Manually calculated R^2 doesn't match up with randomForest() R^2 for testing new data

I know this is a fairly specific `R` question, but I may be thinking about proportion variance explained, R^2, incorrectly. Here goes.

I'm trying to use the `R` package `randomForest`. I have some training data and testing data. When I fit a random forest model, the `randomForest` function allows you to input new testing data to test. It then tells you the percentage of variance explained in this new data. When I look at this, I get one number.

When I use the `predict()` function to predict the outcome value of the testing data based on the model fit from the training data, and I take the squared correlation coefficient between these values and the *actual* outcome values for the testing data, I get a different number. *These values don't match up*.

Here's some `R` code to demonstrate the problem.

```
# use the built in iris data
data(iris)

#Load the randomForest library
library(randomForest)

# split the data into training and testing sets
index <- 1:nrow(iris)
trainindex <- sample(index, trunc(length(index)/2))
trainset <- iris[trainindex, ]
testset <- iris[-trainindex, ]

# fit a model to the training set (column 1, Sepal.Length, will be the outcome)
set.seed(42)
model <- randomForest(x=trainset[ ,-1],y=trainset[ ,1])

# predict values for the testing set (the first column is the outcome, leave it out)
predicted <- predict(model, testset[ ,-1])

# what's the squared correlation coefficient between predicted and actual values?
cor(predicted, testset[, 1])^2

# now, refit the model using built-in x.test and y.test
set.seed(42)
randomForest(x=trainset[ ,-1], y=trainset[ ,1], xtest=testset[ ,-1], ytest=testset[ ,1])
```

Thanks for any help you might be willing to lend.

r | correlation | predictive-models | random-forest | r-squared

edited Jul 13 '12 at 13:05
cardinal ♦
15.8k 4 60 93

asked Feb 18 '11 at 2:32

Stephen Turner
1,830 2 17 26

1 Answer

The reason that the R^2 values are not matching is because `randomForest` is reporting *variation* explained as opposed to *variance* explained. I think this is a common misunderstanding about R^2 that is perpetuated in textbooks. I even mentioned this on another thread the other day. If you want an example, see the (otherwise quite good) textbook Seber and Lee, *Linear Regression Analysis*, 2nd. ed.

A general definition for R^2 is

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}.$$

That is, we compute the mean-squared error, divide it by the variance of the original observations and then subtract this from one. (Note that if your predictions are really bad, this value can go negative.)

Now, what happens with linear regression (*with an intercept term!*) is that the *average* value of the $\hat{y}_i$'s matches $\bar{y}$. Furthermore, the residual vector $y - \hat{y}$ is orthogonal to the vector of fitted values $\hat{y}$. When you put these two things together, then the definition reduces to the one that is more commonly encountered, i.e.,

$$R^2_{\mathrm{LR}} = \mathrm{Corr}(y, \hat{y})^2.$$

(I've used the subscripts LR in R^2_{LR} to indicate *linear regression*.)

The `randomForest` call is using the first definition, so if you do

```
> y <- testset[,1]
> 1 - sum((y-predicted)^2)/sum((y-mean(y))^2)
```

you'll see that the answers match.

edited Feb 18 '11 at 4:21

answered Feb 18 '11 at 3:31

cardinal ♦
15.8k 4 60 93

1 +1, great answer. I always wondered why the original formula is used for R^2 instead of square of correlation. For linear regression it is the same, but when applied to other contexts it is always confusing. – mpiktas Feb 18 '11 at 8:22

(+1) Very elegant response, indeed. – chl ♦ Feb 18 '11 at 9:14

@mpiktas, @chl, I'll try to expand on this a little more later today. Basically, there's a close (but, perhaps, slightly hidden) connection to hypothesis testing in the background. Even in a linear regression setting, if the constant vector is not in the column space of the design matrix, then the "correlation" definition will fail. – cardinal ♦ Feb 18 '11 at 13:26

thanks! this really helps! – Stephen Turner Feb 18 '11 at 20:10

If you have a reference other than the Seber/Lee textbook (not accessible to me) I would love to see a good explanation of how variation explained (i.e. 1-SSerr/SStot) differs from the squared correlation coefficient, or variance explained. Thanks again for the tip. – Stephen Turner Feb 18 '11 at 21:25