A Point-wise linear models and feature importance

In this appendix we introduced the point-wise linear model, our explainable deep learning model, and the method we used to derive the importance score for each feature (i.e., biomarker). Deep learning models are non-linear models that have achieved remarkable successes in the various fields including the medical one [1], because their deep neural network structure can express complex interactions between input features.

A.1 Idea of point-wise linear models

Let $\boldsymbol{x}^{(n)} \in \mathcal{R}^D$ $(n = 1, \cdots, N)$ represent a D-dimensional feature vector with N denoting the sample size and $\mathcal{R}$ denoting the set of all real numbers. Using the feature vectors and their binary target labels, we constructed the following nonlinear deep-learning model to generate a logistic regression model for each sample:

$$y^{(n)} = \frac{1}{1 + e^{-(\boldsymbol{\xi}^{(n)} \cdot \boldsymbol{x}^{(n)})}}, \tag{1}$$

$$\boldsymbol{\xi}^{(n)} \equiv \boldsymbol{\xi}(\boldsymbol{x}^{(n)}), \tag{2}$$

where $\boldsymbol{a} \cdot \boldsymbol{b}$ is the dot product of two arbitrary vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, and $y^{(n)} := \{y^{(n)} \in \mathcal{R} | 0 \leq y^{(n)} \leq 1\}$ is the probability with which the sample with index (n) is classified as label 1. Each element of $\boldsymbol{\xi} \in \mathcal{R}^D$ is a function of $\boldsymbol{x}^{(n)}$ that a deep neural network determines so as to minimize the negative log likelihood loss of the output $y^{(n)}$ for the target label. $\boldsymbol{\xi}$ behaves as the weight vector for the original feature vector $\boldsymbol{x}^{(n)}$, thus giving the explainability (i.e., the importance of each feature) in the model Eq (1). For simplicity of notations, bias parameters have been omitted from Eq (1). Here we should notice that this weight vector is tailored to each sample because $\boldsymbol{\xi}$ depends on $\boldsymbol{x}^{(n)}$. We call Eq (1) the point-wise linear model over the sample index (n).

A.2 Feature importance

In this subsection, we defined importance scores for each feature to extract variables which could clearly distinguish the responder patients from non-responder patients, based on the point-wise linear model. In the following, we considered the point-wise linear model where the patients with target labels 1 and 0 are defined as the responder and non-responder patient, respectively. Hence the output probability $y^{(n)}$ indicates the probability with which the model classifies a patient with index (n) as a responder.

As was discussed in the previous subsection, the point-wise linear model gives the weight vector $\boldsymbol{\xi}^{(n)}$ tailored to each patient. First, we calculated feature importance for each patient from the weight vector. We introduced a sample-wise importance score for a k-th feature x_k of a sample with index (n) as

$$s_k^{(n)} \equiv \xi_k^{(n)} x_k^{(n)} - \frac{1}{|U_{(n)}|} \sum_{i \in U_{(n)}} \left[\xi_k^{(i)} x_k^{(i)} \right], \tag{3}$$

where $U_{(n)}$ is the set of samples whose weights are close enough to the one of sample (n). This sample-wise importance score was inspired by Shapley value [2] and expresses the contribution of a sample (n) to raising the output probability $y^{(n)}$ compared with the average contribution among a sample group whose members obey similar linear models. Positively large $s_k^{(n)}$ suggests that the feature x_k of the sample (n) distinctively raises the output probability $y^{(n)}$, while negatively large $s_k^{(n)}$ suggests that the feature x_k of the sample (n) distinctively lowers the output probability $y^{(n)}$.

In this study, we defined $U_{(n)}$ as follows: a sample (i) is in the set $U_{(n)}$ if $|\boldsymbol{\xi}^{(i)} - \boldsymbol{\xi}^{(n)}|/|\boldsymbol{\xi}^{(n)}|$ is smaller than $3|\boldsymbol{\sigma}|/|\bar{\boldsymbol{\xi}}|$, where $\boldsymbol{\sigma}$ and $\bar{\boldsymbol{\xi}}$ are vectors whose elements are given as the standard deviation and the mean, respectively, of the corresponding element of $\boldsymbol{\xi}$.

Next, we derived importance scores for a group (e.g. responder patients) from the sample-wise importance score. In order to reveal a group's macroscopic property contained in the sample-wise importance scores of all the samples in the group, we performed voting among the group based on

the value of sample-wise importance scores. We conducted two votings, one in which each patient votes to its top 10% features whose sample-wise importance scores are positively largest, and the other one in which each patient votes to its top 10% features whose sample-wise importance scores are negatively largest. For one group, we defined two group-wise importance scores $v_{\rm pos}$ and $v_{\rm neg}$ for each feature as the rate of samples who vote for the feature in the above two votings, respectively.

Finally, we defined the scores which extract the features that separate the responder patients and the non-responder patients, i.e., the features that largely change the output probability for only one of the groups (responders or non-responders). We divided the patients into the responder group and the non-responder one based on their target labels and evaluated the group-wise importance scores $v_{\rm pos}$ and $v_{\rm neg}$ for each group. We introduced relative scores $v_{\rm pos}^{\rm rel}$ and $v_{\rm neg}^{\rm rel}$ to compute how important the feature is in one of the two groups compared with how important it is in the other group:

$$v_{\rm pos}^{\rm rel} \equiv \left(v_{\rm pos}^{\rm r}\right)^2 - \left(v_{\rm pos}^{\rm n}\right)^2 , \tag{4}$$

$$v_{\rm neg}^{\rm rel} \equiv \left(v_{\rm neg}^{\rm r}\right)^2 - \left(v_{\rm neg}^{\rm n}\right)^2 , \tag{5}$$

where $v_{\rm pos}^{\rm r}$ and $v_{\rm neg}^{\rm r}$ are the group-wise importance scores for the responder group, and $v_{\rm pos}^{\rm n}$ and $v_{\rm neg}^{\rm n}$ are those for the non-responder group. Since all of the ranges of $v_{\rm pos}^{\rm r}$, $v_{\rm neg}^{\rm r}$, $v_{\rm pos}^{\rm n}$ and $v_{\rm neg}^{\rm n}$ are $[0, 1]$, both of the ranges of $v_{\rm pos}^{\rm rel}$ and $v_{\rm neg}^{\rm rel}$ become $[-1, 1]$. The features whose relative scores have large absolute values will help to separate the responder patients from the non-responder ones: a feature with positively (negatively) large $v_{\rm pos}^{\rm rel}$ tends to raise the output probability in the responder (non-responder) group, not in the non-responder (responder) group, and a feature with positively (negatively) large $v_{\rm neg}^{\rm rel}$ tends to lower the output probability in the responder (non-responder) group, not in the non-responder (responder) group.

In this study, we considered the features for which the absolute value of either of the relative scores is more than 0.09 as important biomarkers. Eqs. (4) and (5) imply that when an absolute value of a relative score $v_{\rm pos}^{\rm rel}$ or $v_{\rm neg}^{\rm rel}$ of a feature is more than 0.09, one of its two group-wise importance scores $v_{\rm pos}^{\rm r}$ and $v_{\rm pos}^{\rm n}$ or $v_{\rm neg}^{\rm r}$ and $v_{\rm neg}^{\rm n}$ must be at least 0.3, i.e., more than 30% of patients in one of the groups votes for the feature.

References

[1] L. Caroprese, P. Veltri, E. Vocaturo and E. Zumpano, "Deep Learning Techniques for Electronic Health Record Analysis," 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), Zakynthos, Greece, 2018, pp. 1-4.

[2] Shapley, L. S. 1953. A value for n-person games. In Contributions to the Theory of Games. volume II. Princeton University Press. 307-317.