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

Forums / Data Analysis Assignment 2

Another posted analysis

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

No tags yet.+ Add Tag

 Anne Paulson COMMUNITY TA · 6 days ago

Analysis: https://docs.google.com/file/d/0B_bBMWttZpsKZlF1XzByS2QxNVU/edit?usp=sharing

Figure: https://docs.google.com/file/d/0B_bBMWttZpsKbW0wVUVLVzZDc28/edit?usp=sharing

Caption: https://docs.google.com/file/d/0B_bBMWttZpsKNnNRd0xPOTk1QUU/edit?usp=sharing

It's a bit of a different take on the assignment.

28

 Thia, Kai Xin · 6 days ago

Got code?

2

 Anne Paulson COMMUNITY TA · 6 days ago

Nothing pretty enough to post.

-5

 ANIMESH KUMAR · 6 days ago

Then post the dirty one pleasse! It will be very useful. Thanks Anne.

5

Soheila Dehghanzadeh · 6 days ago

Anne have used cross validation to pick best hyper parameters, could somebody kindly explain how to do that?

0

Anonymous · 6 days ago

I like the discussion of the cross-validation technique in this report: removing the data for a subject from the training set and using them in the validation set, so that the mis-classification error is estimated more realistically. I made the cross-validation mistake described: I took random samples of data from all subjects in the training set to form my validation set, and it produced error estimates which were much lower than what I saw later with the test set. I realized my mistake too late.

0

Anne Paulson COMMUNITY TA · 5 days ago

Realizing the mistake is key, no matter when it happened. I made several mistakes in my analysis. Happily, I have now realized some of them and I won't make those mistakes again.

0

+ Add New Comment

Przemysław Maciej Jura · 6 days ago

Thanks a milion!

3

+ Add New Comment

Diego F. Pereira-Perdomo · 6 days ago

Thank you Anne!

2

+ Add New Comment

David Hood · 6 days ago

Here is my one: https://www.dropbox.com/l/CiXlgHwO3XfvivABgELnRc (pdf file)

I could post the code, but to be honest the interesting stuff is in the exploratory code, and the useful bits of that have been reused in the subject effects thread I've been posting in.

7

Diego F. Pereira-Perdomo · 6 days ago

Thank you David!

1

Anonymous · 6 days ago

thank you David. It would be helpful if you could also post the code.

0

Prokhorov George · 6 days ago

Code, please.

0

Mary L Howard · 6 days ago

At least you mentioned, David, that there was a subject effect that was not accounted for in the model- that the data had repeated measures within subject, but was a predictive model, was the most frustrating part ot the analysis to me; I wanted to say, "really we can't use this data to predict new subjects unless we gather a lot more data with different subjects", but I in the end didn't do that :-)

0

David Hood · 6 days ago

Since people have been asking for my assignment code, it is at:

https://www.dropbox.com/s/31mwb3pp6emtd8k/finaldaa2.R

But this is just the final clean code of the model. For those wanting to learn from it, I think you would learn a lot more if you have a look at the exploratory process at how I get to the final version of the prediction tree, and for that see the Subject Effects thread where I have been redoing the exploratory work focusing on how to deal with Subject (and indeed use it to improve the model).

https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=3494

Seriously, it is more useful. And hopefully in the next 36 hours I will have the time to get together a final model built off that.

0

+ Add New Comment

Boran Altincicek · 6 days ago

Really great piece of work. I enjoyed it and learned a lot.

p.s. The reference no. 7 is just perfect for this assignment.

Thank you Anne.

^ 1 v

Anne Paulson COMMUNITY TA · 6 days ago

I thought reference 7 was a serious, scholarly look at the subject;)

^ 0 v

+ Add New Comment

Satyendra Srivastava · 6 days ago

Thank you Anne! As you say- a different take on the assignment; but full of interesting insights. Can we request for the code too?

^ 0 v

+ Add New Comment

Marius Mather · 6 days ago

Very nice Anne. I like the conversational writing style, I get frustrated with the round-about passive voice writing in science but can never quite bring myself to fully break with convention, so it always tends to sneak back in.

You've definitely put a lot of thought into how to deal with the non-independent nature of the data in cross-validation. I can definitely see the merits of your approach, but on the other hand I think your method might allow strange subjects to push the parameters around too much- it seems like by doing 17-fold cross-validation, you are giving equal "weight" to each subject, so odd subjects will be able to influence the parameters just as much as the subjects who have more or less the same activity patterns. Maybe you could combine the by-subject K-fold validation with a normal K-fold somehow?

^ 2 v

Anne Paulson COMMUNITY TA · 6 days ago

Our best guess from the training data is that there are lots of idiosyncratic subjects. We have to assume that the training subjects are drawn randomly from the population, and that other unseen subjects will be similarly varied. We have no reason to believe we've seen most of the variation already. I refer you to the serious scholarly work from M. Python for more on that issue ;)

^ 1 v

Janis Sutherland · 5 days ago

The reason I give you an point up on everything you post!! You are a peach!

0

+ Add New Comment

Martin Müller · 6 days ago

Anne,

Did you investigate in what activities the error rates occured? I noticed that mergind standing/sitting lead to a substantial decrease of the classification rate in training/test data. Moreover the difference in error rate between training/testing was reduced to around 10% and perfect classification of lying/standing-sitting was achieved (average misclassification rate of 0.44% for training and 0.54% for testing no matter which subject was in which data set). So one could argue that the problem with predicting the activity is not the subject, but simply the definition of the activity.

1

Anne Paulson COMMUNITY TA · 6 days ago

I had a whole other section on activities ready to add, but instead I had to cut to get down to 2000 words. And yes, the sitting/standing distinction was most of the errors.

0

Rose Kudlac · 6 days ago

Thanks for the sharing and discussion of assignments. Very helpful.

I presented the misclassification by subject by plotting the predicted activity by the observed activity, differentiating subjects by color. https://docs.google.com/file/d/0BzHgpVMulpneNDBnR2gxLUpWaG8/edit

plot(jitter(as.numeric(as.factor(test$activity))),jitter(as.numeric(as.factor(randomtreepredictionontest))),pch=19,col=test$subject-13,cex=0.5)

In each of the plots in my figures, the misclassifications differed by activity AND subject, so it looks like the prediction problem with my models is more than "the definition of the activity". I noted in my conclusion that there could have been a problem with recording the activity as this was done manually (per the documentation).

0

+ Add New Comment

gadfly1974 · 6 days ago

Wow, how interesting analyzing each subject separately.

This gets to the heart of the dependent nature of this particular data set.

Great work Anne! Thanks for your enthusiasm and passion.

1

+ Add New Comment

Salem L. Engler · 6 days ago

I looked at subject variability in my report as well but came at it from a bit of a different angle. I wanted to see how well each subject's data (from my training set) individually could be used to predict my overall training and validation data. To do that, I broke out each subject's data and created random forests based on those data. From there, I used the resulting subject-specific random forest models to predict the overall training and validation data. The results were fairly interesting, I think. For instance, with predicting "walkup", almost all the subjects are very bad at being useful for predicting the other subjects, except for subject 5 who was able to be used to predict with only 5.3% error all of the training data's "walkup" activities. Or subject 25, whose data were able to predict with 100% accuracy all of the training subjects' "walkdown" activities.

If you're interested in the results, here's my paper: https://coursera-uploads.s3.amazonaws.com/user-a6b6027d40838ed73c1ad6b9/294/asst-5/294-513ce7e533f7e0.88237786.pdf

Be wary, though... somehow my reviewers thought it was only worthy of a 58.5 score. Sigh.

3

Gundas Vilkelis · 6 days ago

The analysis is thorough and has some really interesting findings. I wonder how and why would anyone score it as low??

1

Salem L. Engler · 6 days ago

Good question :) Given the amount of time and effort I put into the thing, I was shocked, to say the least, to see the final score. Luck of the draw, I guess.

1

Prokhorov George · 6 days ago

Code, please.

0

Anne Paulson COMMUNITY TA · 6 days ago

I don't agree with that grade at all. Most of the papers I've seen seem to have more-or-less the right grade, but your grade is way too low. Was there an issue with your figures?

0

Salem L. Engler · 6 days ago

For the figure and caption, I submitted what I have in my report as "Table 1". After reading some of the other forum posts, it seems there was a healthy debate brewing before about whether a table should be considered a figure or not. Maybe I got reviewers that thought the only thing possible to submit for a figure was some sort of plot, even if it wouldn't have made much sense in the context of what I was writing about... not to mention the word limit that I ran up against. Oh well.

0

Salem L. Engler · 6 days ago

Some of the scores: * Are figures labeled and referred to by number in the text? 2.5 * Does the analysis report any missing data or other unusual features? 2 * Does the analysis include description and justification for data transformations? 1.5 * Does the analysis include a discussion of potential confounders/unmeasured variables that could hurt prediction? 2 * Are the prediction models correctly applied? 4 * Are estimators/predictions appropriately interpreted? 2 * Does the analysis make concrete conclusions? 1.5 * Does the analysis specify potential problems with the conclusions? 3 * Is the figure caption descriptive enough to stand alone? 2.5 * Does the figure focus on a key issue in the processing/modeling of the data? 2.5 * Are axes labeled in plain language and large enough to read? 2.5

0

Anne Paulson COMMUNITY TA · 6 days ago

You can look at your grade and see where you lost the points. Is that where you lost all those points? Your grade is a mystery to me, and I would have given you a much higher grade based on that writeup and those tables.

0

Peter C. Fontana · 5 days ago

I didn't grade this one, but I can see certain things that I would have docked off for:

==Does the analysis report any missing data or other unusual features?==

You say:

"The data ultimately used in this analysis was comprised of the original study's training data set which included 7,352 records, each with 561 different measurement values, along with activity label and subject number, from 21 of the original 30 subjects."

However, you did not state why there were only 21 out of the 30 subjects. I consider the non-continuous use of subjects to be an unusual feature.

====Does the analysis include description and justification for data transformations?====

Also, later you say:

"From this analysis, decisions were made as how best to transform the raw data to be of best use for the decision tree analysis. The analysis aided in identifying any missing data and ensuring data integrity."

First, you needed to mention that there were no missing values (this is a 1-line R code to check). Also, since you gave this preface I expected you to list out the transformations that you did do. When reading your document, I don't see any transformations that you applied (other than partitioning the data into a training, test and validations set)

Also, given the rubric wanted cross validation instead of a validations set. I wouldn't dock off for this. However, I see training set and validation classification results, but where are the results for the test set classification? Note: the "training results" are not that relevant due to overfitting (as you mention). I consider the "validation" set to be the "training set" report. I would have also liked to see a confusion matrix (correct by activity, but not subject), but I would not dock off for that.

===Is the figure caption descriptive enough to stand alone?===

Given the previous item, I can understand others considering your caption as insufficient to stand alone. I consider "training set" rates to be compared against the validation set, not the evaluation of how the data did on the set it was trained on.

I was a rather hard grader though (I didn't grade yours)...

Best Wishes, Peter Fontana

^ 1 v

Salem L. Engler · 5 days ago

Many thanks for the feedback, Peter.

I don't understand your comment about the "non-continuous use of subjects". Given that all of the papers I reviewed didn't even mention the fact that the data set we were given didn't actually have 30 subjects in it, I'd like to think it was good to at least have it mentioned in my report. Do you mean that I should have written that our professor thought it best to provide us with only the training set in the RDA file?

As for data transformations, yes, I should have spelled it out more clearly for the graders. I guess I don't like writing things just to write them when they don't really add anything to the discussion. I think it would seem odd if I were to read a real scientific paper and they were to say that there were no missing data and they used as.factor() to transform the activities. But I don't read scientific papers (nor write them, obviously), so maybe they actually do say things like that.

The results for the test set are in the final paragraph of the results section. I probably should have made them into a table to keep things consistent.

I'd be interested to know what grade you would have given me if you were to have received my report. Do you feel that a 58.5 was appropriate?

Again, thanks for the feedback.

^ 0 v

Peter C. Fontana · 4 days ago

I think so. Since the subjects were not 1-21, I found that unusual.

I understand. My point was that if you say that you use exploratory analysis to identify transformations to use, my question is, "what transformation was appropriate?" and, "what did you learn from your exploratory analysis?". If there was nothing to report from your exploratory analysis, then you could have omitted that section.

As for grading, here are my thoughts on some of the categories.

" Are figures labeled and referred to by number in the text?"

If table 1 is your figure, then I would give you 5/5.

"Does the analysis report any missing data or other unusual features?"

I think the score of 2 is reasonable.

"Does the analysis include description and justification for data transformations?"

While I don't see any transformations that you made, I would probably give a 3/5 since there no glaring transformations that I needed.

- Does the analysis include a discussion of potential confounders/unmeasured variables that could hurt prediction? *

This one I am not sure of.

Are estimators/predictions appropriately interpreted?

I would give a 3/5. I don't think your "training set" report is worth including, and on a first read, could be confused with a cross-validation report.

"Does the analysis make concrete conclusions?"

Hard to say. It's definitely worth at least 2 points, but I would probably give you 5/5. Ideally I would like the conclusions to have more support (just because they did better for your set on the validation set) does not necessarily mean that it generalizes.

- Does the analysis specify potential problems with the conclusions?

I'd give a 3/5 for this one.

- Is the figure caption descriptive enough to stand alone?

Is the figure Table 1 and its caption? If so, I would give a 4/5. The fact that "training set" did not mean "cross-validation set" or the "validation set results" means that it is hard for it to stand alone without the report.

I would give a 5/5 for the other two figure categories.

Overall, it seems that you were graded a bit low; I would have given you a higher mark. Since I don't know what "standard" to calibrate grades to, it is hard to say what is appropriate.

Best Wishes, Peter

0

+ Add New Comment

Yuen King Ho · 6 days ago

Hi Anne,

If you've used *glmnet* with default setting for alpha, you might want to note that you are actually using *lasso* instead of simple multinomial regression. It does some sort of automatic variable selection for you.

0

Anne Paulson COMMUNITY TA · 6 days ago

I thought LASSO was a kind of L1 regularization.

0

Yuen King Ho · 6 days ago

Yes, it is. glmnet uses L1 regularization by default.

I am not saying that using L1 is wrong, just that your report implied that you had used simple multinomial regression instead of regularized multinomial regression.

0

Anne Paulson COMMUNITY TA · 6 days ago

You're right, King. I had always intended to regularize, but I didn't fully understand the parameters of the function. For a long while I thought that lambda was the regularization parameter, rather than the learning rate.

0

+ Add New Comment

DICKO Ahmadou · 6 days ago

+1 It's funny because I used also svm (kernlab) and glmnet in my assignment. I enjoyed reading it, it was well written and I congratulate you for respecting the KISS principle..

0

+ Add New Comment

Anonymous · 6 days ago

Hi Anne, What was your score? Thank you.

2

A post was deleted

Anne Paulson COMMUNITY TA · 6 days ago

I got an 81. I got only a 2 for the confounders criterion, which I thought was funny, since the whole analysis is about confounding.

2

Rangarajan Sreenivasan · 6 days ago

Ditto!! I got dinged on the confounder criteria as well. However, I did explicitly state in a few sentences that since this is not a "causal" analysis we don't have to look for confounders in this report. Perhaps my understanding is incorrect.

1

Larry Cahoon · 6 days ago

Anne, I don't think the class ever understood the concept of confounders. All the discussion seemed to focus around variables they had but did not use. So they claimed they solved the problem by their choice of variables. There seems to be very little understanding that the far bigger problem is the variables I don't have. So subject or subject characteristics as confounders just did not cut it. I got hit to, but at least I got a 3 on that one, and like you I focused there on the subject issue specifically mentioning age, sex, and weight of the subject as confounders.

1

Anne Paulson COMMUNITY TA · 6 days ago

I agree that a solid definition of confounders was never offered by the instructor, and as far as I can tell, the word is used in different ways by different actual practitioners of data analysis. I wish we had had more coverage of the issue, including a specific definition that we were to use, with examples, and particular emphasis on your point that "the far bigger problem is the variables I don't have". This was the source of endless confusion and contention; I hope the issue is addressed for the second round of the course.

2

David Hood · 6 days ago

I deliberately wrote mine spelling out "I consider these to be confounders because of evidence X, to address that I did Y". This seemed to pay off, with a 4.5 from peers for that one. I pretty much had a policy of make it clear for the peer.

2

+ Add New Comment

Alex Kumenius · 6 days ago

Hei, everybody, I have been pretty quiet during this Course. However, I have closely followed, read, laugh -- in silence -- with all of you guys. It has been a tremedous course. I learned so much, that knowledge overwhelms me, hehehehe. The lectures and professor Leek have been outstanding. TAs have been the soul and guidance of the course, Anne Paulson you are superb. Anyway, the "icing on the cake" has been for TAs and classmates to share their assigments with us. Sincerelly, rigth now, I have the right yardstick to know my mean, standard deviation, and my error rates, relating to the overall material/knowledge provided by the Data Analysis course. Thak you so much, for this awesome course. I'm going to miss you all.

6

+ Add New Comment

Larry Cahoon · 6 days ago

I'll add my report to the list. It got an 81 overall with hits on the issue of confounders (3.5), people still don't understand these is my take, prediction accuracy (3), and references (3), problems with conclusions(4), and transformations(4). The references issue may be because I got caught in the rubric as I had the references but failed this time around to link them in the report itself.

https://coursera-uploads.s3.amazonaws.com/user-0d6e5642025e88c3ea2466b1/294/asst-5/294-513a54dfccda27.27074234.pdf

https://coursera-uploads.s3.amazonaws.com/user-0d6e5642025e88c3ea2466b1/294/asst-5/294-513a55064fd3a7.11302303.pdf

0

Anne Paulson COMMUNITY TA · 6 days ago

I too got dinged on transformations (3), confounders (2), prediction accuracy (3, I should have written this section differently) and missing data (3).

Nice job, Larry. You actually analyzed the data and made conclusions.

-1

+ Add New Comment

Rebecca Shustef · 6 days ago

Hi Anne, a very interesting analysis, very clearly written. Could you please post your code? I believe it would be very helpful for people to learn, even if you do not consider it pretty. Thank you for sharing your analysis.

0

+ Add New Comment

Alexander Kucherenko · 6 days ago

I'm posting my small analysis here (if Anne doesn't mind). My first version of report was enormous but having read some topics on the forum I've decided to keep it simple.

0

+ Add New Comment

Anonymous · 6 days ago

Here is mine. I was lucky enough to get generous reviewers and scored 90. I'd love any feedback/comments.

Analysis https://docs.google.com/file/d/0B0xebhfwFDptT3R3U2h3eTAyNVk/edit?usp=sharing

Figure https://docs.google.com/file/d/0B0xebhfwFDptOUl4RTFucjJzMFU/edit?usp=sharing

Figure caption https://docs.google.com/file/d/0B0xebhfwFDptckppR3dWQTJwZ2M/edit?usp=sharing

Code https://docs.google.com/file/d/0B0xebhfwFDptY0FCelc0YzBITEk/edit?usp=sharing

I started with the raw data so as to get the full set of 30 subjects. Here is the version I processed to make it the same format as SamsungData.rda https://docs.google.com/file/d/0B0xebhfwFDptNWdqd2pRT1Y5VEU/edit?usp=sharing WARNING: large file. You only need this if you want to run my code without doing any modifications.

2

+ Add New Comment

⬇ scroll down for more ⬇

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