Minimizing Misclassification Cost in Healthcare Associated Infection

Elioth Sanabria
m.elioth@columbia.edu

David D. Yao
yao@columbia.edu

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY10027.

June 13, 2019

Abstract

Motivated by healthcare associated infection (HAI), which is estimated to cost US hospitals $9.8 billion per year, we develop a machine-learning based scheme to administer patient-targeted preventive measures upon admissions to the hospital. The model involves an objective function that acknowledges the asymmetry between the cost of missing an infection and the cost of a "false alarm," the former (real cost to the hospital) could be order-of-magnitude higher than the latter (minor expenses for certain preventive measures, which are nonetheless imperfect). A two-step algorithm is devised to solve the problem in a much more efficient manner than a black-box algorithm such as deep neutral network, and with a clearly interpretable solution. We also provide convergence and rate-of-convergence results, in the form of distribution-free probabilistic guarantees, using a variation of the Dvoretzkey-Kiefer-Wolfowitz bound (as refined by Massart). We illustrate the performance of the model with real data from several largest hospitals in New York City.

Keywords: Machine Learning, Binary Classification, Cost Sensitive Learning, Performance Bounds, Infection Prevention.

1 Introduction

According to Zimlichman et al. 2013, healthcare-associated infections (HAI) are estimated to cost 9.8 billion USD per year. Most of these infections are reasonably preventable Umscheid et al. 2011, making them an ideal candidate for machine-learning (ML) prediction techniques using electronic health records (EHR) of patients (e.g: Wiens et al. 2012, Schoonover et al. 2017).

Most HAI belong to a broader category of adverse conditions known as Hospital Acquired-Conditions. Their categorization into these conditions is not coincidence, the Deficit Reduction Act DRA 2006 highlights three characteristics of Hospital Acquired-Conditions:

- High cost or high volume of occurrence.
- Result in the assignment of a case to a Diagnosis Related Group that has a higher payment when present as a secondary diagnosis.

- Could reasonably have been prevented through the application of evidence-based guidelines.

Yet, there is a conceptual hole in the academic literature unifying ML predictions of HAI and the economic incentives of the hospital to implement ML models (*what-when-how*). In other words, there is little consensus as to *What to do with the predictions and when/how to use them?* in healthcare settings. Quoting Chen et al. 2017:

> "An accurate prediction of a patient outcome does not tell us what to do if we want to change that outcome - in fact, we cannot even assume that its possible to change the predicted outcomes."

We propose an unified approach answering the *what-when-how*, while addressing the economic incentives the hospital faces to implement such system.

1.1 Economic Incentives for HAI Prevention

Beyond the clear public health benefits of preventing HAI (decreased associated mortality rates, shorter associated length of stay), hospitals also benefit monetarily by preventing HAI in the US.

Since October 2008, hospitals do not receive additional payments from Medicare centers if a secondary condition (i.e. a condition not present on the first day) correspond to one of the adverse conditions specified in Deficit Reduction Act. That is, hospitals receive payments as if the Hospital Acquired-Condition was not present as a secondary diagnosis in the first place, consequently, this cost has to be paid in full by the hospital.

To put the problem in perspective in Table 1 the average cost per case of HAI, and average length of stay is presented. This is precisely the cost that the hospital will have to pay that will not be reimbursed from Medicare centers. In some cases, part of this cost will shift to the patient themselves.

Table 1: Average costs per single case and length of stay for most common HAI. Data collected by Zimlichman et al. 2013

Healthcare Associated Infection (HAI)	Cost per case	Length of Stay
Catheter-associated urinary tract infections (CAUTI)	$896	NR
Central line-associated bloodstream infections (CLABSI)	$45,814	10.4
Clostridium difficile infection (CDIFF)	$11,285	3.3
Surgical site infection (SSI)	$20,785	11.2
Ventilator-associated pneumonia (VAP)	$40,144	13.1

Using universal prevention measures hospitals have achieved steady progress towards HAI reduction. In Table 2 the 2020 HAI reduction targets the U.S. Department of Health and Human

Table 2: HAI reduction targets with respect to 2015 baseline by the U.S. Department of Health and Human Services

Measure (and data source)	Progress made by 2016	2020 Target (from 2015 baseline)
CAUTI (NHSN)	56% relative reduction	25% reduction
CLABSI (NHSN)	10% reduction	50% reduction
CDIFF (NHSN)	8% reduction	30% reduction
SSI (NHSN)	13% reduction	30% reduction

Services are presented. Nonetheless, there is still plenty of room for reduction to achieve these goals.

The key question is if by giving patient-targeted preventive measures using ML to patients before the infections are developed, could the costs be further reduced? In the next section we explore the costs of prevention.

1.2 Prevention Costs

An overview of measures (interventions) for HAI prevention with their correspoding success rates were collected by Umscheid et al. 2011. We merge these with estimated costs interpolated from Arefian et al. 2016 in Table 3. Most of these interventions are known as *horizontal* prevention measures (e.g. Septimus et al. 2014).

One remarkable feature of these measures is that their costs is orders of magnitude lower than the infection costs for every infection. This is the rationale behind the value-based incentives of not reimbursing hospitals for infection.

Table 3: Summary of Preventive Intervention Costs (per patient-stay) and Success Rates. Success rates and example interventions estimated from Umscheid et al. 2011. * Estimated cost per patient from total intervention cost from Arefian et al. 2016. ** success rate from Dingle et al. 2017.

HAI	Cost	Success(%)	Example Intervention
CAUTI	\$16.5*	69%	Reduction in placement of catheters; removal of unnecessary catheters.
CLABSI	\$22.26*	66%	Sterile barrier precautions; chlorhexidine disinfection
CDIFF	\$115.62*	80%**	Antibiotic stewardship; limit patient contact (isolation; extra care).
SSI	\$50	54%	Improvement in perioperative glucose control.
VAP	\$215	46%	Hand hygiene; head of bed angle 130; daily interruption of sedation.

1.3 Paper Outline and Contributions

In Section 1 we described the economic incentives a hospital faces to prevent Healthcare Associated Infections. We formalize the *mis-classification cost problem* and present the general model in Section 2. In Section 3 we study and propose a two-step optimization algorithm to solve the data-

driven version of the problem, where we make a connection to the more popular Cross-Entropy minimization for binary classification and the Neyman and Pearson Test. In Section 4 we study the statistical properties related to its convergence and rate of convergence. Our results in this section improve upon known Vapnik-Chervonenkis (see Devroye et al. 1996) and Large Deviations Theory (see Kleywegt et al. 2002) bounds on the same setting, these apply for any family of classifiers (including deep neural networks and families of models with infinite V-C dimension) and are distribution free. We illustrate these results with a numerical example using real data in Section 5 and close with concluding remarks in Section 6.

2 The Mis-classification Problem

Patients become exposed to HAI at the very same moment they are admitted to the hospital. The early days of admission are critical because 80% of HAI happen within the first 10 days of admission. This means most HAI cases become colonized (infected, but asymptomatic) within the first 7 days of admission to the hospital. Then, early action is going to be crucial if an adverse outcome is to be prevented.

We propose a fully automated assessment of the patient at admission (on day 1) based on standard Electronic Health Records, where the patient is automatically classified to receive a preventive measure from the moment of admission until discharge from the hospital. After a patient is classified to receive an infection preventive measure, this will be added to the schedule and care of the patient. In Figure 1 we present a flow chart of how the ML scheme works.

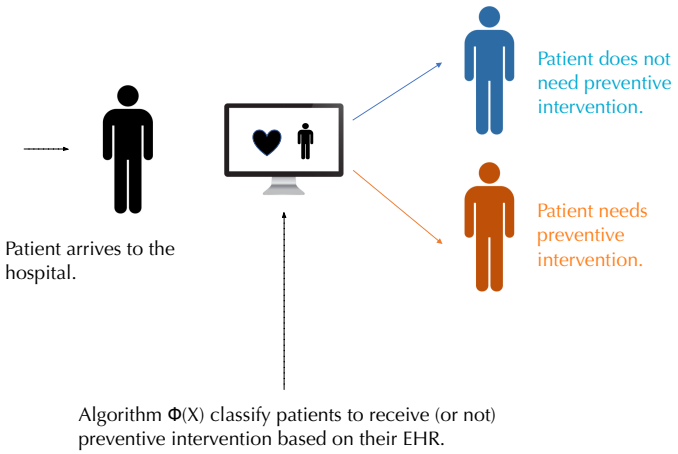


Figure 1: Explanation of Machine Learning scheme to give prevention measures at admission using Electronic Health Records (EHR).

Each patient has Electronic Health Records collected and available at admission codified into a vector $X \in \mathcal{X}$ ($\mathcal{X}$ can be $\mathcal{R}^d$), these records are collected on day 1 of admission. At discharge, for each patient the infection status is codified into the indicator random variable $Y \in \{0, 1\}$. The machine learning scheme $\phi := (p, \rho)$ is a tuple denoting a machine learning model p that outputs the probability of infection and a threshold for classification ρ. A patient with features x is classified to receive a preventive intervention if $p(x) \geq \rho$, that is, the patient receives a preventive treatment if the probability of infection $p(x)$ exceeds a threshold ρ.

Formally, $Y = \{0, 1\}$ is a Bernoulli r.v., with $Y = 1$ representing the patient is infected, and $\pi := \mathsf{P}(Y = 1)$ the infection rate; $X = (X_1, \ldots, X_\ell)$, a random vector, represents the vector of patient's features upon admission collected in the EHR. Given a patient with features X, the cost of mis-classifiying a patient can happen in two ways: on one hand, a preventive measure with cost K_0 will be lost on a patient that receives the preventive measure but doesn't develop an infection, that is, when $p(X) \geq \rho$ and $Y = 0$; on the other hand, an infection with cost K_1 will be developed in patients that did not receive the preventive treatment, that is, when $p(X) < \rho$ and $Y = 1$. Later we will show that this formulation is general enough to accommodate the case when the preventive treatment doesn't have a 100% success rate preventing the infection.

The mis-classification cost problem we want to solve can be presented as follows:

$$\min_{\phi:=(p,\rho)} K_0 \mathsf{E}\big[(1-Y)\mathbf{1}\{p(X) \geq \rho\}\big] + K_1 \mathsf{E}\big[Y\mathbf{1}\{p(X) < \rho\}\big] := \min_{\phi} \mathcal{C}(\phi) := \mathcal{C}(\phi^*), \tag{1}$$

where $\rho \in [0, 1]$, and p is a function that maps X to $[0, 1]$. For instance, p can be chosen from a logistic regression model or a deep neutral network (DNN), or a decision tree/random forest (more about this below). Then, the objective is to select a model ϕ that better minimizes the expected mis-classification cost.

Since the distribution of (X, Y) is unknown, we pursue the following "data-driven" formulation, where (x_i, y_i) for $i = 1, \ldots, n$ are i.i.d. copies of (X, Y):

$$\min_{\phi} \frac{1}{n} \sum_{i=1}^{n} \big[K_0(1-y_i)\mathbf{1}\{p(x_i) > \rho\} + K_1 y_i \mathbf{1}\{p(x_i) \leq \rho\}\big] := \min_{\phi} \hat{\mathcal{C}}_n(\phi_n). \tag{2}$$

3 Algorithms and Solutions

The difficulty in minimizing the objective in (2) lies mainly in two factors: (i) The choice of models ϕ, and (ii) the non-convexity of the objective function. For example, by assuming $p(x_i) := \beta' x_i$ is the set of linear functions (such as the logistic regression family) the data-driven problem can be expressed as:

$$\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \big[K_0(1-y_i)\mathbf{1}\{\beta' x_i > 0\} + K_1 y_i \mathbf{1}\{\beta' x_i \leq 0\}\big] = \min_{\beta} \sum_{i \in S_0} K_0 I_i + \sum_{j \in S_1} K_1 I_i. \tag{3}$$

Where $S_0 := \{i \in [n] : y_i = 0\}$ and $S_1 := \{i \in [n] : y_i = 1\}$ are the sets of healthy and infected patients, and $I_i = 1 \iff \beta' x_i > 0$ for $i \in S_0$ and $I_j = 1 \iff \beta' x_j \leq 0$ for $j \in S_1$ are binary variables conditional on whether the observation is classified one class or the other. Therefore, the problem is an instance of MIP and NP-hard even for linear models. Nonetheless, recent advances in SDP and SOS optimization (e.g. Lasserre 2001) allow to solve exactly this instance of the problem for small values of n (less than 1,000 in modern solvers for a reasonable amount of time and memory).

For larger values of n and different families of models, we explore in this section a tractable optimization scheme that while not guaranteed to find the global minimum, will find a local minimum in a reasonable amount of time without enumerating all possible solutions. We start by describing the optimization of the problem when the model p is fixed and only the threshold ρ varies. We then introduce a two-step algorithm to solve this optimization problem. In addition, we also present a brief overview of two other algorithms, deep neural network and decision tree, which will be used to compare against in the two-step algorithm.

3.1 Solving for ρ

Here we focus on finding the best $\rho \in [0, 1]$ to do the classification assuming $p(x_i)$ is given for all $i \in S_0 \cup S_1$. First, divide the interval $[0, 1]$ into M equal segments, with M given (say, $M = 10, 100$ or any other integers depending on the desired precision). Let n_{0m} denote the number of patients $i \in S_0$ such that $p(x_i)$ takes value in the *m-th* segment, and analogously denote n_{1m} for patients in S_1. Specifically,

$$n_{0m} := \{\, \#i \in S_0 : p(x_i) \in (\frac{m-1}{M}, \frac{m}{M}]\,\} \quad \text{and} \quad n_{1m} := \{\, \#i \in S_1 : p(x_i) \in (\frac{m-1}{M}, \frac{m}{M}]\,\}. \tag{4}$$

Thus, $\sum_{m=1}^{M} n_{0m} = n_0$ and $\sum_{m=1}^{M} n_{1m} = n_1$. While this discretization is not necessary for the analysis below, it does facilitate computation, as well as help with intuition.

Consider a ρ that falls into the *m-th* segment, $\rho \in (\frac{m-1}{M}, \frac{m}{M}]$. Ignoring the constant factor $1/(n_0 + n_1)$ in the cost objective in (2), which is equal to both $(1 - \pi)/n_0$ and π/n_1, the cost in (2) can be expressed as:

$$K_0 \sum_{k=m}^{M} n_{0k} + K_1 \sum_{k=1}^{m-1} n_{1k}. \tag{5}$$

Now, suppose we increase the value of ρ from the *m-th* segment to the $(m + 1)$*-th*. Then, the first sum in (2) will lose its *m-th* term and the second sum will gain its *m-th* term, so the total will decrease by an amount equal to

$$K_0 n_{0m} - K_1 n_{1m}. \tag{6}$$

Therefore, we can keep increasing m until the above difference becomes negative, i.e., the best ρ should fall into the m^*-*th* segment, where

$$m^* = \min\{\, m : \ \frac{n_{0m}}{n_{1m}} < \frac{K_1}{K_0} \,\}. \tag{7}$$

Note, the validity of the above argument depends on $\frac{n_{0m}}{n_{1m}}$ being monotone (decreasing) in m; o.w., m^* might only be locally optimal, in which case a search over all M segments is needed.

As it turns out, there is a connection to the Neyman-Pearson test (see Neyman et al. 1933). Denote:

$$g = \{g_m := n_{0m};\, m = 1, \ldots, M\} \quad \text{and} \quad h = \{h_m := n_{1m};\, m = 1, \ldots, M\}, \tag{8}$$

corresponding to the two distributions in the N-P test. (Note $\{g_m\}$ and $\{h_m\}$ can be normalized by dividing by n_0 and n_1, but there's no need to do so.) Then, $p(x_i) \geq \rho$ becomes $g_m/h_m \geq \rho$ as evident from the above argument; and the best ρ is equal to (approximately, subject to the precision of discretization) $\rho^* = m^*/M$ with m^* following (7). The details are as follows.

Suppose a discrete random variable ξ takes on values v_m, $m = 1, \ldots, M$, with a probability distribution that is either $g = \{g_m\}$ or $h = \{h_m\}$. Draw i.i.d. samples of ξ; and every time classify its distribution by a classifier ϕ_c with $\phi_c = 1$ (resp. 0) mapping ξ to the h (resp. g) distribution. Specifically,

$$\phi_c(m) := \phi_c(\{\xi = v_m\}) = \mathbf{1}\{f_m := g_m/h_m \geq c\},$$

where c is a parameter (to be optimized).

The mis-classification cost is hence,

$$\sum_{m=1}^{M} \big[K_0 \phi_c(v_m) g_m + K_1 (1 - \phi_c(v_m)) h_m\big].$$

The above can be written more explicitly as follows:

$$K_0 \sum_{m: f_m \geq c} g_m + K_1 \sum_{m: f_m < c} h_m. \tag{9}$$

Assume f_m is monotone (decreasing) in m, and let $m = f^{-1}(c)$. Then, as long as

$$K_0 g_m - K_1 h_m \geq 0, \quad \text{or} \quad \frac{g_m}{h_m} \geq \frac{K_1}{K_0},$$

we want to keep increasing m so as to decrease the cost, following the same argument that deals with (5) and (6) above. Thus, it is optimal to stop at m^* with

$$m^* = \min\{\, m : \ \frac{g_m}{h_m} < \frac{K_1}{K_0} \,\},$$

which parallels (7). Thus, the optimal c is $c^* = K_1/K_0$.

In the continuous case, $g(x)$ and $h(x)$ are two density functions; $f(x) := g(x)/h(x)$, and $\phi(x) = \mathbf{1}\{f(x) \geq c\}$. The cost in (9) takes the following form:

$$K_0 \int_{x:f(x)\geq c} g(x)dx + K_1 \int_{x:f(x)\leq c} h(x)dx.$$

When $f(x)$ is decreasing, the above can be rewritten as

$$K_0 \int_{x\leq f^{-1}(c)} g(x)dx + K_1 \int_{x\geq f^{-1}(c)} h(x)dx.$$

Taking derivative w.r.t. c, the first-order optimality condition is

$$K_0 g(f^{-1}(c)) = K_1 h(f^{-1}(c)), \quad \text{or} \quad \frac{K_1}{K_0} = \frac{g(f^{-1}(c))}{h(f^{-1}(c))} = f(f^{-1}(c)) = c.$$

3.2 Cross-Entropy (CE) Minimization

The cost objective in (2) is a sum of indicator functions, which are not continuous (let alone differentiable); hence, it's difficult to minimize. One remedy is to solve a cross-entropy (CE) minimization problem as a surrogate (for a detailed treatment, see Bartlett et al. 2006), meaning we want the estimated probabilities $p(x_i)$ be such that the CE — a measure of distance between two distributions — between y_i and $p(x_i)$ is minimized:

$$\min \; K_0 \sum_{i\in S_0} -\log(1-p(x_i)) + K_1 \sum_{i\in S_1} -\log p(x_i), \tag{10}$$

Intuitively, in minimizing CE, we want the $p(x_i)$'s associated with $i \in S_1$ be as close as possible to $y_i = 1$ and those associated with $i \in S_0$ be as close as possible to $y_i = 0$; in other words, to best match the y_i's and improve the prediction quality.

Another way to appreciate the CE objective is to observe the following inequalities (where we write $p(x_i) := p_i$ for simplicity), which are readily verified

$$\mathbf{1}\{p_i \leq \rho\} \leq \frac{-\ln(p_i)}{-\ln(\rho)} \quad \text{and} \quad \mathbf{1}\{p_i \geq \rho\} \leq \frac{-\ln(1-p_i)}{-\ln(1-\rho)}.$$

Thus, dividing the CE objective by a constant, $-\ln(\rho \wedge (1-\rho))$ (which will not affect the solution to the CE problem), we will have an upper bound on the objective function of the original problem in (2). Not to add, as the CE objective is independent of ρ, it provides a natural way to separate out the two solution approaches, to $p(\cdot)$ and to ρ.

To illustrate how to solve the CE problem, consider the logistic regression model, where the $p(\cdot)$ function is:

$$p(x_i) = \frac{1}{1+e^{-w\cdot x_i}}, \quad \text{where} \quad w \cdot x_i = \sum_{j=1}^{\ell} w_j x_{ij}. \tag{11}$$

Here the vector $w := (w_j)_{j=1}^{\ell}$, is what we want to optimize.

Write $p_i := p(x_i)$ as before. We have

$$\frac{\partial p_i}{\partial w_j} = p_i(1-p_i)x_{ij}. \tag{12}$$

Let g denote the CE objective function in (10). Making use of (12), we have

$$\frac{\partial g}{\partial w_j} = K_0 \sum_{i \in S_0} p_i x_{ij} - K_1 \sum_{i \in S_1} (1-p_i)x_{ij}, \quad j = 1, \ldots, \ell;$$

and

$$\frac{\partial^2 g}{\partial w_j w_k} = K_0 \sum_{i \in S_0} p_i(1-p_i)x_{ij}x_{ik} + K_1 \sum_{i \in S_1} p_i(1-p_i)x_{ij}x_{ik}, \quad j,k = 1, \ldots, \ell. \tag{13}$$

Denote $\theta_i^2 := p_i(1-p_i)\big[1 + (K-1)\mathbf{1}\{i \in S_1\}\big]$, with $\theta \geq 0$. It is readily verified that the Hessian has the following representation

$$\mathcal{H} := \Big[\frac{\partial^2 g}{\partial w_j w_k}\Big]_{j,k=1}^{\ell} = XX', \quad \text{where} \quad X := [\theta_1 x_1, \ldots, \theta_n x_n], \tag{14}$$

and X' denotes the transpose of the matrix X. Thus, we can conclude that the Hessian is positive semi-definite. In fact, it is non-singular, i.e., positive definite, provided the (random) matrix X (of n i.i.d. random vectors, $\theta_i x_i$, $i = 1, \ldots, n$) is full rank ($= \ell$). This is almost certainly true, since $n >> \ell$; in which case, g is strictly convex in w.

In summary, when $p(\cdot)$ follows the logistic regression model, the CE problem is a convex minimization problem, which can be readily solved making use of the Hessian; and the solution is unique.

3.3 A Two-Step Algorithm

Our optimization scheme will take into account two facts: (i) That solving the formulation (10) is an upper bound of the original data driven formulation (2) and (ii) that while we cannot directly perform gradient descent on (2) we can smooth the indicators in the objective function and make the function differentiable with an approximation of desired accuracy.

A natural way to smooth indicator functions is by using a sigmoid function, this can be accomplished to any desired accuracy by approximating an indicator $\mathbf{1}\{x > 0\}$ with $\tilde{\psi}(x) := \big(1 + \exp(-xL)\big)^{-1}$, when $L \to \infty$ the distance between the indicator and the approximation goes to 0. Once the indicators are smoothed we can write the objective (2) as:

$$\min_{\phi} \hat{\mathcal{C}}_n(\phi_n) \approx \min_{\phi} n^{-1}\Big[K_0 \sum_{i \in S_0} \tilde{\psi}\big(p(x_i) - \rho\big) + K_1 \sum_{j \in S_1} \tilde{\psi}\big(\rho - p(x_j)\big)\Big]. \tag{15}$$

That now is amenable for smooth optimization packages using gradient descent. Unfortunately, the smoothing doesn't solve the intrinsic non-convexity of the problem and starting the optimization at an arbitrary point yields poor solutions.

In practice, a way to get a good initialization for the objective in (15) is to first solve the cross-entropy minimization problem in (10) and use this solution as a starting point of the optimization of (15). In summary, we propose the following algorithm to solve the data-driven formulation:

- Step 0: Let $\theta \in \Theta$ be the family of parameters of the model p. Initialize $\theta \leftarrow 0$.
- Step 1: Solve the Cross-Entropy formulation:

$$\theta' \leftarrow \text{argmin}_{\theta\in\Phi}\ K_0 \sum_{i\in S_0} -\log(1-p(x_i)) + K_1 \sum_{i\in S_1} -\log p(x_i),$$

- Step 2: Using θ' as a starting point for the optimization, solve:

$$\min_{\phi}\, n^{-1}\Big[K_0 \sum_{i\in S_0} \tilde{\psi}\big(p(x_i)-\rho\big) + K_1 \sum_{j\in S_1} \tilde{\psi}\big(\rho - p(x_j)\big)\Big].$$

3.4 Deep Neural Network (DNN)

Let $W^{(d)}$, $d = 1,\ldots,D$, be a set of matrices. Let $\psi^{(d)}$ be a sequence of operators on a vector $x = (x_j)_{j=1}^{\ell}$; e.g., $\psi^{(d)}(x) = x$, or $\psi^{(d)}(x) = (\frac{1}{1+e^{-x_j}})_{j=1}^{\ell}$. A deep neutral network (DNN) can be constructed as follows: each layer, the d-th, is represented by a vector $x^{(d)}$, defined recursively as follows:

$$x^{(1)} = \psi^{(1)}(W^{(1)}x) \quad \text{and} \quad x^{(d)} = \psi^{(d)}(W^{(d)}x^{(d-1)}), \quad d = 2,\ldots,D; \tag{16}$$

where $x = (x_j)_{j=1}^{\ell}$ is a given vector, the input data.

In DNN, the matrices $W^{(d)}$ are decision variable to be optimized, whereas the operators $\psi^{(d)}$ are given. First, about the latter: $\psi^{(d)}$ can be the same for all d; although typically, the last one $\psi^{(D)}$ is allowed to be different from the others. As to the matrices, the number of columns for $W^{(d)}$ must be equal to the dimension of $x^{(d-1)}$, for all d (with $x^{(0)} := x$, the input vector); whereas the number of rows can a free parameter of choice, for all $d \neq D$. The last matrix, $W^{(D)}$, is required to have a single row, so that $x^{(D)} = \psi^{(D)}(W^{(D)}x^{(D-1)})$ is a scalar, which is the counterpart of $p(x)$. Below we shall take $\psi^{(d)}(x) = (\frac{1}{1+e^{-x_j}})_{j=1}^{\ell}$, for all $d = 1,\ldots,D$.

When the input is the vectors x_i $(i = 1,\ldots,n)$, the vectors $x^{(d)}$ above are expressed as $x_i^{(d)}$. Note, in particular, that $x_i^{(D)} = [1+\exp(W^{(D)}x_i^{(D-1)})]^{-1}$ is a scalar; and the (CE) objective function is

$$g(W^{(1)},\ldots,W^{(D)}) := \sum_{i=1}^{n} \big[-K_0(1-y_i)\ln(1-x_i^{(D)}) - K_1 y_i \ln(x_i^{(D)})\big].$$

Write $W^{(D)} = (w_j)$. Similar to the logistic regression model, we have $\frac{\partial x_i^{(D)}}{\partial w_j} = x_i^{(D)}(1 - x_i^{(D)})x_{ij}^{(D-1)}$; and hence,

$$\frac{\partial g}{\partial w_j} = K_0 \sum_{i\in S_0} x_i^{(D)} x_{ij}^{(D-1)} - K_1 \sum_{i\in S_1} (1-x_i^{(D)}) x_{ij}^{(D-1)}.$$

Other derivatives, w.r.t. those further down in the network, can be recursively derived, albeit with more complicated expressions.

3.5 Decision Tree (DT)

The DT approach can be viewed as solving the (CE) problem by applying a specific "greedy heuristic" to finding the p_i values so as to minimize the CE objective.

To illustrate, observe from the CE objective function in (10), there are a total of $n(= n_0 + n_1)$ terms in the summation. Suppose all p_i's take on a single (common) value, then the natural choice would be $p_i = n_1/n(= \pi)$. This makes the CE objective value equal to

$$-n_0 \ln(1 - \frac{n_1}{n}) - n_1 \ln(\frac{n_1}{n}) = n\Big[- (1 - \frac{n_1}{n}) \ln(1 - \frac{n_1}{n}) - \frac{n_1}{n} \ln(\frac{n_1}{n})\Big]. \tag{17}$$

Next, we ask what if we allow all p_i's to take on two (distinct) values? One way to do this is to split the n data points into two disjoint subsets, according to some criterion based on say, the j-th component of x, such as, whether $x_{ij} \geq a$ or $x_{ij} < a$ for some specified value a. This will split the n data points into two disjoint subsets of sizes denoted t_j and t_j^c (c for "complement"). For either subset, we can further divide it into non-infected and infected, with the number counts denoted (t_{0j}, t_{1j}) and (t_{0j}^c, t_{1j}^c). We have

$$t_j + t_j^c = n; \quad t_{0j} + t_{0j}^c = n_0, \quad t_{1j} + t_{1j}^c = n_1.$$

Accordingly, we set the two p_i values as $\frac{t_{1j}}{t_j}$ and $\frac{t_{1j}^c}{t_j^c}$. This way, the CE objective value becomes

$$t_j\Big[- (1 - \frac{t_{1j}}{t_j}) \ln(1 - \frac{t_{1j}}{t_j}) - \frac{t_{1j}}{t_j} \ln(\frac{t_{1j}}{t_j})\Big] + t_j^c\Big[- (1 - \frac{t_{1j}^c}{t_j^c}) \ln(1 - \frac{t_{1j}^c}{t_j^c}) - \frac{t_{1j}^c}{t_j^c} \ln(\frac{t_{1j}^c}{t_j^c})\Big]. \tag{18}$$

Clearly, we will pick among all components (of x) the one (say, identified as the j-th) that yields the smallest CE value in (18), and we will only do the splitting if the CE value in (18) is smaller than the one in (17). If the splitting does occur, we will then apply the above procedure to each of the two terms in (18). This may result in four terms, or three terms (if one does not split), or remain as two terms (neither split).

Clearly, every time a split occurs, the CE value will decrease. On the hand, if a term does not split, it corresponds to a "leaf node" in the tree.

4 Convergence and Rate of Convergence

In this section we propose probabilistic bounds for the population *mis-classification cost problem* using the data-driven one. Our argument relies on the assumption that the support of the r.v. X is finite, meaning that X can take only finitely many values. This is certainly the case in our application as most of the variables in X are indicators for disease or discretized quantities such

as age. Our bounds holds for schemes ϕ in the family of all measurable functions (including deep neural networks and most families of parametric models).

Recall the "data-driven" version of the problem we want to solve in (2), as well as the original one in (1). Note that the objective function of the latter is a deterministic value (expectation), whereas its data-driven counterpart in (2), $\hat{\mathcal{C}}_n$, is a random quantity. What we want is the minimized objective function $\min_\phi \mathcal{C}(\phi) := \mathsf{E}\mathcal{C}(\phi^*)$ and the minimizing scheme ϕ^*; instead we will obtain ϕ_n^*, the scheme that minimizes the data-driven problem $\hat{\mathcal{C}}_n$. Note that when ϕ_n^* is applied, in lieu of ϕ^*, to $\mathcal{C}$, the latter becomes random, as ϕ_n^* is determined by the data $(x_i, y_i)_{i=1}^n$, i.i.d. copies of (X, Y). Specifically,

$$\mathcal{C}(\phi_n^*) = \mathsf{E}_{(X,Y)}\Big(\big[K_0(1-Y)\mathbf{1}\{p(X) \geq \rho\} + K_1 Y \mathbf{1}\{p(X) < \rho\}\big] \,|\, (x_i, y_i)_{i=1}^n\Big), \tag{19}$$

in particular, $p(\cdot)$ and ρ are both determined by $(x_i, y_i)_{i=1}^n$.

We can bound the gap between the two as follows:

$$\begin{aligned} 0 &\leq \mathcal{C}(\phi_n^*) - \mathcal{C}(\phi^*) \\ &= \mathcal{C}(\phi_n^*) - \hat{\mathcal{C}}_n(\phi_n^*) + \hat{\mathcal{C}}_n(\phi_n^*) - \mathcal{C}(\phi^*) \\ &\leq \mathcal{C}(\phi_n^*) - \hat{\mathcal{C}}_n(\phi_n^*) + \hat{\mathcal{C}}_n(\phi^*) - \mathcal{C}(\phi^*) \\ &\leq 2 \sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)|, \end{aligned} \tag{20}$$

where the second last inequality takes into account $\hat{\mathcal{C}}_n(\phi_n^*) \leq \hat{\mathcal{C}}_n(\phi^*)$, as ϕ_n^* is the scheme that minimizes $\hat{\mathcal{C}}_n$; the other inequalities are obvious. (See Lemma 8.2 of Devroye et al. 1996.)

Thus, it suffices to bound $\sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)|$. Observe that when both $\hat{\mathcal{C}}_n$ and $\mathcal{C}$ are driven by the same scheme ϕ, the difference is then a sum of i.i.d. terms; indeed each term is mean zero random variable.

To continue, let's take a closer look at the two expectations of $\mathcal{C}$ in (1):

$$\mathsf{E}\big[Y\mathbf{1}\{p(X) \leq \rho\}\big] = \mathsf{E}\mathbf{1}\{Y = 1, p(X) \leq \rho\} = \mathsf{P}(Y)\mathsf{P}\big(p(X) \leq \rho | Y = 1\big) = \pi \mathsf{P}\big[p(X) \leq \rho | Y = 1\big],$$

and similarly,

$$\mathsf{E}\big[(1-Y)\mathbf{1}\{p(X) > \rho\}\big] = (1-\pi)\mathsf{P}(p(X) > \rho | Y = 0).$$

In the same spirit, breaking into $y_i = 1$ (i.e., $i \in S_1$) and $y_i = 0$ (i.e., $i \in S_0$), $\hat{\mathcal{C}}_n$ can be expressed as:

$$\frac{K_0}{n} \sum_{i \in S_0} \mathbf{1}\{p(x_i) > \rho\} + \frac{K_1}{n} \sum_{i \in S_1} \mathbf{1}\{p(x_i) \leq \rho\}.$$

Clearly, we have

$$\begin{aligned} \mathsf{P}\big[p(X) \leq \rho | Y = 0\big] &= \mathsf{E}\mathbf{1}\{p(x_i) > \rho | i \in S_0\} := \mathsf{E}(I_i^0(\phi)), \\ \mathsf{P}\big[p(X) \leq \rho | Y = 1\big] &= \mathsf{E}\mathbf{1}\{p(x_i) \leq \rho | i \in S_1\} := \mathsf{E}(I_i^1(\phi)), \end{aligned}$$

where ϕ refers to $p(\cdot)$ and ρ.

Hence, taking into account $\pi = n_1/n$ and $1-\pi = n_0/n$, we have,

$$\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi) = \frac{K_0 n_0}{n}\frac{1}{n_0}\sum_{i=1}^{n_0}\big[I_i^0(\phi)) - \mathsf{E}(I_i^0(\phi))\big] + \frac{K_1 n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1}\big[I_i^1(\phi)) - \mathsf{E}(I_i^1(\phi))\big]$$

Thus,

$$\begin{aligned}
& \mathsf{P}\Big(\sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)| > \frac{\epsilon}{2}\Big) \\
= \;& \mathsf{P}\Big(\sup_\phi |\frac{K_0}{n}\sum_{i=1}^{n_0}\big[I_i^0(\phi) - \mathsf{E}(I_i^0(\phi))\big] + \frac{K_1}{n}\sum_{i=1}^{n_1}\big[I_i^1(\phi) - \mathsf{E}(I_i^1(\phi))\big]| > \frac{\epsilon}{2}\Big) \\
\leq \;& \mathsf{P}\Big(\sup_\phi |\frac{K_0}{n}\sum_{i=1}^{n_0}\big[I_i^0(\phi) - \mathsf{E}(I_i^0(\phi))\big]| + \sup_\phi |\frac{K_1}{n}\sum_{i=1}^{n_1}\big[I_i^1(\phi) - \mathsf{E}(I_i^1(\phi))\big]| > \frac{\epsilon}{2}\Big) \\
\leq \;& \mathsf{P}\Big(\{\sup_\phi |\frac{K_0}{n}\sum_{i=1}^{n_0}\big[I_i^0(\phi) - \mathsf{E}(I_i^0(\phi))\big]| > \frac{\epsilon}{4}\} \cup \{\sup_\phi |\frac{K_1}{n}\sum_{i=1}^{n_1}\big[I_i^1(\phi) - \mathsf{E}(I_i^1(\phi))\big]| > \frac{\epsilon}{4}\}\Big) \\
\leq \;& \mathsf{P}\Big(\sup_\phi |\frac{K_0}{n}\sum_{i=1}^{n_0}\big[I_i^0(\phi) - \mathsf{E}(I_i^0(\phi))\big]| > \frac{\epsilon}{4}\Big) + \mathsf{P}\Big(\sup_\phi |\frac{K_1}{n}\sum_{i=1}^{n_1}\big[I_i^1(\phi) - \mathsf{E}(I_i^1(\phi))\big]| > \frac{\epsilon}{4}\Big)
\end{aligned} \tag{21}$$

.

[In the above, the first inequality follows from triangular inequality and the subadditivity of sup; the second one from simple event implication ($a+b>\epsilon \Rightarrow a$ and b cannot be both $\leq \epsilon/2$, i.e., we must have either $a > \epsilon/2$ or $b > \epsilon/2$); and the last inequality from bounding union by sum.]

Without loss of generality the probabilistic object we want to bound is:

$$\sup_{\phi\in\Phi} |\frac{1}{n}\sum_{i=1}^{n}\big[I_i(\phi) - \mathsf{E}(I_i(\phi))\big]|. \tag{22}$$

Denote $S(X)$ as the support of X. In the case when $|S(X)| = k$ there is a simplified way to handle the supremum over all measurable functions Φ. Note that the classification function $\phi(x) = \mathbf{1}\{p(x) > \rho\}$ is a map from $S(X)$ to $\{0,1\}$. Since X can take up to k different values then there is also finite number of maps from $\{1,\cdots,k\}$ to $\{0,1\}$. All these maps can be captured by the power set $\{0,1\}^k$, which then makes that the set Φ of all measurable functions is actually the power set represented by all the vector of zeros and ones $\big(\phi(1),\cdots,\phi(k)\big)$ of size k.

Then, X can be seen as a multinomial random variable where each value of $S(X)$ gets mapped to an integer in $\{1,\cdots,k\}$. An i.i.d. sample of size n from X will be a multinomial vector $(n_1,\cdots,n_k)$ with n_1 observations taking value 1, n_2 observations taking value 2, and so on. We have $\sum_{j=1}^k n_j = n$. The *unknown* law of the multinomial is given by the vector $\boldsymbol{p} = (p_1, p_2, \cdots, p_k)$.

Assuming all samples belong to S_0 (in the same way we splitted S_0 and S_1 using the union bound). We abuse notation by denoting $\phi(j) = 1$ is the indicator for misclassification when $X = j$.

Any vector $\big(\phi(1), \cdots, \phi(k)\big)$ will have empirical mis-classification cost $\frac{1}{n}\sum_{j=1}^{k}\phi(j)n_j$ and population mis-classification cost $\sum_{j=1}^{k}\phi(j)p_j$. The furthest away the empirical cost can be from the real value is:

$$\sup_{\phi\in\Phi}|\sum_{j=1}^{k}\phi(j)\left[\frac{n_j}{n}-p_j\right]|, \tag{23}$$

To solve this problem note that $\left[\frac{n_j}{n}-p_j\right]$ will be either positive, negative or zero. To maximize the absolute value of the sum the solution is to select either all the positive values or all the negative values (note that the sum of the positive values will be equal to the sum of the negative values as both $\sum_j \frac{n_j}{n}$ and $\sum_j p_j$ add up to 1 each) and make all the other ones zero by selecting a ϕ that is equal to 1 in the positive indices. We can do this as the family Φ is the power set $\{0,1\}^k$.

Let σ be a permutation of $\{1,\cdots,k\}$ such that all positive elements of $\left[\frac{n_j}{n}-p_j\right]$ are before the negative ones. For example, according to the descending ordering of $\left[\frac{n_j}{n}-p_j\right]$, such that, $\sigma(1)$ is the element j with largest $\left[\frac{n_j}{n}-p_j\right]$. This ordering naturally depends on the sample, and we will have to condition on the ordering σ as we will see.

Note that once the elements are ordered according to σ we have that the solution of the problem is just the partial sum according to the ordering σ until the point the sum does not increase anymore:

$$\sup_{\phi\in\Phi}|\frac{1}{n}\sum_{i=1}^{n_1}\left[I_i^1(\phi)-\mathsf{E}(I_i^1(\phi))\right]| = \sup_{\phi\in\Phi}|\sum_{j=1}^{k}\phi(j)\left[\frac{n_j}{n}-p_j\right]| = \max_{h\in\{1,\cdots,k\}}|\sum_{j=1}^{h}\left[\frac{n_{\sigma(j)}}{n}-p_{\sigma(j)}\right]|. \tag{24}$$

Finally, conditional on σ the last object is a discretized version of $\sup_{x\in[0,1]}|\hat{F}_n(x)-F(x)|$, that is, the known *Kolmogorov's distribution* (see Shiryayev 2012, it can be showed that a discrete kolmogorov distribution is smaller in the stochastic order than the continuous one, for a trivial proof, consider that for any discrete distribution there exist a continuous one that matches the cdf at every discrete point, therefore, the supremum of the continuous one is at least the supremum of the discrete one). The problem is that σ is not independent of the sample, therefore we have using the union bound:

$$\mathsf{P}\left(\max_{h\in\{1,\cdots,k\}}|\sum_{j=1}^{h}\left[\frac{n_{\sigma(j)}}{n}-p_{\sigma(j)}\right]|>\epsilon\right) \leq \mathsf{P}\left(\sup_{x\in[0,1]}|\hat{F}_n(x)-F(x)|>\epsilon\right)N(\sigma). \tag{25}$$

Where $N(\sigma)$ is the number of orderings with different value for the r.h.s of (24). Note that the total number of possible orderings σ is $k!$, yet for many of these orderings the value of (24) is identical. Letting $h^* = \text{argmax}_{h\in\{1,\cdots,k\}}|\sum_{j=1}^{h}\left[\frac{n_{\sigma(j)}}{n}-p_{\sigma(j)}\right]|$ any rearrangement of the ordering σ left of $\sigma(h^*+1)$ or right of $\sigma(h^*)$ will have the same value for the maximum of the r.h.s of (24) as rearranging the contiguous positive or negative values doesn't change the maximum of the sum, and therefore a single sequence σ accounts for the maximum of the sum over all right and left rearrangements of contiguous elements of the same sign. Then, for $h^*=2$ there are $\binom{k}{2}$ orderings σ

with different value for the maximum of the sum, for $h^* = 3$ there are $\binom{k}{3}$ and so on, therefore we get a total number of orderings with different value of the supremum equal to the sum of binomial coefficients, using the binomial theorem:

$$N(\sigma) = \binom{k}{1} + \binom{k}{2} + \cdots + \binom{k}{k} \leq 2^k, \tag{26}$$

Gathering these facts together yields

$$\mathsf{P}\left(\sup_{\phi\in\Phi} |\frac{1}{n}\sum_{i=1}^{n}\big[I_i(\phi) - \mathsf{E}(I_i(\phi))\big]| > \epsilon\right) \leq \mathsf{P}\left(\sup_{x\in[0,1]} |\hat{F}_n(x) - F(x)| > \epsilon\right) 2^k, \tag{27}$$

Which in turn can be bounded by the bound proposed by Dvoresky-Kiefer-Wolfowitz (see Dvoretzky et al. 1956) and further refined by Massart 1990 of the Kolmogorov's distribution. From which we obtain:

$$\mathsf{P}\left(\sup_{\phi\in\Phi} |\frac{1}{n}\sum_{i=1}^{n}\big[I_i(\phi) - \mathsf{E}(I_i(\phi))\big]| > \epsilon\right) \leq 2^{k+1}\exp\big(-2n\epsilon^2\big). \tag{28}$$

A similar approach using Vapnik-Chervonenkis bounds (see Devroye et al. 1996) yields the same constant next to the exponential factor. Nonetheless, the coefficient in the exponential of our bound is better than the denominator of 32 in the V-C bound.

Taking into account $\pi := n_1/n$, we have

$$\mathsf{P}\Big(\sup_{\phi} |\frac{K_1 n_1}{n}\frac{1}{n_1}\sum_{i=1}^{n_1}\big[I_i^1(\phi) - \mathsf{E}(I_i^1(\phi))\big]| > \frac{\epsilon}{4}\Big) \leq 2^{k+1}\exp\Big(-\frac{n_1 n^2\epsilon^2}{8K_1^2 n_1^2}\Big) = 2^{k+1}\exp\Big(-\frac{n\epsilon^2}{8K_1^2\pi}\Big). \tag{29}$$

Analogously, the other probability can be bounded as follows:

$$\mathsf{P}\Big(\sup_{\phi} |\frac{K_0 n_0}{n}\frac{1}{n_0}\sum_{i=1}^{n_0}\big[I_i^0(\phi) - \mathsf{E}(I_i^0(\phi))\big]| > \frac{\epsilon}{4}\Big) \leq 2^{k+1}\exp\Big(-\frac{n\epsilon^2}{8K_0^2(1-\pi)}\Big). \tag{30}$$

Since $K_1 >> K_0$, we expect $K_1^2\pi \geq K_0^2(1-\pi)$. So, the bound in (29) is the larger one (otherwise pick the maximum).

Combining the above with the inequalities in (20) and (21), we have

$$\begin{aligned} &\mathsf{P}\Big(|\mathcal{C}(\phi_n^*) - \mathcal{C}(\phi^*)| > \epsilon\Big) \leq \mathsf{P}\Big(\sup_{\phi} |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)| > \frac{\epsilon}{2}\Big) \\ \leq\ & 2\mathsf{P}\Big(\sup_{\phi} |\frac{K_1}{n}\sum_{i=1}^{n_1}\big[I_i^1(\phi) - \mathsf{E}(I_i^1(\phi))\big]| > \frac{\epsilon}{4}\Big) \leq 2^{k+2}\exp\Big(-\frac{n\epsilon^2}{8K_1^2\pi}\Big). \end{aligned} \tag{31}$$

Also of interest is to bound the difference $\hat{\mathcal{C}}_n(\phi_n^*) - \mathcal{C}(\phi^*)$: having optimized the data-driven version (of $\mathcal{C}$) in (2), we want to know how much it deviates from optimizing the original objective $\mathcal{C}$ in (1). We have, from $\hat{\mathcal{C}}_n(\phi_n^*) \leq \hat{\mathcal{C}}_n(\phi^*)$,

$$\hat{\mathcal{C}}_n(\phi_n^*) - \mathcal{C}(\phi^*) \leq \hat{\mathcal{C}}_n(\phi^*) - \mathcal{C}(\phi^*) \leq \sup_{\phi} |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)|;$$

and, the other way around, from $\mathcal{C}(\phi^*) \le \mathcal{C}(\phi_n^*)$,

$$\mathsf{E}\mathcal{C}(\phi^*) - \hat{\mathcal{C}}_n(\phi_n^*) \le \mathcal{C}(\phi_n^*) - \hat{\mathcal{C}}_n(\phi_n^*) \le \sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)|.$$

Putting the two together, we have

$$|\hat{\mathcal{C}}_n(\phi_n^*) - \mathcal{C}(\phi^*)| \le \sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)|. \tag{32}$$

Thus, similar to (31), we have

$$\mathsf{P}\Big(|\hat{\mathcal{C}}_n(\phi_n^*) - \mathcal{C}(\phi^*)| > \epsilon\Big) \le \mathsf{P}\Big(\sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)| > \epsilon\Big) \le 2^{k+2} \exp\Big(-\frac{n\epsilon^2}{2K_1^2\pi}\Big). \tag{33}$$

Note, from (31), the RHS is summable (over n); hence, following Borel-Cantelli Lemma, and taking into account that $\mathcal{C}(\phi)$ is bounded (by $\bar{K} := \min\{K_0(1-\pi), K_1\pi\}$), we have

$$\mathcal{C}(\phi_n^*) \to \mathcal{C}(\phi^*) \quad \text{a.s.} \quad \text{and} \quad \text{in } L_p \ (p \ge 1). \tag{34}$$

Similarly, from (33), we can conclude

$$\hat{\mathcal{C}}_n(\phi_n^*) \to \mathcal{C}(\phi^*) \quad \text{a.s.} \quad \text{and} \quad \text{in } L_p \ (p \ge 1). \tag{35}$$

Finally, suppose we have obtained the optimal scheme ϕ_n^* as above (for a sufficiently large n), and then continue to apply this scheme on another set of i.i.d. data $(x_i, y_i)_{i=1}^N$. This results in the following cost:

$$\hat{\mathcal{C}}_N(\phi_n^*) := \frac{1}{N}\sum_{i=1}^N \big[K_0(1-y_i)\mathbf{1}\{p(x_i,\phi_n^*) \ge \rho(\phi_n^*)\} + K_1 y_i \mathbf{1}\{p(x_i,\phi_n^*) < \rho(\phi_n^*)\}\big], \tag{36}$$

where we have explicitly written out the dependence on the scheme ϕ_n^*. By SLLN, we have, as $N \to \infty$, $\hat{\mathcal{C}}_N(\phi_n^*) \to \mathsf{E}C(\phi_n^*)$ a.s. (and in L_p too). This, together with the L_1 convergence in (34), which implies $\mathsf{E}\mathcal{C}(\phi_n^*) \to \mathcal{C}(\phi^*)$, leads to, for sufficiently large N (as well as sufficiently large n),

$$|\hat{\mathcal{C}}_N(\phi_n^*) - \mathcal{C}(\phi^*)| \le |\hat{\mathcal{C}}_N(\phi_n^*) - \mathsf{E}\mathcal{C}(\phi_n^*)| + |\mathsf{E}\mathcal{C}(\phi_n^*) - \mathcal{C}(\phi^*)| \le \epsilon, \quad \forall\, \epsilon > 0. \tag{37}$$

That is, $\hat{\mathcal{C}}_N(\phi_n^*) \to \mathsf{E}\mathcal{C}(\phi^*)$ a.s., as $N \to \infty$ (as well as $\hat{\mathcal{C}}_N(\phi_n^*) \to \mathsf{E}\mathcal{C}(\phi_n^*)$ a.s. as argued above).

This is reassuring. It means: once the ML scheme is optimized for a sufficiently large data set (of size n), the scheme can be applied to other data sets — provided the data are i.i.d. and follow the same distribution (X, Y) as the original set does — and still maintain optimal performance. Furthermore, the rate of convergence is similar to what's analyzed above: Pick n to be sufficiently large such that

$$|\mathsf{E}\mathcal{C}(\phi_n^*) - \mathcal{C}(\phi^*)| \le \epsilon/2. \tag{38}$$

Then, combining the above with (37), we have

$$\mathsf{P}\Big(|\hat{\mathcal{C}}_N(\phi_n^*) - \mathcal{C}(\phi^*)| > \epsilon\Big) \leq \mathsf{P}\Big(|\hat{\mathcal{C}}_N(\phi_n^*) - \mathsf{E}\mathcal{C}(\phi_n^*)| > \frac{\epsilon}{2}\Big) \leq \mathsf{P}\Big(|Z| > \frac{\epsilon\sqrt{N}}{2\hat{\sigma}}\Big), \tag{39}$$

where Z follows the standard normal distribution, and

$$\hat{\sigma}^2 := K_0^2(1-\pi) + K_1^2\pi. \tag{40}$$

The bound in (39) follows from CLT, because with the scheme fixed at ϕ_n^*, $\hat{\mathcal{C}}_N(\phi_n^*)$ involves a sum of i.i.d. random variables, so CLT applies. Each term under the summation has a zero mean. To derive its variance, denoted $\sigma^2(\phi)$, write

$$B^0(\phi) := (1-Y)\mathbf{1}\{p(X) \geq \rho\} \quad \text{and} \quad B^1(\phi) := Y\mathbf{1}\{p(X) < \rho\},$$

which are Bernoulli variables, and denote their means as θ_0 and θ_1. Then,

$$\sigma^2(\phi) = \mathsf{Var}[K_0B^0(\phi) + K_1B^1(\phi)] = K_0^2\theta_0(1-\theta_0) + K_1^2\theta_1(1-\theta_1) - 2K_0K_1\theta_0\theta_1,$$

taking into account $B^0 \cdot B^1 = 0$, and hence $\mathsf{Cov}(B^0, B^1) = 0 - \mathsf{E}(B^0)\mathsf{E}(B^1) = -\theta_0\theta_1$. Thus,

$$\sigma^2(\phi) = K_0^2\theta_0 + K_1^2\theta_1 - (K_0\theta_0 + K_1\theta_1)^2 \leq \hat{\sigma}^2,$$

taking into account $\theta_1 \leq \mathsf{E}(Y) := \pi$ and $\theta_0 \leq \mathsf{E}(1-Y) = 1-\pi$.

To compare the above with the bound in (31), making use of $\mathsf{P}(|Z| > x) \approx (\leq)\, 2\phi(x)/x$ (where $\phi(x)$ denotes the pdf of Z), we have

$$\mathsf{P}\Big(|Z| > \frac{\epsilon\sqrt{N}}{2\hat{\sigma}}\Big) \approx (\leq)\, \frac{2\hat{\sigma}}{\epsilon\sqrt{N}} \exp\Big(-\frac{N\epsilon^2}{8\hat{\sigma}^2}\Big). \tag{41}$$

What remains is the question how large n should be to satisfy the inequality in (38). To this end, observe that it suffices to have $\mathsf{E}|\mathcal{C}(\phi_n^*) - \mathcal{C}(\phi^*)| \leq \epsilon/2$, since the LHS dominates the LHS of the inequality in (38), following Jensen's inequality. Write $\mathcal{C}_n := \mathcal{C}(\phi_n^*)$ and $\mathcal{C}^* := \mathcal{C}(\phi^*)$ to lighten notation. We have

$$\begin{aligned}
\mathsf{E}|\mathcal{C}_n - \mathcal{C}^*| &= \mathsf{E}\Big(|\mathcal{C}_n - \mathcal{C}^*|\mathbf{1}\{|\mathcal{C}_n - \mathcal{C}^*| \leq \frac{\epsilon}{4}\}\Big) + \mathsf{E}\Big(|\mathcal{C}_n - \mathcal{C}^*|\mathbf{1}\{|\mathcal{C}_n - \mathcal{C}^*| > \frac{\epsilon}{4}\}\Big) \\
&\leq \frac{\epsilon}{4} + \bar{K}\mathsf{P}\Big(|\mathcal{C}_n - \mathcal{C}^*| > \frac{\epsilon}{4}\Big),
\end{aligned}$$

where we have made use of

$$|\mathcal{C}_n - \mathcal{C}^*| = \mathcal{C}_n - \mathcal{C}^* \leq \mathcal{C}_n \leq \bar{K} := \min\{K_0(1-\pi), K_1\pi\}.$$

Thus, to have $\mathsf{E}|\mathcal{C}(\phi_n^*) - \mathcal{C}(\phi^*)| \leq \epsilon/2$, it suffices to have

$$\bar{K}\mathsf{P}\Big(|\mathcal{C}_n - \mathcal{C}^*| > \frac{\epsilon}{4}\Big) \leq \frac{\epsilon}{4}.$$

Applying the bound in (31), replacing ϵ there by $\epsilon/4$, the above inequality is implied by

$$2^{k+2}\bar{K}\exp\Big(-\frac{n\epsilon^2}{128K_1^2\pi}\Big)\leq\frac{\epsilon}{4},$$

which lead to

$$n\geq\frac{128K_1^2\pi}{\epsilon^2}\ln(\frac{2^{k+3}\bar{K}}{\epsilon}). \tag{42}$$

All the results above are summarized in the following theorem.

Theorem 1 Let ϕ^* be the optimal solution to (1) and ϕ_n^* be the optimal solution to the "data-driven" version in (2), with $\mathcal{C}(\phi^*)$ and $\hat{\mathcal{C}}_n(\phi_n^*)$ denoting the corresponding objective functions, in the case of finite support of size k, for any family of models ϕ:

(i) $\hat{\mathcal{C}}_n(\phi_n^*)\to\mathcal{C}(\phi^*)$ a.s.; and the rate of convergence is exponential, following (33).

(ii) In addition, $\mathcal{C}(\phi_n^*)\to\mathcal{C}(\phi^*)$ a.s.; and the rate of convergence is also exponential, following (31).

(iii) Suppose the solution ϕ_n^*, for a sufficiently large n satisfying (42), is applied to another data set of size N, with the data being i.i.d. and following the same distribution as the original set, and denoting the corresponding cost as $\hat{\mathcal{C}}_N(\phi_n^*)$ following (36). Then, $\hat{\mathcal{C}}_N(\phi_n^*)\to\mathsf{E}\mathcal{C}(\phi^*)$ a.s. (as $N\to\infty$); and the rate of convergence is also exponential, following (39).

(iv) All the convergence results above hold in L_p as well, for any $p\geq 1$.

5 Numerical Studies

As a case study, we illustrate the methodology developed in previous sections for prediction and prevention of Catheter Associated Urinary Tract Infection (CAUTI) using different machine learning models. We assess both the cost performance and prediction quality in and out-of-sample in light of our results.

The expected cost for HAI prevention can be expressed as a mis-classification problem the following way:

$$\min_{(\rho,w)}\mathsf{E}_{(X,Y)}\big[C_0(1-Y)\mathbf{1}\{p(X)\geq\rho\}+C_1Y\mathbf{1}\{p(X)<\rho\}+\big(\lambda C_0+(1-\lambda)(C_0+C_1)\big)Y\mathbf{1}\{p(X)\geq\rho\}\big] \tag{43}$$

The first two terms are analogous to the mis-classification cost presented in the previous sections, where there is a preventive measure with cost C_0 and an infection with cost C_1, the last term means that out of the correctly classified positive cases, the preventive measure works with an independent success probability λ. Noting that $\mathbf{1}\{p(X)\geq\rho\}=1-\mathbf{1}\{p(X)<\rho\}$ the objective can be expressed in terms of Equation 1 by letting $K_2=\big(\lambda C_0+(1-\lambda)(C_0+C_1)\big)$, $K_1=\lambda C_1-C_0$ and $K_0=C_0$. The expected cost for the hospital becomes:

$$\min_{(\rho,p)}\mathsf{E}_{(X,Y)}\big[K_0(1-Y)\mathbf{1}\{p(X)\geq\rho\}+K_1Y\mathbf{1}\{p(X)<\rho\}\big]+\mathsf{E}[Y]K_2 \tag{44}$$

The previous cost must be compared with the deterministic strategies of giving the prevention to every patient or giving no prevention at all:

$$\mathsf{E}[C_0 + Y(1-\lambda)C_1] \wedge \mathsf{E}[YC_1] = \big(\mathsf{E}[(1-Y)K_0] \wedge \mathsf{E}[YK_1]\big) + \mathsf{E}[Y]K_2 \tag{45}$$

Which we will call *base cost*. The base cost will serve as a benchmark for the machine learning model as it would make no sense to implement the machine learning scheme if it has higher cost in expected value, of say, giving prevention to every patient. This is also where the confidence bounds will be useful as they will they with show if the cost of implementing the machine learning scheme is better than implementing a deterministic strategy. To maintain resemblance to the mis-classification cost setting of previous sections we will subtract $\mathsf{E}[Y]K_2$ from the hospital cost and the base cost to ease comparison.

5.1 Data

We use non identifiable individual patient data from 2009 to 2016 collected at NewYorkPresbiterian Hospital. The patient variables used for prediction (available at day 1 of admission) collected in a vector x_i are:

Charlson Comorbidity score [integer], 3M[1] severity of illness (SOI) [integer] and risk of mortality (ROM) [integer] scores averaged over three most recent admissions. As well as age[integer], history of diabetes mellitus [indicator], dialysis [indicator], organ transplant [indicator], malignancy [indicator],admission through the emergency room [indicator], previous HAI episodes [integer], and previous number of visits [integer]. All variables are based on ICD-9/ICD-10 codes.

Using these variables as inputs x_i we optimize the mis-classification cost model $\hat{\mathcal{C}}(\phi_n^*)$ for the remaining of the section and compare its performance with the baseline cost of the hospital in Equation (45).

As parameters for the model we use $C_1 = \$896$ for the cost of infection, $C_0 = \$16$ as the cost of prevention, and $\lambda = 69\%$ as the success rate of the prevention measure.

5.2 Comparing CE Minimization vs the Two-Step Procedure

Here we compare the performance of optimizing only the CE objective (as commonly done in practice) with the proposed two step algorithm for the logit model as solution to the mis-classification cost problem $\hat{\mathcal{C}}(\phi_n^*)$. In Figure (2) and Table (4) the mis-classification cost $\hat{\mathcal{C}}_n(\phi_n^*)$ is computed for different sample sizes n. The two-step procedure outperforms only optimizing the CE objective by around 9%, making it comparable with the performance of the Deep Neural Network Model in the next subsections.

[1]Proprietary assessments of severity of illness and risk of mortality used by the hospital

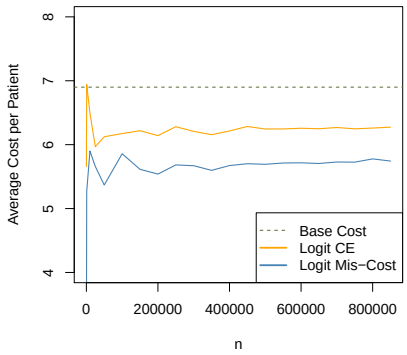


Figure 2: Mis-classification cost $\hat{\mathcal{C}}(\phi_n^*)$ comparison between optimizing only the CE objective and our proposed two-step procedure for the logit model.

5.3 Convergence and Rate of Convergence

Using the two-step optimization scheme outlined in the previous section we minimize the mis-classification cost $\hat{\mathcal{C}}_n(\phi_n^*)$ using logistic regression, deep neural network and decision tree models. For the deep neural network, two hidden layers of size 32 and 16 are used with sigmoid activations in all the layers, including the output one, other activation functions and architectures were tested achieving similar results. For the decision tree, we constraint the trees to be no deeper than 7 levels.

A key question when training a model is to decide at what sample size n should the training be stopped with some confidence on the performance of the model. At the end of Section 4 in Theorem 1 and specifically in Equation (42) the training sample size n that guarantees that the "out-of-sample" performance is within ϵ of the true optimal solution $\mathcal{C}(\phi^*)$ is given explicitly. In practice, this estimate is too conservative (as it applies for the family of all measurable functions Φ and all distributions of (X, Y)), and in this section we will show that in practice convergence seems to occur at a faster rate on n.

As a first example to test convergence, we divide our dataset into 10 equal-sized non-overlapping partitions each of size $89,734$ and estimate the sample mean $\hat{\mu} := \mathsf{E}\,\hat{\mathcal{C}}_n(\phi_n^*)$ and sample variance $\hat{\sigma}^2 := \mathsf{Var}\,\hat{\mathcal{C}}_n(\phi_n^*)$ of independent realizations of $\hat{\mathcal{C}}_n(\phi_n^*)$ for each partition. Theorem 1 implies that $\mathsf{E}\,\hat{\mathcal{C}}_n(\phi_n^*) \to \mathcal{C}(\phi^*)$ meaning that $\mathsf{E}\,\mathcal{C}_n(\phi_n^*)$ also converges to the optimal solution. Then, independent draws of $\mathcal{C}_n(\phi_n^*)$ (using the non-overlapping partitions of data) should converge in average to the optimal solution $\mathcal{C}(\phi^*)$ and their variance should be decreasing with n. In Table 5 and Figure 3 we can see how the mean of the process stabilizes around 5.7 while the variance $\hat{\sigma}^2$ decreases with n, implying a fast and comparable rate of convergence to the a.s. limit for all models. As n increases the mean $\hat{\mu}$ seems to stabilize around 5.7 after $n = 50,000$. Note that for all models the interval

n	$\hat{\mathcal{C}}_n(\phi_n^*)$ (CE)	$\hat{\mathcal{C}}_n(\phi_n^*)$ (Two-step)
50,000	6.12	5.37
100,000	6.17	5.86
150,000	6.22	5.61
200,000	6.14	5.54
250,000	6.28	5.68
300,000	6.21	5.67
350,000	6.16	5.60
400,000	6.21	5.67
450,000	6.28	5.70
500,000	6.25	5.69
550,000	6.25	5.71
600,000	6.26	5.72
650,000	6.25	5.71
700,000	6.27	5.73
750,000	6.25	5.73
800,000	6.26	5.78
850,000	6.27	5.74

Table 4: Mis-classification cost comparison between optimizing only the CE objective and the proposed two-step procedure for the logit model.

$\hat{\mu} \pm 1.96\sigma$ is well below the base cost for the hospital, implying that the performance of the ML model is better than the base-line cost in around 17%.

n	5,000	10,000	25,000	50,000	89,734	
$\hat{\mu}$	5.01	5.29	5.56	5.78	5.74	Logit
$\hat{\sigma}^2$	0.099	0.178	0.131	0.087	0.054	
$\hat{\mu}$	5.41	5.48	5.66	5.66	5.73	DNN
$\hat{\sigma}^2$	0.146	0.204	0.092	0.078	0.054	
$\hat{\mu}$	4.69	5.08	5.46	5.59	5.75	DT
$\hat{\sigma}^2$	0.253	0.096	0.054	0.080	0.053	

Table 5: Convergence of $\mathsf{E}\,\hat{\mathcal{C}}_n(\phi_n^*)$ for different paths of $\hat{\mathcal{C}}_n(\phi_n^*)$ for the logit, deep neural network and decision tree models using the two-step procedure for optimization.

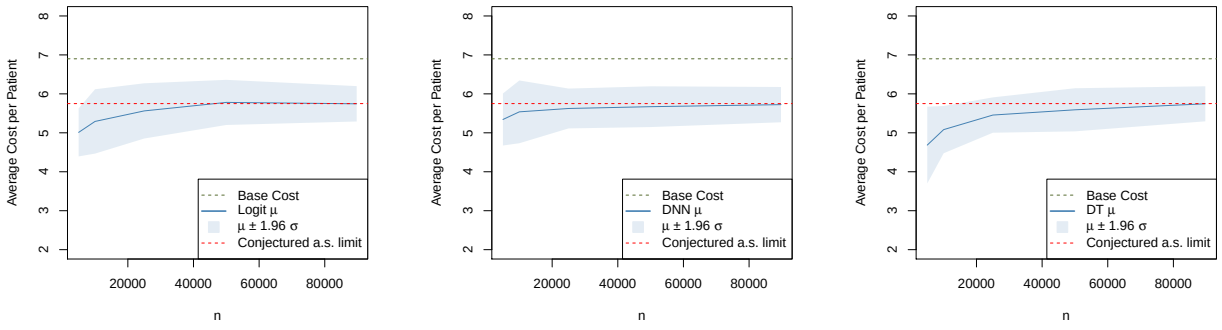


Figure 3: Convergence of $\mathsf{E}\,\hat{\mathcal{C}}_n(\phi_n^*)$ for different paths of $\hat{\mathcal{C}}_n(\phi_n^*)$ for the logit, deep neural network and decision tree models using the two-step procedure for optimization.

In summary, an empirical way to check the convergence of $\hat{\mathcal{C}}_n(\phi_n^*)$ is:

- Given $\epsilon > 0$ check the convergence of $\mathsf{E}\hat{\mathcal{C}}_n(\phi_n^*)$ as a function of n. Pick n_0 such that $\mathsf{E}\hat{\mathcal{C}}_n(\phi_n^*)$ is a cauchy sequence for $n > n_0$ for the given ϵ.
- Given a confidence level α, and stop training at $n^* > n_0$ such that $\mathsf{Var}\,\hat{\mathcal{C}}_n(\phi_n^*) \leq \alpha$.

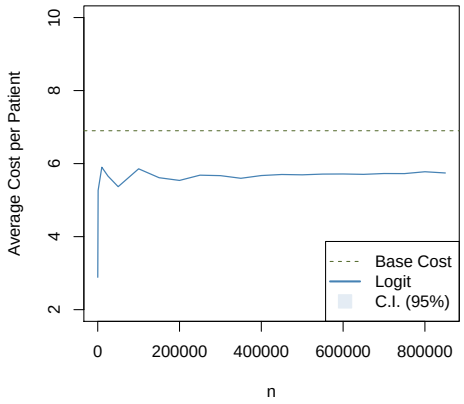
Following this procedure, we stop training at $n = 250,000$ where $\mathsf{E}\hat{\mathcal{C}}_n(\phi_n^*)$ has converged for all models and the variance $\mathsf{Var}\,\hat{\mathcal{C}}_n(\phi_n^*)$ is below 10^{-3}. We further examine convergence by analyzing the performance "out-of-sample" in the next section.

5.4 Comparison of the Models

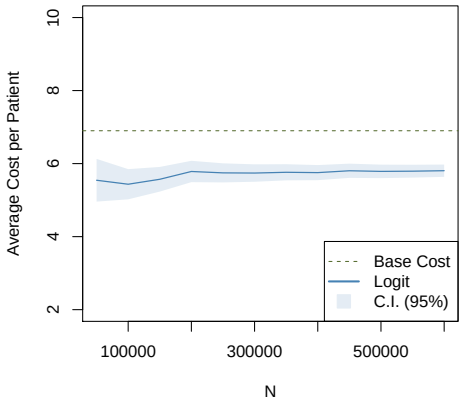
According to Theorem 1, after stopping training at a suitable n where $\hat{\mathcal{C}}_n(\phi_n^*)$ is within $\frac{\epsilon}{2}$ of its a.s. limit $\mathcal{C}(\phi^*)$. The performance of testing the model on a testing dataset of size N, that is, $\hat{\mathcal{C}}_N(\phi_n^*)$ should also converge within ϵ of $\mathcal{C}(\phi^*)$ for a large enough N. In this section, we test this convergence for single paths in and out-of-sample.

In Table 6 we present the training mis-classification cost $\hat{\mathcal{C}}_n(\phi_n^*)$ for different values of n. As n increases, the sample mis-classification cost $\hat{\mathcal{C}}_n(\phi_n^*)$ converges to its a. s. limit for each model, all around 5.7. The accuracy of the models is around 89%. It is important to note that the accuracy of the model in itself doesn't give the full picture of the performance of the model captured by the mis-classification cost (as predicting all patients as 0 implies an accuracy of $(1 - \pi) \approx 98\%$).

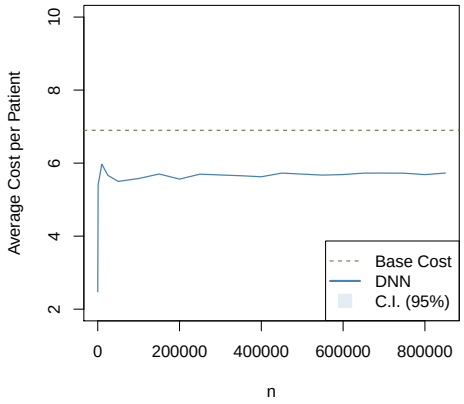
In Table 7 we present similar results after stopping training at $n = 250,000$ on the rest of the data as test dataset. In Figure 4 we present the trajectories of the training mis-classification cost $\hat{\mathcal{C}}_n(\phi_n^*)$ and the test mis-classification cost $\hat{\mathcal{C}}_N(\phi_n^*)$ after stopping the training at $n = 250,000$ using the analysis in the previous subsection. The confidence bounds for the test mis-classification cost $\hat{\mathcal{C}}_N(\phi_n^*)$ are based on Equation (39), and as we can see they all guarantee that the performance of the machine learning scheme is below the base cost (dotted line), which is the deterministic strategy of either giving prevention to everyone or to no-one, whichever is less costly. This is an important use of the methodology as it gives a simple criterion to assess the performance of the model and allows to help decision making on whether to implement the model or not. Also, the sample training and test mis-classification cost for each model seem to be relatively close to its a.s. limit for each model around 5.7 as conjectured in the previous section.



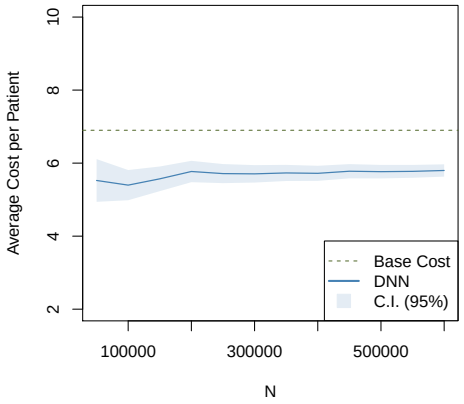
(a) Training Cost $\hat{\mathcal{C}}_n(\phi_n^*)$ Logit.



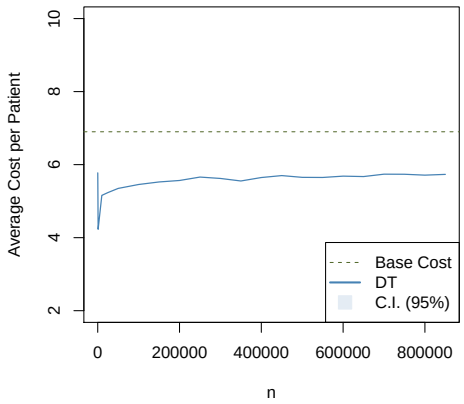
(b) Test Cost $\hat{\mathcal{C}}_N(\phi_n^*)$ Logit.



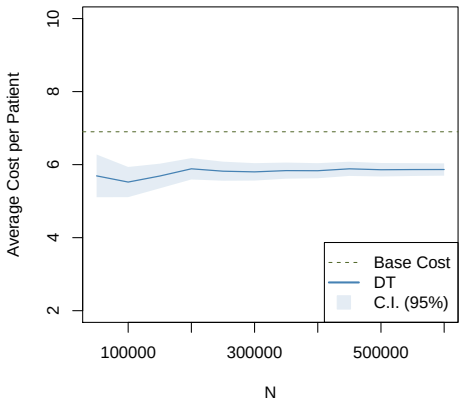
(c) Training Cost $\hat{\mathcal{C}}_n(\phi_n^*)$ DNN.



(d) Test Cost $\hat{\mathcal{C}}_N(\phi_n^*)$ DNN.



(e) Training Cost $\hat{\mathcal{C}}_n(\phi_n^*)$ DT.



(f) Test Cost $\hat{\mathcal{C}}_N(\phi_n^*)$ DT.

Figure 4: Training and test average mis-classification cost per patient for logistic regression, deep neural network and decision tree models. For the Test Costs in Column 2 the training is stopped at $n = 250,000$. The confidence interval for testing is based on Equation (39).

n	$\hat{\mathcal{C}}_n(\phi_n^*)$ Logit	$\hat{\mathcal{C}}_n(\phi_n^*)$ DNN	$\hat{\mathcal{C}}_n(\phi_n^*)$ DT
50,000	5.37	5.50	5.35
100,000	5.86	5.58	5.46
150,000	5.61	5.70	5.53
200,000	5.54	5.56	5.57
250,000	5.68	5.70	5.66
300,000	5.67	5.68	5.62
350,000	5.60	5.65	5.55
400,000	5.67	5.63	5.65
450,000	5.70	5.73	5.70
500,000	5.69	5.70	5.65
550,000	5.71	5.67	5.65
600,000	5.72	5.69	5.68
650,000	5.71	5.73	5.67
700,000	5.73	5.73	5.74
750,000	5.73	5.72	5.74
800,000	5.78	5.69	5.71
850,000	5.74	5.73	5.73

Table 6: Training mis-classification cost for logistic regression, deep neural network and decision tree models for different values of n using the two-step procedure for optimization.

N	$\hat{\mathcal{C}}_N(\phi_n^*)$ Logit	$\hat{\mathcal{C}}_N(\phi_n^*)$ DNN	$\hat{\mathcal{C}}_N(\phi_n^*)$ DT
50,000	5.54	5.52	5.69
100,000	5.44	5.40	5.52
150,000	5.57	5.57	5.69
200,000	5.78	5.77	5.89
250,000	5.75	5.71	5.82
300,000	5.74	5.71	5.80
350,000	5.76	5.73	5.84
400,000	5.75	5.72	5.83
450,000	5.80	5.78	5.89
500,000	5.79	5.77	5.86
550,000	5.79	5.77	5.87
600,000	5.81	5.80	5.87

Table 7: Test mis-classification cost for logistic regression, deep neural network and decision tree models for different values of N after stopping training at $n = 250,000$.

6 Concluding Remarks

In this study we have formulated the mis-classification cost problem as a loss function that captures the intrinsic trade-off between type-I and type-II errors incurred by a machine learning algorithm. We have developed a two-step algorithm that solves the classification problem efficiently and effectively – much faster than the traditional training of deep neutral networks while achieving virtually the same cost objective. Furthermore, we prove the convergence of the algorithm and also characterize the rate of convergence, via both analytical and numerical results.

There are two avenues that require further research effort. The first one is to continue exploring good algorithms—i.e., those that can overcome the multitude of local optimal solutions and quickly converge to the global optimal solution— that are even better than the two-step algorithm proposed here. The second one has to do with the performance bound. Here, the real object of interest is

$|\hat{\mathcal{C}}_n(\phi_n^*) - \mathcal{C}(\phi^*)|$, the gap between the ML scheme ϕ_n^* derived from the given data set (of size n)—the "training" problem—and the (unknowable) best scheme ϕ^* for the original problem. Here, this gap is bounded via $\sup_\phi |\hat{\mathcal{C}}_n(\phi) - \mathcal{C}(\phi)|$, which is essentially a worst-case' bound (due to the sup). Thus, it is no surprise that the observed performances in our numerical study are far better than the analytically derived bounds. It remains a worthy task to improve the latter to a level that commensurates the former.

Acknowledgments

This research is support in part by AHRQ-R01-HS024915-01 (PI: Elaine Larson). We thank Elaine Larson, Philip Zachariah, Bevin Cohen, Jianfang Liu, Jingjing Shang and the rest of the team at Columbia University Medical Center for sharing the data used in the paper and for many insightful discussions that have enhanced our understanding of HAI and related issues.

References

Arefian, Habibollah, Monique Vogel, Anja Kwetkat, and Michael Hartmann (2016). "Economic evaluation of interventions for prevention of hospital acquired infections: a systematic review". *PloS one* 11 (1), e0146381.

Bartlett, Peter L, Michael I Jordan, and Jon D McAuliffe (2006). "Convexity, classification, and risk bounds". *Journal of the American Statistical Association* 101 (473), 138–156.

Chen, Jonathan H, Steven M Asch, et al. (2017). "Machine learning and prediction in medicine-beyond the peak of inflated expectations". *N Engl J Med* 376 (26), 2507–2509.

Devroye, Luc, László Györfi, and Gábor Lugosi (1996). *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media.

Dingle, Kate E et al. (2017). "Effects of control interventions on Clostridium difficile infection in England: an observational study". *The Lancet Infectious Diseases* 17 (4), 411–421.

DRA (2006). *Deficit Reduction Act Sec. 5001. Hospital Quality Improvement.*

Dvoretzky, Aryeh, Jack Kiefer, Jacob Wolfowitz, et al. (1956). "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator". *The Annals of Mathematical Statistics* 27 (3), 642–669.

Kleywegt, Anton J, Alexander Shapiro, and Tito Homem-de Mello (2002). "The sample average approximation method for stochastic discrete optimization". *SIAM Journal on Optimization* 12 (2), 479–502.

Lasserre, Jean B (2001). "Global optimization with polynomials and the problem of moments". *SIAM Journal on optimization* 11 (3), 796–817.

Massart, Pascal (1990). "The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality". *The annals of Probability* 18 (3), 1269–1283.

Neyman, Jerzy and Egon Sharpe Pearson (1933). "IX. On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231 (694-706), 289–337.

Schoonover, Heather, Kristen Kelley, and Levi Thatcher (2017). "Accurately Predicting Risk of Central Line-Associated Bloodstream InfectionApplication of Machine Learning to Predict and Minimize Incidence of Central Line-Associated Bloodstream Infection". *American Journal of Infection Control* 45 (6), S46.

Septimus, Edward, Robert A Weinstein, Trish M Perl, Donald A Goldmann, and Deborah S Yokoe (2014). "Approaches for preventing healthcare-associated infections: go long or go wide?" *Infection Control & Hospital Epidemiology* 35 (7), 797–801.

Shiryayev, Albert Nikolaevich (2012). *Selected Works of AN Kolmogorov: Volume II Probability Theory and Mathematical Statistics*. Vol. 26. Springer Science & Business Media.

Umscheid, Craig A et al. (2011). "Estimating the proportion of healthcare-associated infections that are reasonably preventable and the related mortality and costs". *Infection Control & Hospital Epidemiology* 32 (2), 101–114.

Wiens, Jenna, Eric Horvitz, and John V Guttag (2012). "Patient risk stratification for hospital-associated c. diff as a time-series classification task". *Advances in Neural Information Processing Systems*. 25, 467–475.

Zimlichman, Eyal et al. (2013). "Health care–associated infections: a meta-analysis of costs and financial impact on the US health care system". *JAMA internal medicine* 173 (22), 2039–2046.