Forums / Assignments Help

Question on using randomForest for prediction for the course project

Subscribe for email updates. UNRESOLVED

randomForest × + Add Tag Sort replies by: Oldest first Newest first Most popular

john jiang · 9 days ago

I have been trying tree and also randomForest to do the modeling for the course project."pml-training" Dataset was partitioned in training set and test set. Then I build tree model and randomForest model. Everything works well.
With the model , I can do prediction with "predict" function and also check the accuracy(sensitivity,specificity and so on..).

However when I load the "pml-testing" data, using that dataset to do predict with the model I got, I got the following error:

Error in predict.randomForest(activity.rf, temp.t, type = "class") :
  Type of predictors in new data do not match that of the training data

Then I googled the error, found some links like http://permalink.gmane.org/gmane.comp.lang.r.geo/10008, I have read the function code https://r-forge.r-project.org/scm/viewvc.php/*checkout*/pkg/extendedForest/R/predict.randomForest.R?root=gradientforest

```
if (!all(object$forest$ncat == cat.new))
            stop("Type of predictors in new data do not match that of the training data.")
```

But I still not able to figure out what's wrong. Meanwhile I compared my "pml-testing" dataset variable type with the training set , no difference, not able to understanding why with my testing data the predict will give the "Type of predictors" error.

And finally , I tried just use rbind to combine my "pml-testing", with the previous testing dataset I have,then running the predict, then there are no "Type of predictors" error.

Just no able to figure out the reason, can anyone give some hints?
Thanks!

↑ 4 ↓ · flag

john jiang · 9 days ago

Alright, I just go through my code again, found that the problem was caused by including some not needed factor variable in my training model when create the randomForest , with these factor variable in, it caused the problem. This link have more details.
http://permalink.gmane.org/gmane.comp.lang.r.geo/10008, .
With the randomForest, I got 99% accuracy for my validation set. Then for the test set , 100% correction for the project submission. I guess it might be that the instructor have intentionally chosen the test set, as in real world I don't we can get such high accuracy for test set.

I have tried PCA/SVD also to do feature compression, however sounds I am not able to beat the randomForest model. With tree model I can only get roughly 67% accuracy for test set.

↑ 3 ↓ · flag

Brandon Verkennes · 8 days ago

I came to the same conclusions John, regarding the accuracy for validation and test correction. Can anyone else confirm the same? I find it a bit weary that the accuracy is this high for a data set (unless the instructor planned for this).

↑ 0 ↓ · flag

+ Comment

valerio orfano · 8 days ago

HI guys thanx a lot for ur hints. Did u use the caret package or random forest package?

↑ 0 ↓ · flag

+ Comment

john jiang · 5 days ago

I think they are the same package for the randomforest.

↑ 0 ↓ · flag

+ Comment

Patrick Cronin Signature Track · 4 days ago

Same deal here. I am able to load the training set, split it, and come up with a 0.8% error rate, but when I try to apply it to the test set I am getting all kinds of grief. Has anybody has success in applying the training model to the predict function to process the test set any insight would be appreciated.

↑ 0 ↓ · flag

Shaun Brophy · 4 days ago

Can you give some details re. the kind of grief you're having? I'm sure someone will be able to help. It's pretty straightforward getting from where you are already to test set predictions, so it's probably something minor.

↑ 0 ↓ · flag

Patrick Cronin Signature Track · 4 days ago

I am able to build the model and I attempted to create the prediction using the model built with the following:

```
pred <- predict(modFit, testing)
```

□Hide Traceback
□Rerun with Debug
Error in predict.randomForest(modelFit, newdata) : newdata has 0 rows

I then get the error newdata has 0 rows

I am at then end of my tether with this one.

↑ 0 ↓ · flag

Shaun Brophy · 4 days ago

I believe it's saying that you passed an empty data set to predict.randomForest to predict from. Have you looked at dim(newdata)?

↑ 0 ↓ · flag

+ Comment

john jiang · 3 days ago

I think that you might need to check your training model, what's your input to your training set?

↑ 0 ↓ · flag

+ Comment

Christian Grant · 3 days ago

In the following thread people are reporting a 100% accuracy rate, when applying their model to the test set:
https://class.coursera.org/predmachlearn-002/forum/thread?thread_id=119

↑ 0 ↓ · flag

+ Comment

New post

To ensure a positive and productive discussion, please read our forum posting policies before posting.

B	*I*	☰	☰	Link	<code>	Pic	Math		Edit: Rich ▾	Preview

☐ Make this post anonymous to other students

☑ Subscribe to this thread at the same time

Add post