Related work

BindSpace(2019)

High dimensional embedding

not choose because feature it choose is not representative enough for different cell type.

To make epigenetic data not only easier to analysis but also explainable, IMPACT utilizes SHAP to make feature selection very clear and easier.IMPACT tries to predict the death risk for people with different age and habit. The data set is describe as below.

| Feature | Type | Display\_Name |
| --- | --- | --- |
| Demographics\_Age | Demographics | Age |
| Demographics\_Citizenship\_2.0 | Demographics | Not a citizen of the US |
| Questionnaire\_AlcoholFreqDays | Questionnaire | Avg # alcoholic drinks/day - past 12 mos |
| Laboratory\_UrineAlbumin | Laboratory | Albumin, urine (ug/mL) |
| Examination\_Weight | Examination | Weight (kg) |

pic()Examples of Features in Dataset, containing 47,261 samples. Collected from NHANES

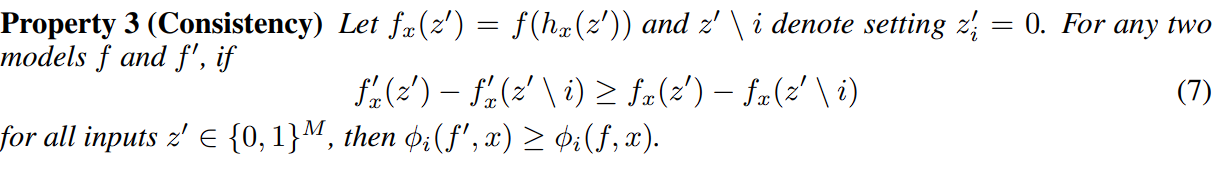
Though this is not a high dimensional case, the evaluation of risk requires evaluation of importance for different features. Thus explainable feature selection is important.

To understand the secret of SHAP, we first review the original paper where this concept is unified and well organized(1). SHAP combines the concept of Shapley value, which is hotly discussed in the field of machine learning, with different models (i.e. linear, DNN…).

Shapley values identify the importance of a feature by evaluating different combinations of features. Thus Shapley value wont ignore the cooperation between features. However, combination calculation is a disaster in computation, so the author(1) uses a linear model to approximate the combinatorial effect.

Then, the linear model will have weight for each feature. The weight here is the importance of the feature (at local area).Thus the explainable model has additive property of all features(i.e. the combine average importance of 2 feature is the sum of the weight)

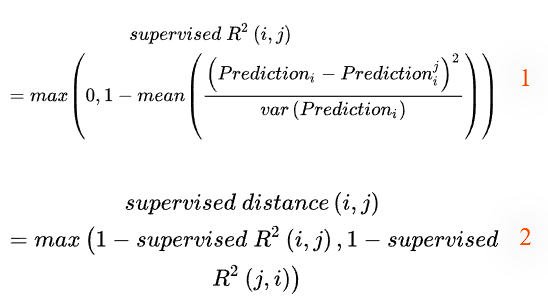
In the reference(1), it ensures this is the only way to guarantee local accuracy, missingness(if value goes 0 at the point, the weight should be 0), and consistency(as below). Thus this system is very reliable to be used to generate explainable feature selection.



pic() the Consistency property of SHAP

Now we get back to IMPACT. IMPACT uses gradient boosted tree as a model to work with sharpley values. To generate risk evaluation, IMPACT use generated model with SHAP to calculate importance. And the risk evaluation is simply the sum of each risk in different features( by additive property of SHAP).

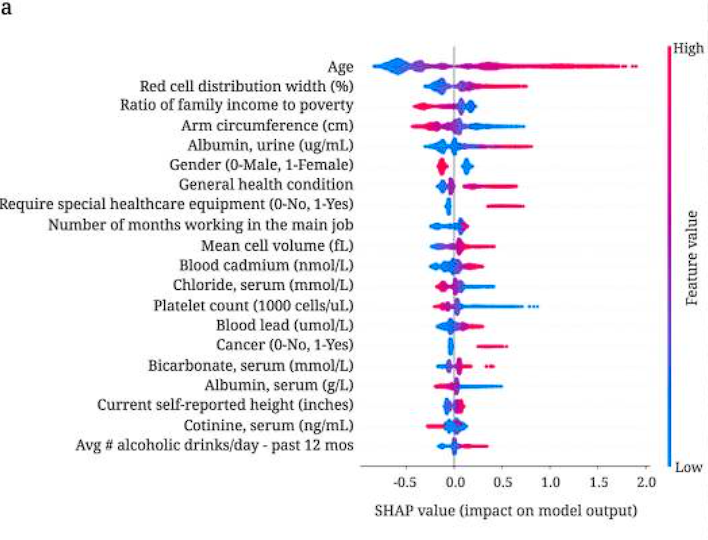
There is also another funny process of feature when generating a gradient boosted tree. To evaluate the relation between features, in order to kill redundant features, it defined a term supervised distance.



pic() supervised distance definition. predictioni means use i as 1-var model and predict with i; predictionij means use i as 1-var model and predict with j. 1-var model means only 1 feature is real data, other features use random selection from other data.

After having supervised distance, it uses hierarchical clustering to find out the most redundant features, that is, the pair with low supervised distance. Then remove the one with a smaller SHAP value. By repeating clustering and removal, it can eventually find out the feature set is predictive and less redundant.

The result is very surprising. First, The high SHAP value features have been shown to have correlation with mortality previously. For example, red cell distribution width (RDW) is identified as an indicator in experiment(2). Second, two of the twenty highest SHAP value features are not observed in the past medical reviews. which can be viewed as an important feature for medical use in the future.



pic() The 20th features with highest SHAP value, color indicate value of feature. length indicate the importance, width indicate the number of data.

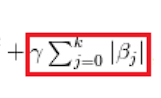
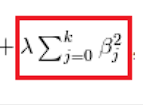
Reference:

1. [A Unified Approach to Interpreting Model Predictions](https://arxiv.org/abs/1705.07874)
2. Felker, G. M. et al. Red cell distribution width as a novel prognostic marker in heart failure: data from the CHARM Program and the Duke Databank. J. American College. Cardiol. 50, 40–47 (2007)

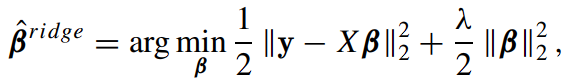
The second paper serves for general biological high dimensional cases with prior information. Motivated by want to learn result by classical models, author tries to use an alternative method to regularize, and tries to utilize prior knowledge while having regularization.

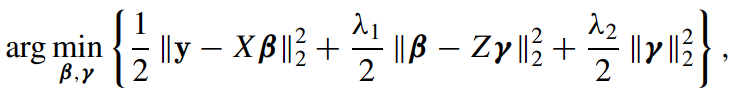
To utilize prior knowledge, the author invented a matrix that concentrates data into lower dimensions. And then regularize the weight of the outcoming features. The feature coming out from the matrix is called the meta features.

Like other regularizations, the author adds a penalty term to the loss function (as below).However, since there are two layers of features, there are also two layers of regularization.



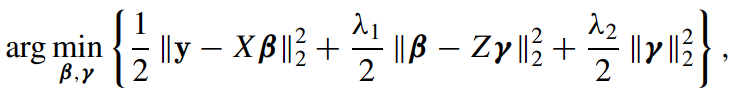
pic() Example for penalty in loss function. The left is for normal regularization, the right one is for Lasso regularization.

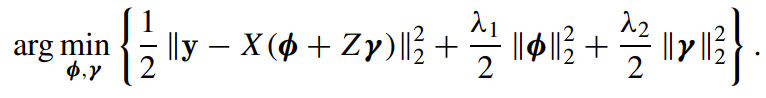


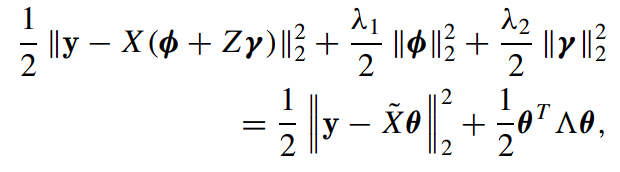


pic() The upper is the regular normalization loss function, the lower is the normalization author proposed. Z is matrix from prior knowledge, B: coefficient of subjective feature,γ: coefficient of meta feature

Just like ordinary regularization, this Hierarchical Ridge Regression also has closed form solution. By changing the parameters, closed form solution can be easily derived.



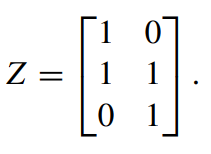
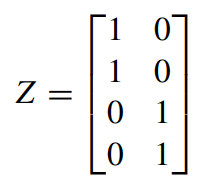






pic() closed form solution for hierarchical ridge regression.

The behavior of the mate matrix can have 3 common situations: totally uncorrelated, at the same pathway, appears in 2 pathways. For a totally uncorrelated case, the meta matrix is just a diagonal matrix with different eigenvalues. For the second case, the features in the same pathway will be assigned in the same colume, while each row has one nonzero value. For cases where a feature appears in multiple pathways, it is just like the previous case, besides there might be multiple nonzero terms in the meta matrix.

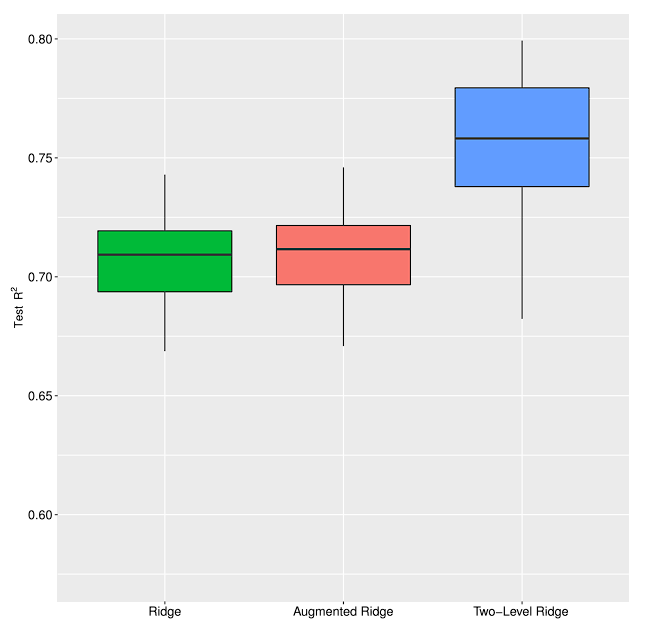


pic() Meta feature matrix with different behaviors. The left one show cases in the same pathway might be assigned in the same column. The right one shows cases that appear in multiple pathways.

The result was amazing compared to other regularizations.

DNA methylation is long been believed to have correlation with age. In 2013, Horvath use methylation pattern to build a highly predictive model for human real age, indicating methylation is indeed a reliable biomarker for age. In the analysis, they both use data with 250,000 features, which is the location of methylation. Meta matrix takes locations on DNA that belongs to the same gene to be same features, and find concentrated 250,000 features into 6,766.

Horvath Horvath use high dimensional regularization to enhance his predictor(2), while hierarchical ridge excels his result. This is one of the best way of showing the power of hierarchical ridge regularization.



pic() The comparison of different regularizations.

Though hierarchical ridge regulation is easy to apply and seems to be useful, The downside is also very obvious. This method only applies to linear model, and the quality of prior knowledge is very important.

reference

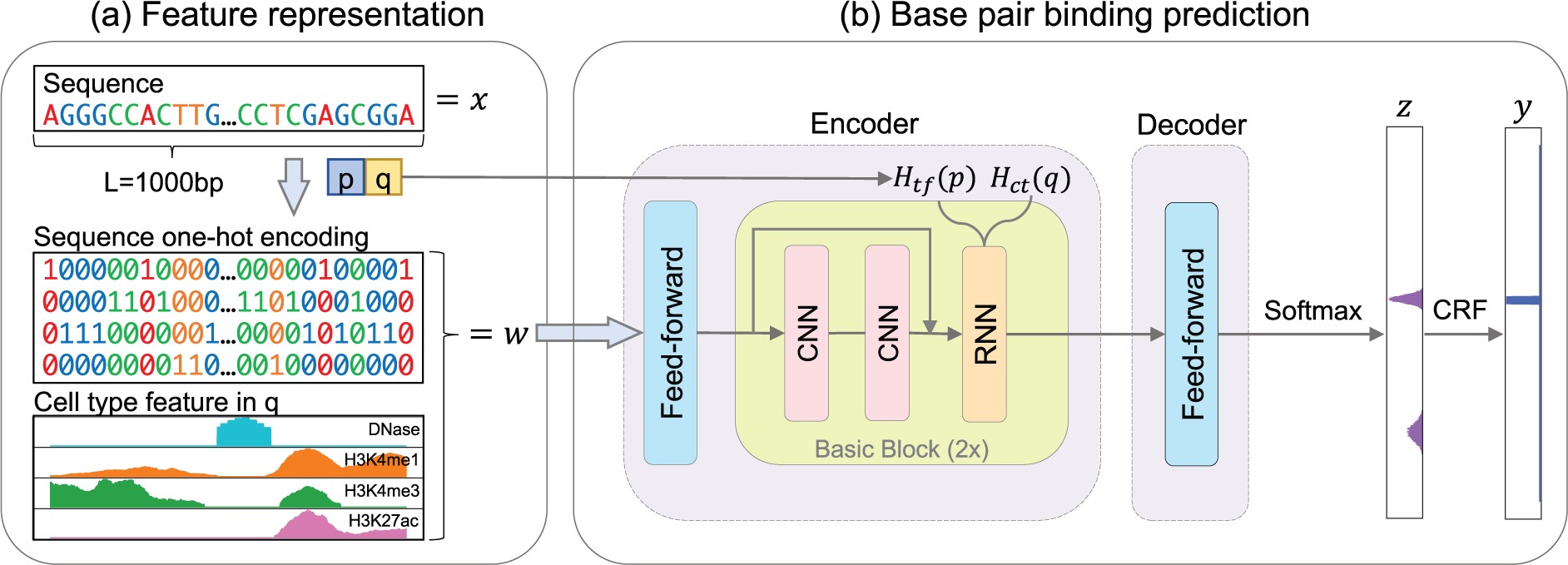
1. Dna methylation age of human tissues and cell types
2. Accelerated epigenetic aging in down syndrome

Nettime is quite different from other feature engineering. instead of lowering dimension, it tries to finding some new features to simplify the analysis.

Nettime aims to make DNA-protein binding prediction form cell type specific into one same question. DNA-protein binding prediction has long been researched(1).This topic mainly discuss how to make prediction whether a protein will bind on given DNA sequences. While sharing the DNA sequence, different cell type has different function, thus behave very differently in DNA-protein binding. Thus the problem has always been researched with specific types. In the process of organizing similarity between cell types, author found that some of the feature of a cell type might be used as a representative of cell type. By trying and prior knowledge, author found there are 4

cell type specific data. That is DNase, H3K4me1, H3K4me3, H3K27ac. These are modifications of DNA, which have different patterns at different loci on DNA.

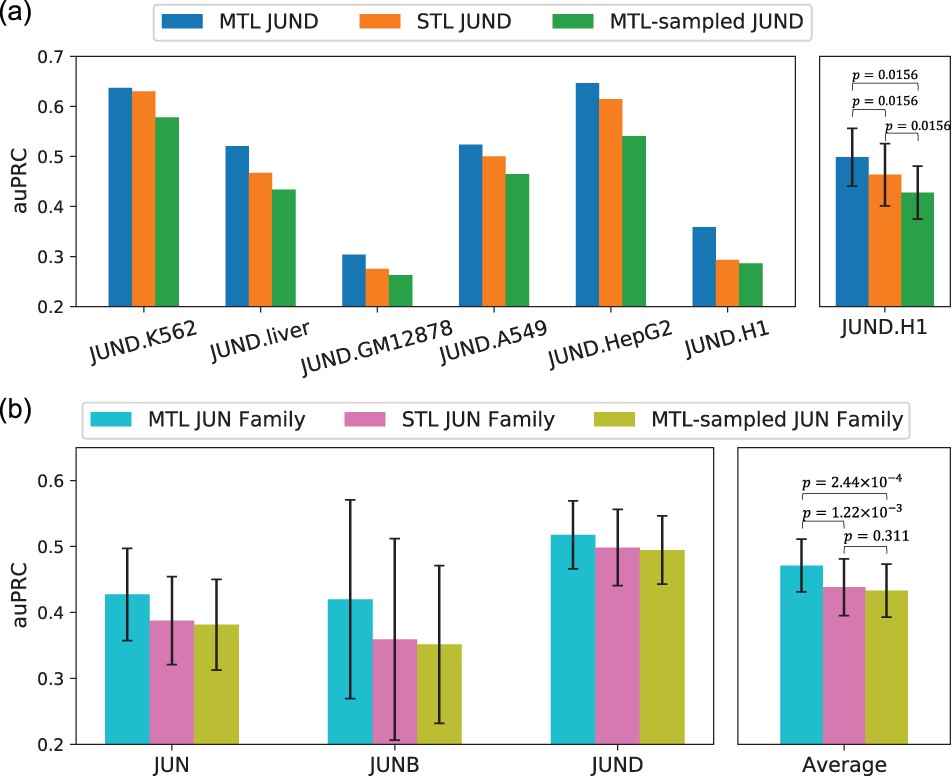
To preprocess the feature, the author turns DNA sequences into one-hot-encoded data, and each position includes 4 data mentioned above. Then the cell type represented feature is done.



pic() one-hot-encoded DNA sequence and the 4 cell type representing data

The result is quite stunning. If all cell type is put into training set, the performance is even better than cell type specific training prediction.

To conclude, the feature representing different cell type is very representative. and the result shows it can make better prediction than original method. However,due to the requirement of prior knowledge to accomplish this. This means the method is hard to improve and selection of these prior knowledge is also a challenge.



pic() Result for comparing cell type specific training.

reference

1. **Convolutional neural network architectures for predicting DNA–protein binding**