We've made a few improvements to the forums. You can read read more on the blog.

Forums / Data Analysis Assignment 2

How I achieved 87.5/90 - walkthrough of analysis with final report and code

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

DataAnalysis × assignment × + Add Tag

Thia, Kai Xin · 6 days ago

Hi all,

Seems like many people found my previous thread on my assignment 1 pretty useful. I have learnt a lot from the feedbacks of everyone as well, thank you so much =)

Here's my code and report:

- Report (Note that the PCA part is wrong, see walkthrough below)
- Code (take note that I commented away the parallel processing sections. If you have a multi core machine, remember to uncomment them before running the code)

FYI: There is an interesting in house Kaggle like competition going on using the data for this assignment

FYI 2: If you like to learn more about data analysis, I have put together this website that collects all the best materials (ebooks, online courses) that I can find (from MIT, Stanford, Caltech, Coursera, professors' sites etc), all completely free:Check it out: Whizage.org

FYI 3: Check out our TA, Anne's report. I made the assumption that the trials are independent. That's a bad assumption (read her paper to find out way) but I redeemed myself by using caret's confusionMatrix() which is pretty robust (see walkthrough below + discussion)

FYI 4: There is a discussion on black box vs white box A black box will be something like random forest that has good accuracy but hard to explain while a white box will be like decision tree, not as accurate at random forest but easy to explain and plot.

Basic Initializing

- First, I created a color blind friendly palette for my graphs
- Multi core processing code was deployed to speed up random forest (I ran 2001 trees in about 3mins on my quad core machine). Multi core does not increase accuracy, just makes things faster as I hate waiting =P

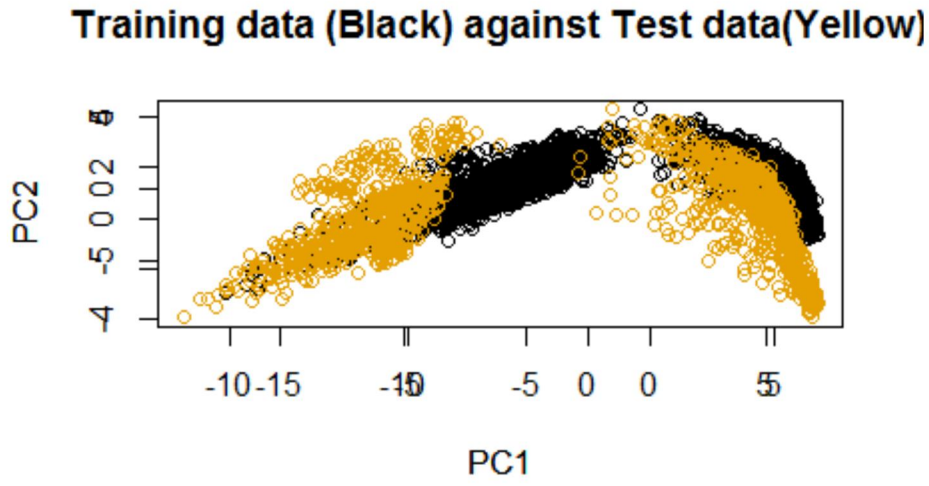
- Basic renaming using gsub("[[:punct:]]"…) to remove all punctuations from the column names (see code)
- Convert activity and subject to factor
- Use subjects <27 as training, rest as test

Advance Initializing

- I noticed that within the training set, subject 25 had the highest observation count of 409, 45% more observations compared to subject 8 which had the lowest observation count of 281. To prevent any individual subjects from skewing the data (in case some subjects had weird moving habits or their device was not tuned correctly), I did a random sampling across the training data again, this time ensuring every subject in the training data had equal counts of observation, 281.
- I further subsetted ¼ of the training data to be validation data via one last random sampling and drop subject column as subject number should not be factored into analysis.
- **Important Note**: This section below on PCA is wrong. I should not have done PCA on the test data as it will tend to overfit the data (remember test data MUST always be locked away from everything at start of analysis, even from my human eyes/brain). A much better way to use PCA may probably be to generate ideal independent data points and compare my training data to that generated data instead of test data. By doing so, I can see what the distribution of truly independent data points will look like and if my training data has any resemblance of independence.

Kept original wrong section below for your reference, don't do this

I needed to check if the training data and test data had similar distribution. This is important as most modeling methods like decision trees, K-nearest neighbors etc works best if training and test data have similar distribution. By using Principal Components Analysis, I summarized the complex data set of over 500 features into 2 principal components that represent about 90% of the data distribution. Fig1 below shows the plot of the first two principal components of training data (black) against first two principal components of test data (yellow). As we can see, the two data sets had similar trends and overlapped each other. This meant that training data and test data had similar distribution.

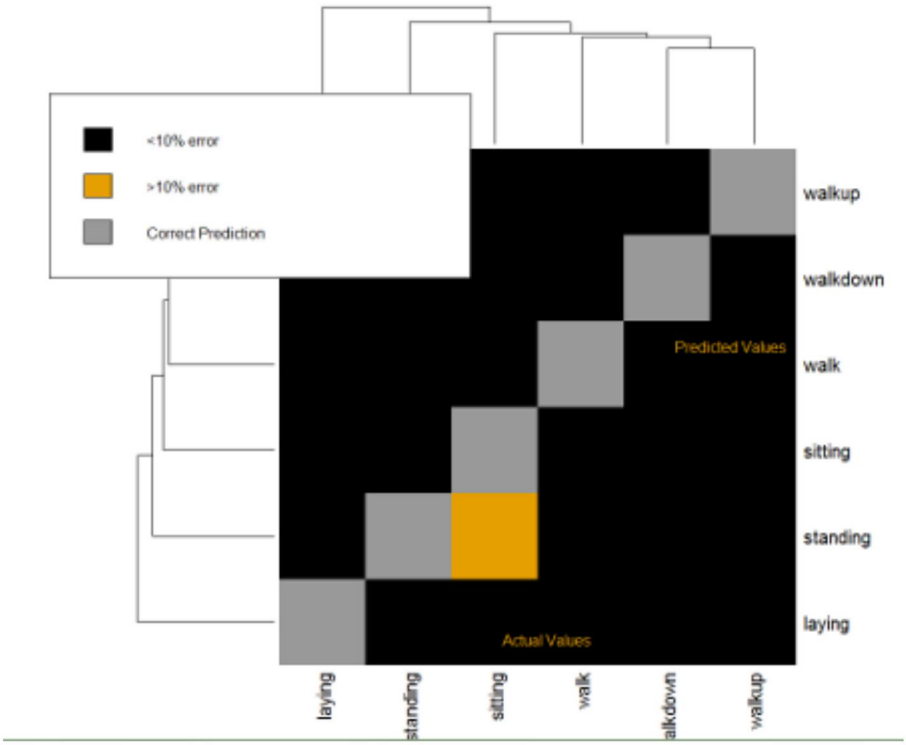


Modeling:

- My final model is a random forest with 2001 trees (odd number of trees help break tie votes), "mtry" = $\sqrt[3]{}$(number of features). I arrived at it via trying different combinations using the validation set (see code).
- I did no feature selection as I want to use as much data as possible, based on the idea of "unreasonable effectiveness of data by google" (see report)
- I decided that simple, single validation was sufficient and cross validation was not needed as long as I set a large enough number of trees and subsample the data well in the random forest; both techniques should result in a sufficiently unbiased random forest.
- I created a benchmark random model for comparison (see report and code)

Result:

- Created a confusion matrix using the caret library confusionMatrix() command
- I obtained an overall accuracy of 96.16% with my random forest model.
- (Note: A much better way will be to calculate the individual subject error rates of each subject vs the overall error rate I have below, this would have helped me identify subjects that our model have problem with) We had a very small P value and much higher accuracy than our benchmark, which signaled that our model was significantly better than random guessing. The accuracy of the model was between 95.06% and 97.08% at 95% confidence interval. Finally, we saw that most of the activities obtained close to 100% prediction; except for subjects who were actually sitting, as there was more than 10% chance that we would incorrectly predict them as standing.
- Final diagram for me was a heatmap



P.S If you are too shy to post your questions here, you can add me on linkedin and drop me a mail: http://www.linkedin.com/in/thiakx

66

Monika Jakubczak · 6 days ago

Thanks for that! Really useful:) And amazing work btw!!!!!:)

7

Deepak Chaturvedi · 6 days ago

Congratulation ,I have gone through your analysis ,really great..

2

Michael Lane · 6 days ago

Thanks for sharing your Assignment 2 - great approach to analysis

1

Dana Barberio · 6 days ago

Thanks, Thia, this is very useful to see your analysis and code.

1

+ Add New Comment

Diego F. Pereira-Perdomo · 6 days ago

Great!

Congratulations and thanks for sharing it.

6

+ Add New Comment

Przemysław Maciej Jura · 6 days ago

Thanks. Very useful.

2

+ Add New Comment

Anonymous · 6 days ago

A little modesty in the title wouldn't have hurt. You seem to have gotten graders who could read as well. Not everyone was so fortunate.

-21

Monika Jakubczak · 6 days ago

Haters gonna hate ;)

12

Diego F. Pereira-Perdomo · 6 days ago

Oh, I didn't like the grade I received, but who cares???

Let's feel happy for Thia for a moment and learn from his analysis.

He is being very kind in sharing it with us.

25

+ Add New Comment

Anonymous · 6 days ago

A very nice report - it was a pleasure to read.

2

+ Add New Comment

Steven O'Neill · 6 days ago

Thanks for providing this to all of us.

3

+ Add New Comment

Anne Paulson COMMUNITY TA · 6 days ago

Do you believe that "accuracy between 95.06% and 97.08% at 95% confidence interval" is correct? I don't see how you can possibly make such a claim on the basis of testing the data from only four subjects. I personally think that confidence interval is preposterous. Moreover, if you, for example, put subject 16 in the test set and subject 30 in the training set, you would do the same procedure, produce more or less the same trees, and discover a confidence interval that didn't even overlap with 95%-97%.

3

Thia, Kai Xin · 6 days ago

Good question. Personally, I see them as a bunch of individual trials rather then subjects trials (so i am looking 1485 independent trials rather than 4 subjects so my claim can be statistically valid) Are the trials really independent...well that's a tough question, I was a little lazy to explore that for my test data haha (for training at least I tried to take equal samples from each subject so no 1 subject will skew the model too much)

I didn't know about the subject 16 and 30 in training set thing...haha I am a lazy person. Anne I know you did a crazy lot of cross validation. Perhaps you can do a writeup of your own and show us your validation steps?

2

Anne Paulson COMMUNITY TA · 6 days ago

What I'm asking is, do you really believe that your tree would have 95-97% accuracy 95% of the time with any four subjects? Check my thread where I posted my analysis, and you'll see that the test subjects were exceptionally easy to model.
https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=3664

If you think the test set observations were independent, I have a bridge to sell you. It's lovely, goes from Brooklyn to Manhattan, low low price.

5

Gareth Robins · 6 days ago

I would like to see the submissions from all the TA's and Prof Leek himself, so that we might learn from our mistakes and their ingenuity

4

Chandramouleeswaran Srinivasan · 6 days ago

I presume you get this confidence interval for accuracy by using the function 'confusionMatrix()' from package:caret If so, these results use a one-against-all approach for a multi-class classification. So it doesn't really depend on the number of observations or the subject. Rather, confidence intervals for overall accuracy is reported by taking into consideration accuracy for each class.

If you have an alternative explanation, please let me know.

▲ 2 ▼

Anne Paulson COMMUNITY TA · 6 days ago

I was assuming the confidence interval came from bootstrapping.

▲ 1 ▼

Chandramouleeswaran Srinivasan · 6 days ago

Usage is something like this: confusionMatrix(predicted.testClasses, actual.testClasses)

▲ 0 ▼

Thia, Kai Xin · 6 days ago

Anne is right in the sense that I used confusionMatrix() with the assumption of independence among data. And her paper pointed out correctly that Bill's 20 trials on walking will be based on Bill's walking habit, which may not be useful to predict Sally's walking behavior.

I was just lucky that caret's confusionMatrix() is a little more robust than I thought

▲ 1 ▼

Larry Cahoon · 6 days ago

Compute the error rate by subject with the test sample. Is that anywhere near the confidence interval you got? My error rates were widely spread - 7.7% overall, but 2.6% for one subject and 9.9%, 10.5%, and 10.2% for the other three.

▲ 2 ▼

Thia, Kai Xin · 5 days ago

Hi Larry, good idea to calculate error rate by subject. I did not do that as I just lumped all the trials together, assume (statistically naively) that all trials are independent and calculate the lump sum error rate across all the trials. I was being stupid and I should have at least calculated the individual error rates for each subject and compared them

to this lump sum error rate that I have above.

1

+ Add New Comment

Boran Altincicek · 6 days ago

Thank you very much. It was a pleasure to read and very helpful.

1

+ Add New Comment

 Walter Hop · 6 days ago

Interesting approach! Thanks for posting. How did you arrive at p-values?

1

 Thia, Kai Xin · 6 days ago

Hi, I used the caret library's confusionMatrix() command to get the p-values. Check it out, caret is a great library. From the help file: "The overall accuracy and unweighted Kappa statistic are calculated. A p-value from McNemar's test is also computed"

1

+ Add New Comment

 Anne Paulson COMMUNITY TA · 6 days ago

Nice analysis.

1

Michael J Hannon · 3 days ago

Linear Discriminant Analysis gave me 97.9% correct on the 4 subject test set. Try it yourself:

library("MASS") ldaMdl <- lda(factor(activity) ~ ., data=trainSet, na.action="na.omit")

Then predict against your test set.

http://en.wikipedia.org/wiki/Linear_discriminant_analysis http://www.statmethods.net/advstats/discriminant.html

0

Robert Blanford · 3 days ago

Tidy. What led you to find it?

0

Anne Paulson COMMUNITY TA · 3 days ago

Why does this work so well? LDA assumes that the independent variables are normally distributed. Are they? I didn't think they were.

0

Michael J Hannon · 2 days ago

I really just stumbled on LDA looking for something else as a possible technique. I had already turned in my paper with RandomForest as my go to approach. Like Anne I questioned the whole normal distribution on the variables. But perhaps the data does follow at least a bell shape curve. For instance I kind of have a normal walking speed sometimes a little slower, sometimes faster but usually centered on some speed. So perhaps the other variables followed a similar pattern. LDA works when the measurements made on independent variables for each observation are continuous quantities. So this fit. Anyway I decided to give it a try. Trained on everything but the 4 test subject set. Low and behold it gave great results on the test set. I'm no expert on LDA in fact I wasn't aware of it, so maybe someone else wants to chime in on why it worked so well.

0

+ Add New Comment

Satyendra Srivastava · 6 days ago

Thanks Thia- I enjoyed reading your paper and checking out your code as well.. Many thanks for sharing. And dont worry about modesty. Artificial modesty is no good..

3

Thia, Kai Xin · 6 days ago

Thanks for your support =)

2

+ Add New Comment

Ran Locar · 6 days ago

Great analysis. How did you calculate your P values in this case?

1

Thia, Kai Xin · 6 days ago

*Copied from above, similar qn: I used the caret library's confusionMatrix() command to get the p-values. Check it out, caret is a great library. From the help file: "The overall accuracy and unweighted Kappa statistic are calculated. A p-value from McNemar's test is also computed"

1

+ Add New Comment

Eoin P Sharkey · 6 days ago

Thanks for sharing that. Very nice work, well explained and with useful graphs.

I am interested in the PCA graph.

Did you make this graph * a priori* , i.e. before building and cross-validating the model or *a posteriori * i.e. after building and testing the model on the test data-set. ?

I am interested to know to what extent one can analyse test data-set without it being a violation of the (now sacred-to-me) principle of locking away the test data-set.

0

Thia, Kai Xin · 6 days ago

Good qn. Hmm I did PCA on test data, I interpret that the sacred no-touch-testData rule basically means you should not train/test your model on the test data but it is ok to do exploratory analysis on the testData. So I counted my PCA as pre cross

validation exploratory analysis to help me decide what modelling techniques will be suitable to handle similar test data sets. This does expose me to overfitting, so I will need an algo like random forest and do many repetitions to downplay the odds of overfitting.

1

Anne Paulson COMMUNITY TA · 6 days ago

It's actually not OK to do exploratory analysis on the test data. And the reason is a funny one: your brain is a part of your modeling. So if your brain picks one model over another based on the test data, then you have used the test data to choose your model, and your test error rate is no longer valid.

You're supposed to be testing your model on unseen data. Unseen by you, too. And this really can make a difference.

7

Thia, Kai Xin · 5 days ago

I sat down to think about it..yes you are right, I will edit my walkthrough above. A much better way to use PCA will probably be to generate ideal independent data points and compare my training data to that generated data instead of test data.

2

+ Add New Comment

Anonymous · 6 days ago

Thanks for sharing that. You did put a lot of effort for this assignment and knew better too. With my limited time, English and knowledge, I should be pretty happy with grade 82. A simply work and approach, actually. Just a weird feeling that some graders assigned score 4 for the first evaluation question while I literally have them all in the text.

1

Thia, Kai Xin · 6 days ago

Maybe you will like to put up your report / code to share as well? We should help each other learn on the forum

1

Anonymous · 5 days ago

You could find it here: https://dl.dropbox.com/u/76774311/Final%20Text.pdf I didn't put the codes because I messed up the seeds - sorry, not so organized but this is a good lesson to have tidy work next time.

0

+ Add New Comment

 Mykola Dolgalov · 6 days ago

Thanks for sharing. I god 86 this time (79 last time) and also shared my work, I'd also appreciate your feedback on my work.

0

+ Add New Comment

Tony Ha · 6 days ago

Thanks for sharing. It is a very good report! I run your parallel R code on by my AMD's 8 cores windows 8 PC. it only managed to utilize 40% of the all CPUs. i.e. 2.8 CPUs. Do you have a better utilize figure?

0

 Thia, Kai Xin · 6 days ago

Ah sorry I made a mistake. i set foreach(ntree=rep(round(667,3), ...

which means it will always only use 3 cores. the key is to update this part of the code (i have updated the code, you can redownload from above or change this yourself):

foreach(ntree=rep(round(2001/coreNumber),coreNumber), ...

0

Tony Ha · 5 days ago

The change make use of all my 7 cores! i.e. 90% of all my CPUs. Thanks again for sharing the code, now I know how to utilize multi-cores to run regressions!

0

+ Add New Comment

 Mykola Dolgalov · 6 days ago

I have 2-core Intel Core i5 with multithreading notebook, and the 4 virtual CPUs showed utilization no more than 40% during calculations. I have Windows 7, 32 bit edition.

0

 Thia, Kai Xin · 6 days ago

Same as above:

Ah sorry I made a mistake. i set foreach(ntree=rep(round(667,3), ...

which means it will always only use 3 cores. the key is to update this part of the code (i have updated the code, you can redownload from above or change this yourself):

foreach(ntree=rep(round(2001/coreNumber),coreNumber), ...

0

 Thia, Kai Xin · 6 days ago

Remember to uncomment the parallel processing sections of the code

0

+ Add New Comment

gadfly1974 · 6 days ago

Bravo!

0

+ Add New Comment

⬇ scroll down for more ⬇

PDF-XChange Click to buy NOW! www.docu-track.com