Course Project: How exactly to implement cross-validation?

Subscribe for email updates. UNRESOLVED

R × caret × + Add Tag

Sort replies by: Oldest first Newest first Most popular

Anonymous · 6 days ago

Hi all.

I'm wondering what exactly is meant by the requirement to use cross-validation in the Course Project.

My procedure was as follows:

1. Split the training dataset into a training and a testing data set. (I'll call the test data set with the 20 observations for submission "validation data set" from here on.)
2. Train some models on the training data set using "train" from the "caret" package. Here, I used cross validation for model assessment by using the option "trControl = trainControl(method = "cv")".
3. Calculate the accuracy on the testing dataset.
4. If the model is satisfactory, use it to predict the classe of the 20 observations in the validation data set.
5. Submit 20 predictions for grading (I got all 20 right).

I'm not sure if that is meant by using cross validation or if anybody interpreted the assignment differently?

I don't want to lose points because I misunderstood something.

Thanks for you input.

1 · flag


Sandor Bende · 5 days ago

hi, I think it is about point 3 in your list. If you use the whole data set to calculate the accuracy you might not see the performance of your model on a new data set due to overfitting. To get a more realistic estimate you can split the training data set 40-60 % say, train on the 60% and see how the rest fits .

It is probably the one legit way to chose between several models that give100% accuracy on the training set,mfor example.

BTW, how did you check that your predictions are ok on the test set? I do not see a classe variable there.

regards

0 · flag

Anonymous · 5 days ago

Hi Sandor.

Thanks for your reply. It seems I was on the right tracks.

```
BTW, how did you check that your predictions are ok on the test set? I do not see a c
lasse variable there.
```

I simply uploaded my 20 files after predicting.

0 · flag

+ Comment

Anonymous · 5 days ago

This question has 2 parts:
(1) Do I mis-understand the cross-validation requirement for the Course Project? I interpret cross-validation as a technique that is applied to the training data set (AFTER the initial data set has been subset into training/testing/validation). For this project, i am using the 20 observations data as the validation data set and taking the initial Train data set and splitting it (80/20) into training and testing. I thought the cross-validation requirement was a required option that needed to be applied to each of the models when they are trained?
(2) If cross-validation is a technique that is applied during training, how do i apply it if i use the rpart function to create the decision tree? I decided to use rpart instead of train function because of the nice graph produced by rpart.plot and because train seems to take a very long time on this data set. Wondering if i need to revert to train in order to satisfy the cross-validation requirement...

1 · flag

Anonymous · 5 days ago

After thinking about this some more, maybe there is no need to do cross-validation for the decision tree. In my results, accuracy of the tree is poor (in the 70s) so I am not-so-much concerned about over-fitting. Would be interested in how others are thinking about this....

0 · flag

 Louis Bush Signature Track · 4 days ago

I ended up splitting the dataset they gave us into a train and test dataset. After removing some variables and rows I went about cross validation buy creating my model on the training dataset and cross validating by using the test dataset and the model from the train dataset in the predict function. I then ran that through the predict function and produced a confusion matrix.

I repeated this process another two times for my other models each using a different sample for the test and training routine above.

I had an obvious winner from the three trials but in order to produce my final estimate I took my three predictions and my three sets of actual values from the test dataset and created one final confusion matrix. This was my estimate for the overall process.

This is what I think Dr. Leek was getting at for the week two lectures.....

Either way I got 20 out of 20 for my final model so at least that part was right!

0 · flag

+ Comment

Anonymous · 4 days ago

Actually, I am working on the project at the moment. I am a bit frustrated about crossvalidation: I split the given training data into three (training, test, validation), and I was going for the Combined Predictors approach in week 4 lecture 2. This gave rise to a burden:

Since the combined predictor is trained on the test set, aren't the diagnostics calculated on the test set inflated in the first place (slides 11-12)? This is why I opt to split the training data set into three, test the combined predictors against the validation data set (as in slides 13-14), and use my absolutely final predictors for the 20 cases...

Please feedback on this approach if you feel so, because I don't feel ultimately convinced about it yet.

0 · flag

 Trent Baur Signature Track · 13 hours ago

I'm all for diving into the deep end for the sake of learning. If that's your intent, good on you.

If not, you're making it much harder than it needs to be. Nothing from week four is needed for the project. Keep it simple, the necessary steps are rather straight forward if you want them to be.

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