# PH 240C/STATS 245:
# Online Learning: Streaming Data Analysis in Healthcare

Jingshen Wang

November 10, 2021

According to Wikipedia, "in computer science, online machine learning is a method of machine learning in which data becomes available in a **sequential order** and is used to update the best predictor for future data at each step, as opposed to batch learning techniques which generate the best predictor by learning on the entire training data set at once. Online learning is a common technique used in areas of machine learning where it is computationally infeasible to train over the entire dataset, requiring the need of out-of-core algorithms. It is also used in situations where it is necessary for the algorithm to dynamically adapt to new patterns in the data, or when the data itself is generated as a function of time, e.g., stock price prediction. Online learning algorithms may be prone to catastrophic interference, a problem that can be addressed by incremental learning approaches."

In healthcare, online learning techniques enable effective analyses of massive streaming data assembled through smart phone apps, wearable devices, electronic medical record data and infectious disease surveillance programs. Healthcare research centered around these heterogeneous streaming data aims to answer scientific questions including assessing disease progression, studying drug adverse effect and patient dynamic health outcome, which allows real-time decision-making.

## 1 An example with HIV care engagement in electronic medical record data

The World Health Organization (WHO) reports that in the treatment of human immunodeficiency virus (HIV) and acquired immunodeficiency syndrome (AIDS), adherence to antiretroviral therapies (ART) varies between 37% and 83% depending on the drug under study, and lifelong ART success, including retention in care, is often undermined by stigma, food insecurity, negative clinic experiences, anticipated or actual side effects, and myriad factors potentially related to poverty (WorldHealthOrganization, 2003).

The increased availability of routinely collected electronic medical record data (EMR) enables the construction of accurate prediction algorithms of non-adherence, which allows us to proactively support patients struggling to comply with HIV care. As more healthcare systems are going through digital transformation and patient EMR information is updated periodically through linked pharmacies, once the prediction algorithm based on EMR data is scalable, it can be integrated into the clinical workflow in an online fashion and be used to identify patients in need of additional support, whether that be counseling or other supportive interventions.

To fully develop a dynamically updated generalizable predictive model using online EMR data, there are

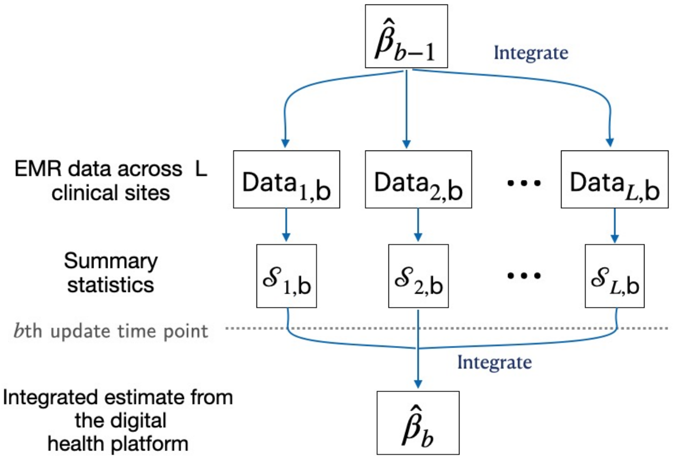

Figure 1: Ideal case for EHR prediction problems described in Section 1.

many challenges we face in practice. First, EMR data analysis often encounters privacy constraints in that individual patient data typically cannot be shared across different clinical sites, as breach of privacy arising from data sharing is a growing concern in general for scientific studies (Duan et al., 2018; Cai et al., 2021). Second, non-adherent patients living with chronic conditions may default and reengage in care numerous times over a lifetime, consequently, healthcare adherence is a dynamic process and our prediction algorithms must take this into account. Third, actionable prediction algorithms based on massive EMR need to be scalable and computationally efficient to be incorporated into the clinical workflow. To date, limited work on dynamic risk/health outcome prediction that overcomes above mentioned challenges simultaneously. See a ideal depiction of the clinical workflow in Figure 1.

## 2 Online streaming data analysis with continuous outcomes

We start with a simple (may be unrealistic, but always good to start with something simple and then generalize) scenario that we can update the health system on a daily basis Suppose in a streaming data environment, we have access to a continuous outcome variable $y_{it}$ and a vector of attributes $x_{it} \in \mathbb{R}^p$ observed at $t = 1, \ldots, m$ for subject $i = 1, \ldots, n$. Typically, medical practitioner might be interested in

1. How do outcome measures change over time? Is there a pattern associated with it?
2. How do the outcome measure depend on the covariates over time?

For example, the severity of a disease often depends on the patient's nutritional status, age, gender and family income, and these information might be observed once every few month. Naturally, the dependence of the outcome variable, severity of disease on the covariates is of interest.

In classical statistics, we would work with a linear model that assumes

$$y_{it} = x_{it}'\beta + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon_{it}|x_{it}] = 0, \; \mathbb{V}\mathrm{ar}[\varepsilon_{it}|x_{it}] = \sigma^2.$$

The problem of such a modelling assumption is that it restricts the flexibility of our analysis. For example, in a situation that $y_{it}$ measures the patients CD4 counts in HIV patients, and $x_{it}$ measures the transportation cost to travel to the clinic for picking up medication. We would expect the influence of transportation cost on the CD4 count changing over time when COVID hits.

A natural alternative of this linear model is to replace the constant $\beta$ with $\beta_t$ so that the coefficient is allowed to change over time:

$$y_{it} = x_{it}'\beta_t + \varepsilon_{it}, \quad \mathbb{E}[\varepsilon_{it}|x_{it}] = 0,\ \mathbb{V}\mathrm{ar}\big[\varepsilon_{it}|x_{it}\big] = \sigma^2.$$

Traditionally speaking, such a dynamic changing coefficient model is referred to as the varying coefficient model. If we have the data for the patient entire trajectory (Wu et al., 1998), we may simply estimate the coefficient through

$$\widehat{\beta}(t;h) = \underset{b\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}\sum_{j=1}^{m} K\left(\frac{x_{ij}-x_{it}}{h}\right)(y_{ij}-x_{ij}'b)^2 \triangleq \underset{b\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}\sum_{j=1}^{m} K_{ij}(t;h)(y_{ij}-x_{ij}'b)^2$$

where $K(\cdot)$ is a kernel function in a nonparametric sense and is a non-negative real-valued integrable function. Typically, it is desirable to define the kernel function that satisfies two constraints:

1. Normalization: $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$
2. Symmetry: $K(-u) = K(u)$ for all values of $u$

Think about why we put down these two constraints in practice? And the resulting estimator has the form:

$$\begin{aligned}\widehat{\beta}(t;h) &= \left(\sum_{i=1}^{n}\sum_{j=1}^{m} K_{ij}(t;h)x_{ij}^2\right)^{-1}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} K_{ij}(t;h)x_{ij}y_{ij}\right)\\ &= \left(\sum_{i=1}^{n} X_i'K_i(t;h)X_i\right)^{-1}\left(\sum_{i=1}^{n} X_i'K_i(t;h)Y_i\right),\end{aligned}$$

where $K_i(t;h) = \mathrm{Diag}\left(K_{i1}(t;h),\ldots,K_{i,}(t;h)\right)$. However, in an online learning environment, at time point $t$, we cannot expect to have access to the patient future data. Such an estimate is thus not feasible in practice. This motivate us to consider a one-sided kernel function instead of the traditional kernel function:

$$\widehat{\beta}(t;h) = \underset{b\in\mathbb{R}^p}{\arg\min}\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda^{t-j}(y_{ij}-x_{ij}'b)^2, \quad \lambda\in(0,1).$$

Incorporating the weighting function $\lambda^{t-j}$ into the loss function enables us to dynamically down weight the observations far away from the current time updating time point $t$. Although such a weighting scheme is just one choice among a broad class of weighting functions, its benefit will be apparent in the following derivations. Recall our task in the streaming data environment is to find scalable, it admits a recursive expression which allows to sequentially update the previous batch estimator $\widehat{\beta}_{t-1}$ when the new data batch arrives:

$$\widehat{\beta}_t^{\lambda} = \widehat{\beta}_{t-1}^{\lambda} + \widehat{\Sigma}_t^{-1}\sum_{i=1}^{n} x_{it}'\big(y_{it} - x_{it}'\widehat{\beta}_{t-1}^{\lambda}\big), \quad \widehat{\Sigma}_t^{\lambda} = \lambda\widehat{\Sigma}_{t-1}^{\lambda} + \sum_{i=1}^{n} x_{it}'x_{it}.$$

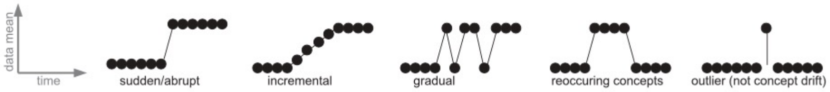

**Fig. 2.** Patterns of changes over time (outlier is not concept drift).

Figure 2: Example of changes shown in Gama et al. (2014).

And the prediction interval for future patient can be constructed by

$$x'_{i,t+1}\widehat{\beta}_t \pm 1.96 \cdot \widehat{\sigma}_t \cdot \left(x'_{i,t+1}\widehat{\Sigma}_t^{-1}x_{i,t+1}\right)^{1/2}.$$

Thus, such a recursive form allows us to conduct online statistical (predictive?) inference. Nevertheless, finding an optimal tuning parameter $\lambda$ is not an easy task in practice. Intuitively, what kind of $\lambda$ appeals e to you?

### 2.1 Connection with existing machine learning literature

In a recent ML survey, Gama et al. (2014) has discussed the issues of online learning when "concept drift" happens. Formally, because data is expected to evolve over time, especially in dynamically changing environment, where non-stationary is typically, its underlying distribution can change dynamically over time. Suppose between time point $t_0$ and $t_1$, there exist an attribute such that the joint distribution of $(X, y)$ change over time, that is

$$p_{t_0}(X, y) \neq p_{t_1}(X, y), \quad \exists X,$$

where $p_t(\cdot, \cdot)$ denotes the joint distribution at time $t_0$ between the set of attribute variable $X$ and the target variable $y$. Changes in data can thus be characterized as changes in the component of this relation. In the literature, there are two types of drift:

1. Real concept drift refers to the changes in $p(y|X)$.
2. Virtual drift happens if the distribution of the incoming data changes, i.e., $p(X)$.

There are, of course, different kind of drift/changes may happen over time. For example, Figure 2 demonstrates different types of changes. Does the introduce model setup cover these changes? If not, how would you propose to change this model?

## 3 More challenging online streaming data analysis in healthcare with binary outcomes

In a real world scenario, it is hard to imagine that any digital platforms will allow daily update on the machine learning algorithms. At best, we would expect the system can be updated every week or every few weeks. We refer to this case as batch update. In this case, between batch updating time points, different patients would have different number of updates; see Figure 3 for an illustration. In other words, within a batch update time period, each patient has repeated and correlated measurements.

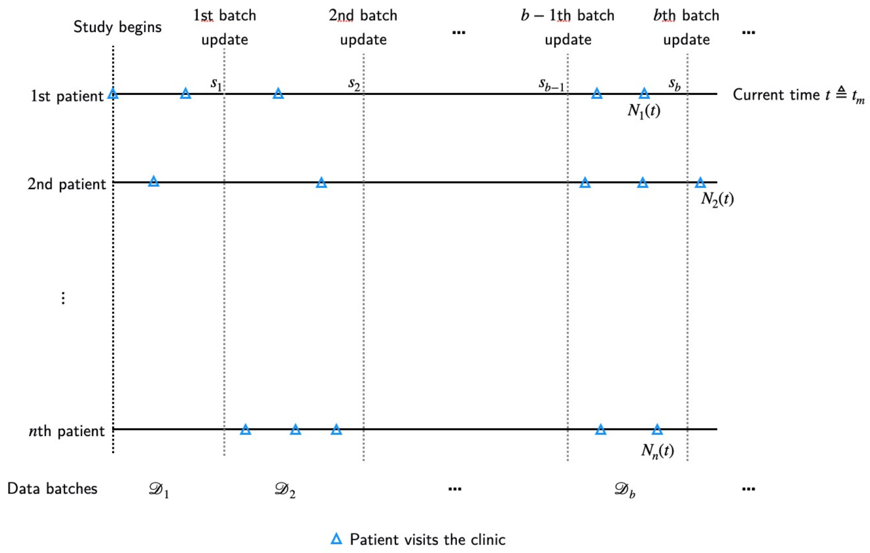

Figure 3: Illustration of online learning with electronic medical record data.

To establish notation, suppose we have different batch update time points $s_1, s_2, \ldots, s_b$, we let $Y_i = (y_{i1}, \ldots, y_{in_i})'$ be the vector of outcome values and $X_i = \big(x_{i1}, \ldots, x_{in_i}\big)' \in \mathbb{R}^{n_i \times p}$ matrix of covariates values for the $i$th subject $i = 1, \ldots, m$. Suppose our outcome is a binary random variable $y_{it} \in \{0, 1\}$, and we further assume a marginal generalized linear model as

$$\mathbb{E}\big[y_{it}|x_{it}\big] = \mu(x_{it}).$$

The conditional density of $y_{it}|x_{it}$ is

$$f_{y|x}(y_{it}|x_{it}) = \exp\left(y_{it} \cdot \log \frac{\mu(x_{it})}{1-\mu(x_{it})} - \log\Big(1-\mu(x_{it})\Big)\right) = \exp\Big(y_{it} \cdot x_{it}'\beta - \log\Big(1-\mu(x_{it})\Big)\Big)$$

with

$$\theta_{it} \triangleq \log \frac{\mu(x_{it})}{1-\mu(x_{it})} = x_{it}'\beta_b.$$

Then, under the working independence assumption, the score equation from a likelihood perspective has the form (Liang and Zeger, 1986):

$$\sum_{i=1}^{n} X_i' S_i \triangleq \sum_{i=1}^{n} X_i' \left(Y_i - \mu(X_i)\right) = 0, \quad \mu(X_i) = \big(\mu(x_{i1}), \ldots, \mu(x_{in_i})\big)',$$

where $X_i = (x_{i1}', \ldots, x_{in_i}')' \in \mathbb{R}^{n_i \times p}$ and $Y_i = (Y_{i1}, \ldots, Y_{in_i})' \in \mathbb{R}^{n_i \times p}$. An estimator that solves the above equation is easy-to-compute with existing software, and the estimator is *consistent* as long as the model is correctly specified. Next question is how to carry out this estimator in an online fashion? Motivated by our previous construction, suppose our data are collected in different batches $\mathcal{D}_1, \ldots, \mathcal{D}_b$ and within each batch, we may work with the weighted estimating equation:

$$\sum_{j=1}^{b} \lambda^{b-j} \sum_{i=1}^{n} X'_{i,j} \left(Y_{i,j} - \mu(X_{i,j})\right) = 0, \quad \mu(X_{i,b}) = \big\{\mu(x_{ij}), j \in \mathcal{D}_b\big\}.$$

Given this estimating equation, how can we dynamically revise the model to get an estimate of the coefficient $\beta_b$?

# References

Tianxi Cai, Molei Liu, and Yin Xia. Individual data protected integrative regression analysis of high-dimensional heterogeneous data. *Journal of the American Statistical Association*, (just-accepted):1–34, 2021.

Rui Duan, Mary Regina Boland, Jason H Moore, and Yong Chen. Odal: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. In *BIOCOMPUTING 2019: Proceedings of the Pacific Symposium*, pages 30–41. World Scientific, 2018.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

WorldHealthOrganization. *Adherence to long-term therapies: evidence for action*. World Health Organization, 2003.

Colin O Wu, Chin-Tsang Chiang, and Donald R Hoover. Asymptotic confidence regions for kernel smoothing of a varying-coefficient model with longitudinal data. *Journal of the American statistical Association*, 93 (444):1388–1402, 1998.