,r}$ - Phenotype assignment for diagnosis code i in record r

θ_p - Words distribution for phenotype p
N_r - Number of words in free-text notes in record r
$w_{n,r}$ - Word instance n in record r
$\delta_{n,r}$ - Phenotype assignment for word n in record r

κ_p - Laboratory test distribution for phenotype p
M_r - Number of laboratory tests in record r
$y_{m,r}$ - Laboratory test instance m in record r
$\zeta_{m,r}$ - Phenotype assignment for test m in record r

Given these variables, the UPhenome model assumes multinomial distributions for observations and phenotype assignments, and Dirichlet distributions for the phenotype distributions. In particular, we have the following generative process:

1. For each phenotype p in $\{1, \ldots, P\}$, sample phenotype distributions for each data type:

$$\begin{aligned} \eta_p &\sim Dirichlet(\mu) \\ \theta_p &\sim Dirichlet(v) \\ l_p &\sim Dirichlet(\xi) \\ \kappa_p &\sim Dirichlet(\pi). \end{aligned}$$

2. For each record in $\{1, \ldots, R\}$, sample patient phenotype composition $\beta_r \sim Dirichlet(\alpha)$. Then, for each diagnosis instance i, sample

$$\begin{aligned} \gamma_{i,r} &\sim Mult(\beta_r) \\ \nu_{i,r} &\sim Mult(\eta_{\gamma_{i,r}}). \end{aligned}$$

For each word instance n, sample

$$\begin{aligned} \delta_{n,r} &\sim Mult(\beta_r) \\ w_{n,r} &\sim Mult(\theta_{\delta_{n,r}}). \end{aligned}$$

For each medication instance o, sample

$$\begin{aligned} \epsilon_{o,r} &\sim Mult(\beta_r) \\ \chi_{o,r} &\sim Mult(l_{\epsilon_{i,r}}). \end{aligned}$$

For each lab test instance m, sample

$$\begin{aligned} \zeta_{m,r} &\sim Mult(\beta_r) \\ y_{m,r} &\sim Mult(\kappa_{\zeta_{i,r}}). \end{aligned}$$

The UPhenome model is a promising starting point, and there are a number of important ways in which we plan to build upon it. First, the UPhenome model does not account for the temporality of patient records. This is a major limitation because longitudinal records and diseases are not time-invariant. Pivovarov et al. (2015) propose dynamic topic modeling to incorporate temporality, based on the approach of Blei and Lafferty (2006). Second, it is important to determine the most effective way to incorporate clinical knowledge into phenotyping. We plan to tackle this problem by considering different prior distributions for phenotypes.

It remains to be determined whether combining phenotyping and sparse coding prediction will yield better models than each model on its own. Because the phenotypes are higher-level, constructed variables over raw data, they will be less sparse than raw data, but most patients have no data for long periods, and they may indeed benefit from the special processing offered by sparse coding.

References:

Blei, D. and Lafferty, J. (2006). Dynamic topic models. ICML.

Hripcsak, G. and Albers, D.J. (2013). Next-generation phenotyping of electronic health records. Journal of the American Medical Informatics Association, 20(1), 117-121.

Mairal, J., Bach, F., Ponce, G., and Sapiro, G. (2009). Online dictionary learning for sparse coding. ICML.

Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381, 607-609.

Pivovarov, R., Perotte, A.J., Grave, E., Angiolillo, J., Wiggins, C.H., and Elhadad, N. (2015). Learning probabilistic phenotypes from heterogeneous EHR data. Journal of Biomedical Informatics.

Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A.Y. (2007). Self-taught learning: Transfer learning from un-labeled data. ICML.

Impact: Methods for predicting health outcomes in patients, recommending treatment pathways, and identifying causal relationships.

Experience: Mandev S. Gill, PhD, Postdoctoral Research Scientist, Department of Statistics, Columbia University

I will be working on this project with the following faculty:

David Madigan, PhD, Professor, Department of Statistics, Columbia University

George Hripcsak, MD, Professor and Chair, Department of Biomedical Informatics, Columbia University

Patrick Ryan, PhD, Adjunct Assistant Professor, Department of Biomedical Informatics, Columbia University

Timeline: 24 months (From April 2016)