
Supplementary Material to “Nonparametric variable importance using an augmented neural network with multi-task learning”

Jean Feng ^{* 1} Brian D. Williamson ^{* 1} Marco Carone ^{1 2} Noah Simon ¹

1. Proof of Lemma 1

Using the classic result from [Leshno et al. \(1993\)](#), we show that there is a neural network that approximates the function $g_{P_0} : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$ arbitrarily well, where g_{P_0} combines the conditional means into a single function

$$g_{P_0}(x, m) := \mu_{P_0}(x) \mathbb{1}\{m = 0\} + \sum_{s \in \mathcal{S}} \mu_{P_0, s}(x) \mathbb{1}\{m = e_s\}. \quad (1)$$

Since g_{P_0} contains all the information from the conditional means, such a neural network is also a good approximation of all the conditional means.

Proof. For any \mathcal{S} , define the augmented conditional function $g_{P_0}(x, m)$ as in (1). Let $\mathcal{E} = \{0\} \cup \{e_s : s \in \mathcal{S}\}$, and let $\tilde{g}_{P_0}(x, m)$ be any continuous function defined over the domain $K \times [-1, 2]^p$ that shares the same values as $g_{P_0}(x, m)$ over all $K \times \mathcal{E}$. Using the result of [Leshno et al. \(1993\)](#), there exists a sequence of neural networks $\{f_j\}_{j=1}^\infty \in \mathcal{F}$ with parameters $\{\theta_j\}_{j=1}^\infty \in \Theta$ such that

$$\lim_{j \rightarrow \infty} \|f_j(x, m; \theta_j) - \tilde{g}_{P_0}(x, m)\|_{L^\infty(K \times [-1, 2]^p)} = 0.$$

Our desired result follows from the fact that

$$\begin{aligned} & \|f_j(x, m; \theta_j) - \tilde{g}_{P_0}(x, m)\|_{L^\infty(K \times [-1, 2]^p)} \\ & \geq \max_{s \in \mathcal{S}} \|f_j(x, e_s; \theta_j) - \mu_{P_0, s}(x)\|_{L^\infty(K)}. \end{aligned}$$

2. Experiments on simulated data

2.1. A non-additive six-variable function

We consider a situation where X is composed of six features and the conditional mean is a multivariate function that only depends on the first four features:

$$f(x_1, \dots, x_6) = x_1 \sin(x_1 + 2x_2) \cos(x_3 + 2x_4). \quad (2)$$

We are interested in estimating the variable importance of the groups $\{x_1, x_2\}$, $\{x_3, x_4\}$, and $\{x_5, x_6\}$, given by 0.820, 0.838, and zero, respectively.

In Table 1, we display the neural network structures that we cross-validated over to obtain the optimal neural network structure.

2.2. A sum of univariate functions

Here we consider the sum of eight univariate functions:

$$f(\mathbf{x}) = x_1 + x_2^2 + \sin(x_3) + \cos(x_4) + (x_5 + 1)^2 - 2x_6 + \max(x_7, 0) + x_8. \quad (3)$$

We estimate the importance of all groups with cardinality up to four.

We train the augmented neural network by minimizing the sum of the losses for estimating the full and reduced conditional means:

$$\begin{aligned} \hat{\theta} \in \arg \min_{\theta \in \Theta^{(2p)}} & \frac{1}{n} \sum_{i=1}^n \left[\left\{ y^{(i)} - f(x^{(i)}, 0; \theta) \right\}^2 \right. \\ & \left. + \sum_{s \in \mathcal{S}} \mathbb{E}_{W_s} \left(\left[y^{(i)} - f(\xi(x^{(i)}, W_s; s), e_s; \theta) \right]^2 \right) \right], \end{aligned} \quad (4)$$

^{*}Equal contribution ¹Department of Biostatistics, University of Washington, Seattle, Washington, USA ²Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. Correspondence to: Jean Feng <jean.feng@uw.edu>, Brian Williamson <brianw26@uw.edu>.

□

	NN structures to cross-validate over
Multiple networks	Full: 6,40,20,1;6,40,40,1;6,20,20,20,1;6,40,20,20,1
	Reduced $\{x_1, x_2\}$: 4,5,5,1;4,10,5,1;
	Reduced $\{x_3, x_4\}$: 4,5,5,1;4,10,5,1;
	Reduced $\{x_5, x_6\}$: 4,20,20,20,1;4,40,20,20,1;4,40,40,20,1
Augmented MTL network	12,40,40,20,1;12,40,40,40,1;12,80,40,40,1

Table 1. Network structures used for multiple networks vs the augmented MTL network in the non-additive six-variable function example. The fitting times are similar for fitting a single network (80-200 seconds) since the NN structures used for the two approaches are similar. However, there is an increase in time in the multiple networks approach: the practitioner first must spend time finding the correct structures; then, more structures must be cross-validated over, since different covariates have different optimal NN structures.

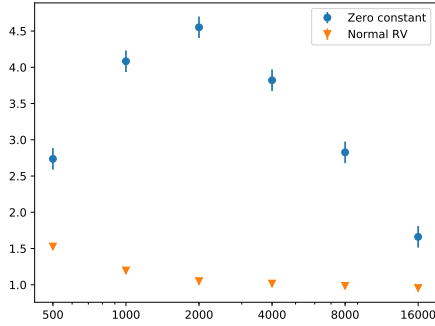


Figure 1. The multi-task loss (4) for the simulation specified by (3) when fitting MTL augmented networks with $W_s \equiv 0$ vs. $W_s \sim N(0, 1)$. The points and error bars represent the mean multi-task loss and its 95% confidence interval; the errors bars do not show for the normally distributed inputs since the CI is very narrow.

where W_s is a random variable with value in $\mathbb{R}^{|s|}$; and the function $\xi(x, w; s)$ maps (x, w) to \mathbb{R}^p by defining $\{\xi(x, w; s)\}_{(-s)} = x_{(-s)}^{(i)}$ and $\{\xi(x, w; s)\}_s = W_s$.

In Figure 1, we compare the multi-task loss (4) when we use 0 for W_s versus when we use a standard normal random variable instead. The minimum loss function value in this setting is 1, due to the variance of the outcome. We see that using random noise results in much improved performance over simply using zero.

3. Predicting Mortality of ICU Patients

Here, we analyze the importance of features measured during the first two days of patients’ ICU stays for predicting in-hospital mortality using data from Event 1 of the PhysioNet/CinC Challenge 2012 (Silva et al., 2012). The challenge provided a training dataset composed of four thousand records with five general descriptors collected upon admission (gender, age, weight, height, ICU admission type) and 37 features measured over the course of the first 48 hours after admission to the ICU, including the Glasgow Coma

Score, blood urea nitrogen, and heart rate, among others. Some patients never had particular features measured, and some only had them measured once, leading to a large degree of missing data. Estimating variable importance may provide guidance on which measurements are most informative for predicting in-hospital mortality.

We computed new features based on those proposed in a neural-network submission to the challenge (Xia et al., 2012) and those used to calculate SAPS I and SAPS II scores (Le et al., 1984; Le Gall et al., 1993). Xia et al. (2012) chose to use 18 of the 37 original variables and compute from them a total of 27 features, such as mean, min/max, and the last measurement; their model was then fit on these 27 computed features. We included these 27 computed features as well as the minimum, maximum, and mean (from fitting linear regression) from the time series of the 18 original variables if they were not already included. In addition, we (1) added five variables that are used in SAPS I and SAPS II but were not in this set of 18 original variables and (2) included all general descriptors measured at admission.

This procedure resulted in a total of 55 computed and original features in our model (Table 2). Though Xia et al. (2012) did a thorough evaluation to decide which variables to include in the model using univariate classifiers, such classifiers may fail to capture the potentially more complex relationship between the features and mortality; we include a more complete set of features to address this gap, and provide more information by estimating the importance of feature groups.

To understand the importance of running particular tests, we estimate the variable importance of variables grouped by the tests that provide measurements of those variables. For instance, comprehensive tests like metabolic panels or complete blood counts measure multiple variables. We group all extracted features for all these variables together as it will help us to understand the importance of running that particular test. (In addition, those measurements are likely to provide similar information for predicting the response so it would be helpful to relieve that correlation by putting them in a single group – otherwise we are estimating a “de-

Variable importance via neural networks

Variable (Meta)-Group	Variable	Summary (computed or original)
GCS	GCS	last, weighted mean, max, min, slope
Metabolic panel	HCO3	min, max, last, weighted mean
	BUN	min, max, last, weighted mean
	Na	min, max, weighted mean
	K	min, max, weighted mean
	Glucose	min, max, weighted mean
SysABP	SysABP	min, max, last, weighted mean
CBC	WBC	min, max, last, weighted mean
	HCT	min, max, weighted mean
Temp	Temp	min, max, last, weighted mean
Lactate	Lactate	min, max, last, weighted mean
HR	HR	min, max, weighted mean
Respiration	RespRate	min, max, weighted mean
	MechVent	max
	FiO2, PaO2	ratio of means
Urine	Urine	sum (based on SAPS II urine item)
General Desc.	Gender	measured at admission
	Height	measured at admission
	Weight	measured at admission
	Age	measured at admission
	ICU admission type	measured at admission

Table 2. Features included for analysis of the PhysioNet/CinC Challenge 2012. CBC = complete blood count test. Weighted mean = fit linear regression of response vs. time and get the estimate at the mean measurement time. Slope = fit linear regression of response vs. time and get slope. Last = last measurement. Impossible values were dropped (zero or lower for many of these variables).

flated” variable importance.) Thus we estimated the variable importance of groups (first column in Table 2) as well as individual variables (second column in Table 2). Note that individual variables may be extracted into multiple features – we group these extracted features together to assess the importance of that variable. Thus we estimate a total of 25 variable importance values. To estimate variable importance, we fit an augmented network structure with MTL.

As the variables measured across different patients is not always the same, we must handle missing data. Our method is well-suited for data that is missing at random. However we expect that many many measurements are not missing at random – doctors are likely to collect measurements for ones they are concerned about in the patient and ignore collecting measurements that they believe are normal in the patient. Therefore for features that are likely not to be missing at random, we imputed random values uniformly within the normal ranges of that measurement at each step of stochastic gradient descent. For features that are likely missing at random, e.g. age, height, weight, we indicated that covariate was missing via the binary augmented/missingness vector and fed in random values for the missing covariate.

We tuned the network structure via training/validation split (80/20) and chose layer sizes 110,4,3,2,1 with relu activation functions for the hidden nodes and a sigmoid function for

the output. The final variable importance estimates are based on models fit on all the data.

Our method estimates that among the medical test variable groups, the Glasgow Coma Score test had the highest variable importance score by far (Figure 2 top). This makes sense as the Glasgow Coma Score scores the consciousness of a patient; the scoring system is based on whether the patient can open their eyes, talk, and move in response to various levels of stimuli. Among all the medical tests, the GCS intuitively seems to have the most direct link to in-hospital mortality. Our conclusion is also supported by the fact that GCS score can contribute the most number of points (up to 26 points) to the SAPS II score whereas most items in the SAPS II score contribute at most 10 points.

The metabolic panel scored next highest in terms of variable importance. This also makes sense as (1) this test is perhaps one of the most commonly ordered lab tests to assess patient health and (2) this test includes measurements of many different variables. The primary driver of the importance of the metabolic test is blood urea nitrogen (BUN) which assesses kidney function. Again, our conclusion is also supported by the SAPS II scoring method – HCO3, BUN, Na, and K combined can contribute up to 24 points.

The confidence intervals for variable importance for urine

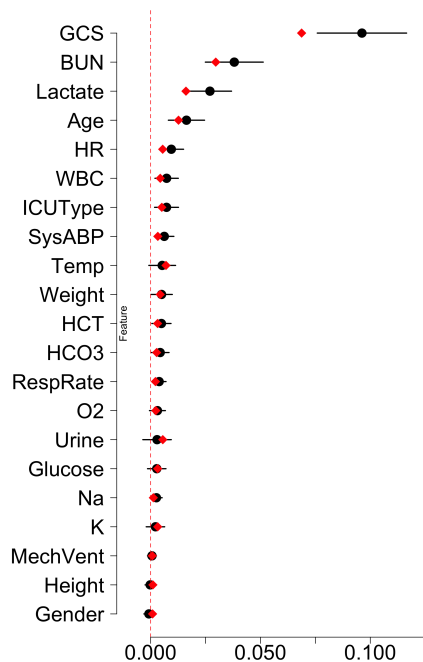


Figure 2. Variable importance estimates for tests in the ICU data (naive = red diamonds; corrected = black circles). Confidence intervals for the true importance, based on the corrected estimator only, are displayed as black bars.

and temperature include zero. The low-importance of temperature makes sense – temperature is one of the lowest-scoring items on SAPS II with a maximum contribution of 3 points. Urine can contribute up to 11 points in SAPS II, but more modern mortality scoring methods like APACHE IV (Zimmerman et al., 2006) do not use urine output for predicting mortality. As such, our results also suggest that to predict in-hospital mortality, we can drop measuring urine or temperature. In fact, dropping urine measurements for mortality could relieve burden from patients and nurses; urine is typically quite taxing to monitor since (1) the patient must either be hooked up to a urine monitoring device (painful) or the patient must urinate into containers (annoying) and (2) the nurses must regularly measure the amount of urine over the course of a day.

References

- Le, JRG, Loirat, P, Alperovitch, A, Glaser, P, Granthil, C, Mathieu, D, Mercier, P, Thomas, R, and Villers, D. A simplified acute physiology score for icu patients. *Critical care medicine*, 12(11):975–977, 1984.
- Le Gall, J-R, Lemeshow, S, and Saulnier, F. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24): 2957–2963, 1993.
- Leshno, M, Lin, VY, Pinkus, A, and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.
- Silva, I, Moody, G, Scott, DJ, Celi, LA, and Mark, RG. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *Computing in Cardiology (CinC)*, 2012, pp. 245–248. IEEE, 2012.
- Xia, H, Daley, BJ, Petrie, A, and Zhao, X. A neural network model for mortality prediction in icu. In *Computing in Cardiology (CinC)*, 2012, pp. 261–264. IEEE, 2012.
- Zimmerman, JE, Kramer, AA, McNair, DS, and Malila, FM. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for today's critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.