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?

```r
# 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	-0.6859925	-0.6846651	-0.5202080	-0.7108130	-0.4904983
beta_vec2	0.02	2.4700201	2.2680094	2.4442414	1.6678653	2.9547735
beta_vec3	0.02	1.5423244	1.6459899	1.6960535	1.5396684	1.8918639
beta_vec4	0.02	0.1086057	0.1309746	0.0933700	1.1913073	0.1548061
beta_vec5	0.02	0.6066107	0.7695696	2.0689442	0.8991037	0.8233475
beta_vec6	0.02	1.0825592	0.9841494	1.3327553	0.9768157	1.6699939
beta_vec7	0.02	-0.5217764	-0.3350776	-1.5663942	-0.2512495	-0.8098719
beta_vec8	0.02	1.0885347	1.2501449	2.3188669	1.5697889	0.9636958
beta_vec9	0.02	2.1922951	1.9632482	1.6208382	1.4671648	2.7033725
beta_vec10	0.02	0.4242812	0.5513924	0.5472317	0.5297819	0.5455744
beta_vec11	0.02	-0.4955606	-0.6031343	-0.0903676	-0.4954602	-0.2339802
g.stat_tr	Inf	238.9887331	270.9964837	166.3436015	344.1339201	229.2555310
auc_te	0.00	0.9934211	0.9916472	0.9811912	0.9844183	0.9747280
g.stat_te	Inf	20.9787985	22.6434852	241.7250075	55.0375829	119.9943143
MSE_test	Inf	3.5422803	3.9095202	5.6402349	5.6143326	8.2576478

instead of using MSE, using pearson chi-square

validation

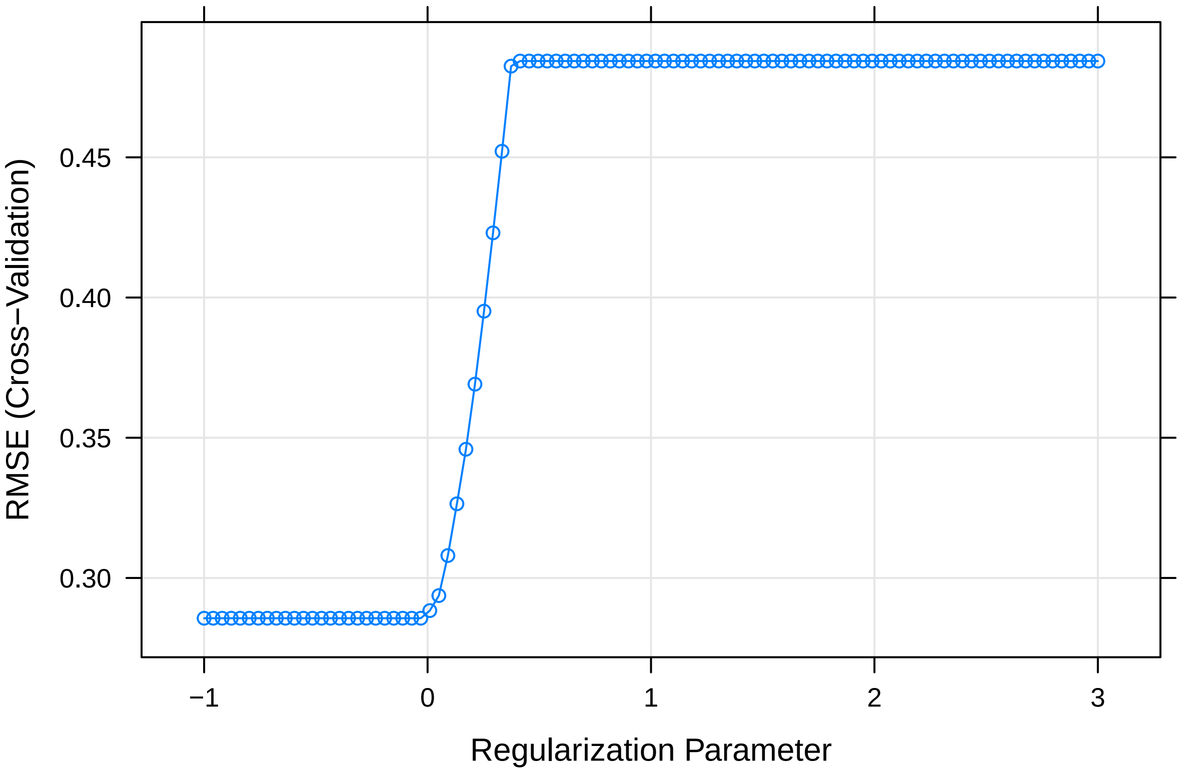
```
x.mat <- model.matrix(diagnosis~., cancer_package[-1])[,-1]
y.class <- cancer_package$diagnosis

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()
```