Predicting Complete Remission of Leukemia Treatments

Willy Chang
Department of Electrical and
Computer Engineering
University of Hawaii at Manoa

***Abstract*—The best binary classifier trained achieved a testing loss of 0.5292 and a testing accuracy of 0.7568.**

I. Problem

The dataset provided contains information regarding the medical data of patients with leukemia. The given information includes data for 181 patients with 37 pieces of clinical information and a patient's result from 231 antibodies, for a total of 268 initial input columns. With this information, a binary classifier is to be trained to identify if a patient will respond to a treatment or not (i.e. have complete remission or resistance). The data types for the 37 clinical information can be seen in Figure 1, and the data types for the proteomic measurements of the 231 antibodies are numerical.

Clinical Covariate	Values	Description
SEX	M, F	Patient gender
Age.at.Dx	numeric	Patient age at the time of diagnosis
AHD	numeric	Prior antecedent hematologic disorder
PRIOR.MAL	YES, NO	Whether the patient has been diagnosed with a prior cancer
PRIOR.CHEMO	YES, NO	Whether the patient has had prior chemotherapy
PRIOR.XRT	YES, NO	Whether the patient has had prior radiation therapy
Infection	YES, NO	Whether the patient was diagnosed with an infection
cyto.cat	"-5", "-5,-7", "-5,-7,+8", "-7", "-7,+8", "11q23", "21", "8", "diploid", "IM", "inv16", "inv9", "Misc", "t6;9", "t8;21", "t9;22"	The cytogenic category of the patient
ITD	NEG, POS, ND	Whether the patient was found to have a ITD FLT3 mutation
D835	NEG, POS, ND	Whether the patient was found to have a D835 FLT3 mutation
Ras.Stat	NEG, POS, NotDone	Whether the patient was found to have a Ras.Stat mutation
Chemo.Simplest	Anthra-HDAC, Anthra-Plus, Flu-HDAC, HDAC-Plus non Anthra, StdAraC-Plus	The specific Anthra based treatment administered
resp.simple	CR, RESISTANT	Patients were categorized as having a complete response or to be resistant to treatment.
Relapse	Yes, No, NA	Whether a patient with complete response later relapsed
vital.status	A, D	The final outcome of each patient at the end of the study, either alive or deceased.
Overall_Survival	numeric	A patient's overall survival time measured in weeks from diagnosis to exiting the study.
Remission Duration	numeric or NA	The duration of time spent in remission measured in weeks.
WBC	numeric	The white blood cell count
ABS BLST	numeric	The total number of myeloid blast cells measured in blood samples
BM.BLAST	numeric	The number of myeloid blast cells measured in bone marrow samples
BM.MONOCYTES	numeric or NA	The number of monocytes measured in bone marrow samples
BM.PROM	numeric or NA	The number of promegakaryocytes measured in bone marrow samples
PB.BLAST	numeric or NA	The number of myeloid blast cells measured in blood samples
PB.MONO	numeric or NA	The number of monocytes measured in blood samples
PB.PROM	numeric or NA	The number of promegakaryocytes measured in blood samples
HGB	numeric or NA	hemoglobin count measured in blood samples
PLT	numeric or NA	platelet count measured in blood samples
LDH	numeric or NA	Lactate dehydrogenase levels measured in blood samples
ALBUMIN	numeric	Albumin levels measured in blood samples
BILIRUBIN	numeric or NA	Bilirubin levels measured in blood samples
CREATININE	numeric	Creatinine levels measured in blood samples
FIBRINOGEN	numeric or NA	Fibrinogen levels measured in blood samples
CD13	numeric or NA	Levels of cell surface marker CD13 detected
CD33	numeric or NA	Levels of cell surface marker CD33 detected
CD34	numeric or NA	Levels of cell surface marker CD34 detected
CD7	numeric or NA	Levels of cell surface marker CD7 detected
CD10	numeric or NA	Levels of cell surface marker CD10 detected
CD20	numeric or NA	Levels of cell surface marker CD20 detected
HLA.DR	numeric or NA	Levels of cell surface marker HLA.DR (human leukocyte antigen) detected
CD19	numeric or NA	Levels of cell surface marker CD19 detected

Fig. 1. Clinical feature description.

II. Methodology

The provided dataset was divided into an 80-20% split at random, with 80% (144 patients) allocated for the training set and 20% (37 patients) allocated for the testing set. The column for prediction, resp.simple, contained two values: 'CR' and 'RESISTANT', corresponding to 'Complete Remission' and 'Complete Resistant' to leukemia treatment, respectively. These values were converted to 0 and 1 using the sklearn LabelEncoder() method, where 0 replaced 'CR' and 1 replaced 'RESISTANT'.

TABLE I
Base binary classifier implementation.

Parameters	Values
Layers	Dense, 256 Nodes, ReLU Dense, 1 Node, Sigmoid
Loss Function	Binary Cross Entropy
Optimizer	Adam
Epochs	100
Batch Size	10
Early Stopping	Validation Accuracy, 20 epochs

In order to train the best fitting model for the given problem set, an iterative approach was used to determine the best model parameters. That is, holding all other parameters constant, one parameter was modified between a series of possibilities. The value for that parameter that produced the best validation accuracy would become the new constant when testing for the other model parameters. The parameters modified were done in the following order: layers, optimizer, and batch size. Considering generalization error, 5-fold cross-validation was implemented.

Some categories for the dataset were also missing values such as BMBLAST, BMMONOCYTES, and CD7, a total of 18 columns. To resolve this issue, a simple imputation using the mean of the values for that feature were used as a starting point [1]. Furthermore, all categorical-based columns were dropped (total of 11 features) and the rest of the numerical categories were normalized to a mean of 0 and a variance of 1. The base implementation of the neural network can be seen in Table I. The resulting cross-validation error was 0.6353 and the cross validation accuracy was 0.7426.

III. Results

A. Missing Value Imputation

For the missing values, three simple imputations were conducted on the base network to determine the best method as well as k-nearest neighbors (kNN). The results from filling the missing values with these methods are listed in Table II.

TABLE II
IMPUTATION RESULTS.

Methods	CV Error	CV Accuracy
Mean	0.6353	0.7426
Median	0.6270	0.7079
Mode	0.0.6344	0.7153
kNN (k=2)	0.6378	0.7086

With the highest cross validation accuracy coming from the mean method, it was used for future tests.

B. Neural Network Layers

Numerous possibilities exist for the construction of the neural network, including, but not limited to, the type of layers, features per layer, and activation function. Listed in Table III are some of the tests conducted to find the optimal structure for the binary classifier.

TABLE III
LAYER MODIFICATION RESULTS.

Layers	CV Error	CV Accuracy
Dense, 256 Nodes, ReLU Dense, 1 Node, Sigmoid	0.6353	0.7426
Dense, 128 Nodes, ReLU Dense, 1 Node, Sigmoid	0.6090	0.7362
Dense, 512 Nodes, ReLU Dense, 1 Node, Sigmoid	0.6556	0.7084
Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 1 Node, Sigmoid	0.6009	0.7153
Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 128 Nodes, ReLU Dropout, 0.5 Dense, 1 Node, Sigmoid	0.5942	0.7153
Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 512 Nodes, ReLU Dropout, 0.5 Dense, 1 Node, Sigmoid	0.5916	0.7084
Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 128 Nodes, ReLU Dropout, 0.5 Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 1 Node, Sigmoid	0.5953	0.7224
Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 128 Nodes, ReLU Dropout, 0.5 Dense, 128 Nodes, ReLU Dropout, 0.5 Dense, 1 Node, Sigmoid	0.5937	0.7224
Dense, 256 Nodes, ReLU Dropout, 0.5 Dense, 128 Nodes, ReLU Dropout, 0.5 Dense, 64 Nodes, ReLU Dropout, 0.5 Dense, 1 Node, Sigmoid	0.5928	0.7224

The layer with the highest cross validation accuracy was the Dense layer with 256 nodes followed by an output, sigmoid classification layer.

C. Optimizer

The optimizers that were tested include Adam, SGD, and Adadelta. The results can be seen in Table IV

TABLE IV
OPTIMIZER MODIFICATION RESULTS.

Optimizer	CV Error	CV Accuracy
Adam	0.6353	0.7426
SGD	0.6050	0.7224
Adadelta	0.6556	0.6259

From these results, the optimizer Adam was used.

D. Batch Size

For the batch size, experiments were conducted with the following values: 1, 5, 10, and 20. The results from these experiments can be seen in Table V

TABLE V
OPTIMIZER MODIFICATION RESULTS.

Optimizer	CV Error	CV Accuracy
1	0.8265	0.7219
5	0.6451	0.7291
10	0.6161	0.7498
20	0.5900	0.7153

Using these results and the deciding criteria of maximizing validation accuracy, a batch size of 10 was chosen for the final model.

E. Final Model

Compiling the results from the test, the final model was created and pickled with the parameters seen in Table VI.

TABLE VI
FINAL MODEL PARAMETERS.

Parameters	Values
Layers	Dense, 256 Nodes, ReLU Dense, 1 Node, Sigmoid
Loss Function	Binary Cross Entropy
Optimizer	Adam
Epochs	100
Batch Size	10
Early Stopping	Validation Accuracy, 20 epochs

Using normalization of the data along with a mean imputation for missing values, the model produced a testing loss of 0.5292 and a testing accuracy of 0.7568. The model was saved into 'respsimple'.

IV. CONCLUSION

While not every parameter was hypertuned, the best results given the tests done provided sufficient results with a testing accuracy of 0.7568. Other optimizations that could be done to improve the neural network include changing the early stopping parameters, including the categorical features, and using efficient hyperparameter tuning algorithms such as grid search or random search.

REFERENCES

[1] Tlamelo Emmanuel et al. “A survey on missing data in machine learning”. In: *Journal of Big Data* 8.1 (Sept. 2021), p. 140. ISSN: 2196-1115. DOI: 10.1186/s40537-021-00516-9. URL: https://doi.org/10.1186/s40537-021-00516-9.