200301-EDA_and_model-yuqi

Yuqi Miao ym2771

3/1/2020

data and manipulation

$$\log(\frac{\pi_i}{1-\pi_i}) = \mathbf{x}_i\boldsymbol{\beta}$$

validation using glm

questions or modify:

1. normalize or standardize?
2. how to standardize easily?

```
# cleaning the above x
library(sjmisc)
y=as.data.frame(x$cv_result)
y_y=rotate_df(y)
names(y_y)=c("Enter","Fold1","Fold2","Fold3","Fold4","Fold5")
knitr::kable(y_y)
```

	Enter	Fold1	Fold2	Fold3	Fold4	Fold5
k	0.00	1.0000000	2.0000000	3.0000000	4.0000000	5.0000000
best_lambda	0.00	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000
beta_vec1	0.02	-0.5851898	-0.5921021	-0.6782398	-0.6305423	-0.5250743
beta_vec2	0.02	1.9370600	1.8598428	1.8706748	1.0162875	1.7805325
beta_vec3	0.02	0.8620696	0.9438123	0.9124855	0.8276538	0.9191474
beta_vec4	0.02	-0.0027197	0.0110750	0.0517034	1.0505263	-0.0067178
beta_vec5	0.02	-0.0045932	-0.0077140	-0.0160055	-0.0010837	-0.0105529
beta_vec6	0.02	0.4154545	0.4039963	0.3839284	0.3662432	0.5432791
beta_vec7	0.02	-0.0406285	0.0146759	-0.0267084	0.0795153	-0.0452543
beta_vec8	0.02	0.1820403	0.1956254	0.3122696	0.2686984	0.1299537
beta_vec9	0.02	2.0666740	1.9595822	1.9458032	1.8233049	2.3142211
beta_vec10	0.02	0.0725989	0.1110710	0.1140780	0.1066402	0.1031025
beta_vec11	0.02	-0.1413081	-0.1589079	-0.1699873	-0.1369779	-0.1387135
g.stat_tr	Inf	137.3082360	135.9975379	116.9462145	135.4900541	114.5090024
auc_te	0.00	0.9913435	0.9923154	0.9849530	0.9826870	0.9743770
g.stat_te	Inf	22.6159703	22.1514224	48.1975784	34.5067570	48.0076306
MSE_test	Inf	4.2783563	4.1138120	4.9735803	5.9718400	6.8409613

instead of using MSE, using pearson chi-square

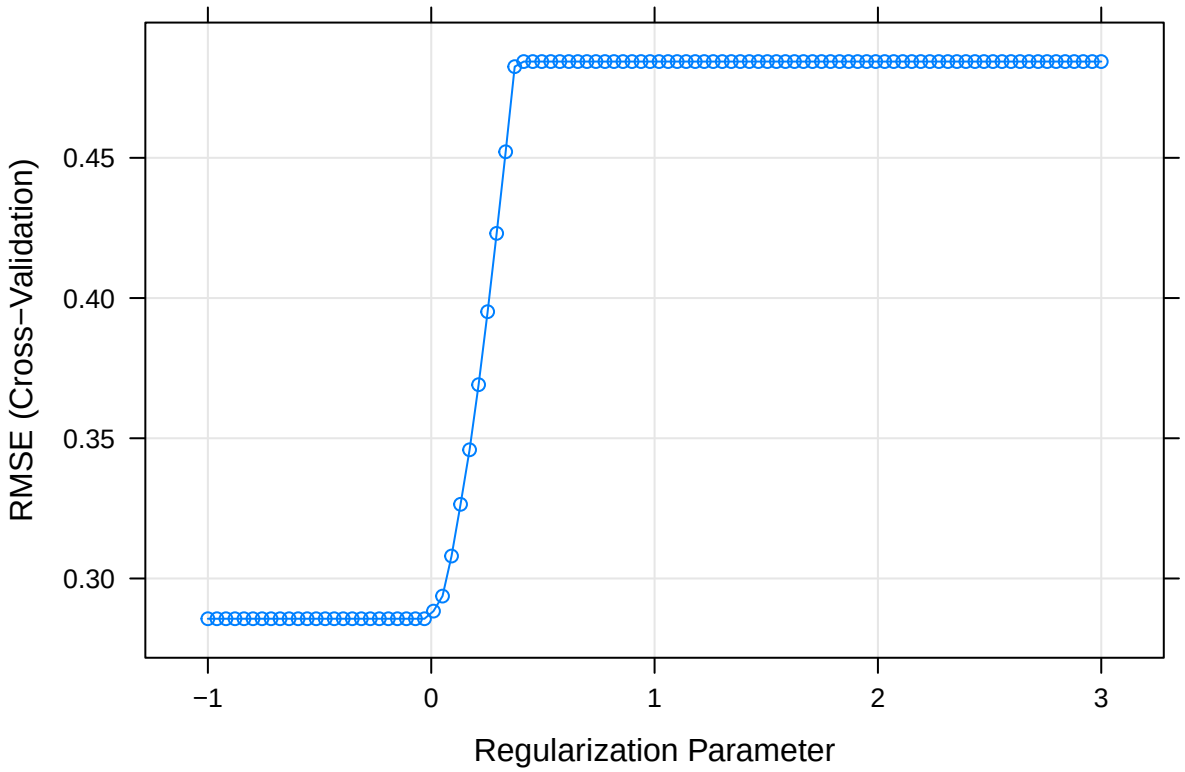
validation

```
x.mat <- model.matrix(diagnosis~., cancer_package[-1])[,-1]
y.class <- cancer_package$diagnosis

ctrl1 <- trainControl(method = "cv", number = 5)
lasso.fit <- train(x.mat, y.class,
```

```
                  method = "glmnet",
                  tuneGrid = expand.grid(alpha = 1,
                                         lambda = seq(3, -1,length = 100)),
                  # preProc = c("center", "scale"),
                  trControl = ctrl1)

lasso.fit$bestTune
```

```
##    alpha      lambda
## 25     1 -0.03030303
```

```
plot(lasso.fit)
```




```
# min(lasso.fit$results$RMSE)
# co=coef(lasso.fit$finalModel,lasso.fit$bestTune$lambda)
# co2=co@x
#
# names(co2)=co@Dimnames[[1]]
# co2 %>% as.data.frame() %>% knitr::kable()
```