iEN: integrating mechanistic immunological information into a machine learning framework

Anthony Culos and Nima Aghaeepour

March 17, 2020

tculos@stanford.edu and naghaeep@gmail.com

Contents

1 Licensing

2 Introduction

Supervised predictive analysis method for flow and mass cytometry data which integrates well known information on cell-signalling responses into the model fitting process. This method is an extension of the well studied Elastic-Net algorithm [2]

Through the prioritization of well understood cell-signalling pathways predictive models of human immunity can more accurately predict responses than the agnostic EN method. The prior knowledge refered to are likelihood coefficients generated by a panel of expert immunologists such that features more consistent with known biology have a higher value with numbers ranging from 0 to 1.

A two-layer K-fold Cross-Validation (CV) approach is used to optimize and estimate model performance in a robust and statistically stringent manner, with the parameter foldid controlling CV behaviour. The models generated by this package are regression model's optimized via grid search during cross-validation

on the provided parameters for alphaGrid, phiGrid, and nlambda. Prediction useing the generated "iEN" object will use the mean of out-of-sample models (which is the collection of optimal models generated for each fold of cross-validation) to predict new data, with default prioritization of the new data being the mean of optimal scaling from the K-fold CV.

3 Example

This example uses mass cytometry data previously published by Aghaeepour et al [1] which studied the gestational age during pregnancy. Using the data matrix X and the vector of prior knowledge, our task is to estimate the vector of gestational age at time of sample collection Y.

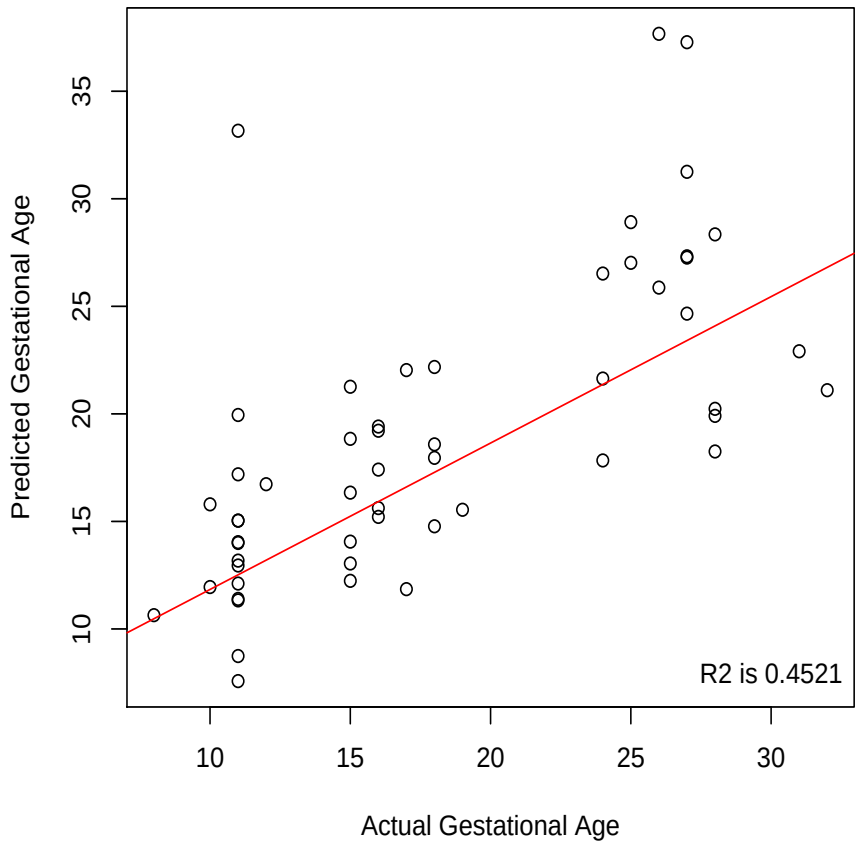
```
> library(iEN)
> install.packages("caret", repos='http://cran.us.r-project.org')
> library(caret)
> data(test_data)
> alphaGrid <- seq(0,1, length.out=2)
> phiGrid <- exp(seq(log(1),log(10), length.out=2))
> nlambda <- 3
> lambdas=NULL
> ncores <- 2
> eval <- "RSS"
> family <- "gaussian"
> intercept <- TRUE
> standardize <- TRUE
> center <- TRUE
> #define 10-fold cross-validation folds
> temp.folds <- createFolds(unique(foldid),k=10)
> folds <- vector()
> for(k in seq(length(temp.folds))){
+         folds[which(foldid %in% temp.folds[[k]])] <- k
+ }
> model <- cv_iEN(X, Y, folds, alphaGrid, phiGrid, nlambda, lambdas, priors, ncores, eval, f
> Y.hat <- model@cv.preds
> print(model)

Estimated model performance, as defined by 'eval' parameter, from out of sample predictions


| features| alpha|    lambda| phi|
|--------:|-----:|---------:|---:|
|      961|     0| 0.0000029|  10|
|      961|     0| 0.2903123|   1|
|      961|     0| 0.0000028|  10|
|      961|     0| 0.3337476|   1|
```

```
|       961|      0| 0.0000028|  10|
|       961|      0| 0.2912042|   1|
|       961|      0| 0.0000028|  10|
|       961|      0| 0.2686643|   1|
|       961|      0| 0.0000026|  10|
|        51|      1| 0.0009299|   1|

> plot(Y,Y.hat, ylab ="Predicted Gestational Age", xlab="Actual Gestational Age")
> abline(fit <- lm(Y ~ Y.hat,), col='red')
> legend("bottomright", bty="n", legend=paste("R2 is",  format(summary(fit)$adj.r.squared, c
```



References

[1] Nima Aghaeepour, Edward A Ganio, David Mcilwain, Amy S Tsai, Martha Tingle, Sofie Van Gassen, Dyani K Gaudilliere, Quentin Baca, Leslie McNeil, Robin Okada, Mohammad S Ghaemi, David Furman, Ronald J Wong, Virginia D Winn, Maurice L Druzin, Yaser Y El-Sayed, Cecele Quaintance, Ronald Gibbs, Gary L Darmstadt, Gary M Shaw, David K Stevenson, Robert Tibshirani, Garry P Nolan, David B Lewis, Martin S Angst, and Brice Gaudilliere. An immune clock of human pregnancy. *Sci Immunol*, 2(15), sep 2017.

[2] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, apr 2005.