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

Forums / Data Analysis Assignment 2

My 86 (out of 90) Points Analysis

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

No tags yet.+ Add Tag

Mykola Dolgalov · 6 days ago

Hi peer courserians!

I'd like to share my second data analysis assignment, which was **rated 86 points** by my reviewers. This is my first attempt to do such a work.

Main writing: samsungDataAnalysisFinal.pdf

Figure caption: **Figure 1. Importance of Variables for Random Forest Model.** The plot demonstrates that most of the variables have little influence and only first 100 variables or so have important influence on the outcome variable. The plot helps to decide at which level to cut off using the Gini Coefficient. Red horizontal line represents filtering the variables whose Gini Coefficient is less than 5 that gives us a list of 111 most important variables that play the major role in growing the forest. This is confirmed by testing error coefficients and cross-validation as described in Writing in more details. Please see Appendix A of the main Writing for the list of the selected 111 variables.

Figure 1: MedianDecreaseGiniFinal.pdf

My first analysis graded 79 points, and this is my pleasure that my efforts scored this high because these are my first data analysis works I've ever done.

I will appreciate your comments, suggestions. I will be glad to find areas for improvements, in particular, I will appreciate suggestions regarding application of models and suggestions what other algorithms could have been used in this prediction analysis besides Random Forests.

Have a great day!

Mykola

8

Mykola Dolgalov · 6 days ago

Other people's work that they've shared:

87.5 points: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=3659

86 points: https://class.coursera.org/dataanalysis-001/forum/thread?thread_id=3665

⌃ 1 ⌄

+ Add New Comment

Monika Jakubczak · 6 days ago 🔗

Great! Thanks for sharing!

⌃ 1 ⌄

Mykola Dolgalov · 6 days ago 🔗

Thanks for your feedback.

⌃ 0 ⌄

+ Add New Comment

Mike Neville · 6 days ago 🔗

Mine. 80.5 points. https://coursera-uploads.s3.amazonaws.com/user-51b62984fde7101bb92f3320/294/asst-5/294-513d3504eac647.33910620.pdf

⌃ 0 ⌄

+ Add New Comment

Thia, Kai Xin · 5 days ago 🔗

Great report! 2 quick qn:

1) How did you do 10-fold CV on randomforest? Did you self code your own for loop or is there a R library for it?

2) How you decided on 111 variables? Why not 110? Or 112? (I know you used 10 fold CV but based on 10 fold CV, how did you decide the cut off pt aka red line on your graph)

⌃ 0 ⌄

Anonymous · 5 days ago

You can used the rfcv function to do cross validation with the number of folds you want. http://cran.r-project.org/web/packages/randomForest/randomForest.pdf

There may be some other packages as well.

⌃ 0 ⌄

Mykola Dolgalov · 5 days ago

Thia, thanks for your feedback, I appreciate it! :-)

1) I wrote a universal function to which you pass a parameter K - how many folds you want, it creates a list with K members for results. I must say that 10-fold was a bit too much for my data set because the validation set had been a bit too small.

2) I decided based on inspecting the graph. It was a bit tough because the descent of the curve is pretty evenly becoming flat, and it is arbitrary to choose the cut off point. I judged that it was pretty practical to get Gini Coefficient greater than 5.

⌃ 0 ⌄

Mykola Dolgalov · 5 days ago

Anonymous, thanks for the information. R is so developed that you bet there must be a ready function not to do it yourself and spend a sleepless night before deadline as I did to finish the analysis and produce a clean report. :-)

⌃ 0 ⌄

Anonymous · 5 days ago

A guide I read elsewhere was to use the absolute value of the negative values in prediction accuracy as the cutoff point. Since there were no negative values, it suggested to use all variables. Doesn't that appear to be correct in this case since having 111 variables appeared to perform not quite as well? It's close, but still a little bit worse?

⌃ 0 ⌄

Mykola Dolgalov · 5 days ago

You are right, having all the variables gives us better precision, but it takes considerably more time to calculate. I did not measure exactly, but it was about 5-10 times faster. So I wrote in my report that I considered a compromise between speed, accuracy and flexibility of the model and thus chose the stripped 111-variable variant.

⌃ 0 ⌄

Anonymous · 4 days ago 🔗

I'm not sure I understand the random forest concept fully and the choice to run the rf twice. Can you explain why the original random forest of all variables would differ from the random forest of the Top 111 variables created from the original random forest, assuming same seed was set? In other words, if rf through its algorithm finds the most accurate from the full set and finds the most accurate from the Top 111 set, why isn't the "best" of the full set the same as the "best" of the Top 111 set?

⌃ 0 ⌄

Mykola Dolgalov · 4 days ago 🔗

Hi! The Random Forest built on all variables (let's call it RFa) will be different from RF built on just 111 variables (let's call it RF1) because RFa takes into account more useful variables. RFa will be more precise than RF1 because it is based on more useful data (more variables). RF1 is a simplified variant, I deliberately limited the number of variables to make the model run faster. In fact that is the only reason to limit the number of variables to work on when you use Random Forest.

Please note that both models are built on the data independantly from each other. My steps were: 1) build and train a RF model on all the training data (RFa) and check which variables it considers the most valuable/important for accurate prediction (using GINI index), 2) build a plot on that data to see how variable importance is distributed (if there are a small group of very important variables and a majority of marginally important variables that could potentially be excluded from the model to make it faster). 3) after looking at the plot I chose an arbitrary point of GINI index = 5. This gave me those 111 most important variables. It was challenging to choose this point because the descent on the curve is pretty even. If I included 200 variables, my resulting model would be more precise because the lower importance variables still contributed to the accuracy of the forest. 4) I constructed a list of variables and trained a new Random Forest model with only those 111 variables included (RF1). 5) I broke down the training set into 5 (and then into 10) parts and repeatedly performed training and testing two types of the models - with all the variables (RFa) and with only 111 variables (RF1). My table in the analysis shows this comparison for 5-fold cross validation. I saw that the RFa surprisingly had worse average accuracy than the model RF1, but RFa was more stable (had lower standard deviation of error). But on the final test set the full model had better accuracy. But as long as I had chosen RF1 to be my final model, I reported its confusion matrix, and I was happy that it generally works much faster than the full model.

So to answer your question briefly, the full model producess different results from the fimplified 111-vars model because the full model takes into account more variables, so it is a bit "wiser" and more stable. :-) That is my current undertanding of the matter.

Do not hesitate to ask. :-)

⌃ 0 ⌄

+ Add New Comment

Stephanie Yeah · 5 days ago

Can someone smart comment on what I did? I think I messed up because I trained on all the train data instead of creating an outside validation set. I got 76.

Main Text: https://coursera-uploads.s3.amazonaws.com/user-3d90c52b52b31d39ce3f819a/294/asst-5/294-513d855d226944.12457574.pdf

Picture: https://coursera-uploads.s3.amazonaws.com/user-3d90c52b52b31d39ce3f819a/294/asst-5/294-513d856342eaf1.29568455.png

Caption: **Figure 1 (Left)** An example plot of 2 variables colored by the 6 activities performed illustrating why the dataset cannot be partitioned using test conditions involving single attributes since the data does not exhibit a clear linear pattern. No line exists that could accurately segment data by each activity. **Figure 1 (Middle)** A graph plotting the size of a tree and the resulting number of misclassifications (represented by the X-val Relative Error). Creating multiple trees from the training data through cross-validation proved that the optimal number of branches was 6. **Figure 1 (Right)** A decision-tree model with 5 variables and 6 branches that had a reasonably high prediction accuracy of 87.5%. If the condition specified at a node is satisfied, then the branch to the left is taken. The height of the vertical lines is proportional to the reduction in deviance. The number of observations is the total classified at each terminal node representing each of the 6 activities.

⌃ 0 ⌄

Mykola Dolgalov · 5 days ago

Your report looks great, good job! I will read carefully and get back to you a bit later.

⌃ 0 ⌄

Stephanie Yeah · 5 days ago

Thank you so much!

⌃ 0 ⌄

Robert Blanford · 5 days ago

While you're waiting for someone smart to comment, I'll pitch in!

If I were Jeff looking at this paper to judge the quality of my class, I'd be breaking out the champagne! The lack of a validation set was a mistake and I think you can make an argument for assigning a couple more participants to the test set.

Overall, you show an excellent appreciation of the models you are using and an impressive grasp of the the raw data. The only let-down for me is the conclusion section. Did you just run out of words? The report was like a firework display that builds and builds and instead of a finale it just stops.

I didn't understand what you meant by intensive cross-validation leading to over-fitting. Cross-validation is one of our weapons against over-fitting. Apologies if I'm being daft.

Although it isn't in the mark scheme, I disagree with your conclusion. I don't think the experiment does show that "daily life activities can be classified even when they are not monitored in a restricted experimental environment." The data were collected under lab conditions and extrapolating to a natural setting is dangerous because we don't all keep our phones strapped to our waists, we have lots of weird things in our environment like escalators, we weave in and out of crowds and are confronted with all sorts of situations which will give different readings. You've demonstrated that if you artificially control any number of extraneous variables by using a lab environment, you can distinguish different activities with a good degree of accuracy. It's an excellent proof of concept. The next step is to see if you can statistically control for those extraneous variables when they present in the real world.

Purely stylistic, but I'd have liked to see all of your references at the end as well as in footnotes. I use Zotero to capture, insert and reference my sources. It allows you to switch styles with a couple of clicks which can be handy.

I really enjoyed your report, great language, clear figures and good understanding. Without trying to apply the rubric, 76 feels on the low side of fair.

˄ 2 ˅

Robert Blanford · 5 days ago

Haha, sorry Mykola, you commented while I was writing my response!

˄ 0 ˅

Stephanie Yeah · 5 days ago

Wow, thank you for the Zotero rec, it is so amazing. More people should know about it, Watch the movie https://www.zotero.org/support/quick_start_guide

1. At the end, I ran out of time and also had big problems with my figure (the first lattice graph should have had a linear line like this / but I couldn't understand how to draw that in lattice, then, the other 2 figures couldn't coexist with the lattice plot so I had to give up and paste them together in photoshop)
2. Thanks for the thoughts on the conclusion, that makes sense. You're right, mine is not right
3. I didn't have a good conclusion because I knew (1) my models were not great to compare so I really only had 1 tree as an option (I couldn't figure out how to make an SVM work), (2) I ran out of time to calculate statistical significance to and confidence intervals compare models so couldn't really conclude anything (see pp 188-193 http://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&

cd=1&ved=0CDUQFjAA&url=http%3A%2F%2Fwww-users.cs.umn.edu%2F~kumar%2Fdmbook%2Fch4.pdf&ei=S6ZIUdnRAqXNiwK_1IGwCQ&usg=AFQjCNEoT6hO3thIrf1cX-zaxSP20yZMhA&sig2=Ehp8wCzP46JLZWY6ompKaQ&bvm=bv.44011176,d.cGE), (3) also gave up cuz I thought I did it wrong by cross-validating on training

4. Can't find where I got the intensive cross-validation leads to overfitting right now, but doesn't running a model on 20 datapoints 100 times to pick the "best" one have the same worth as creating a model tailored to the validation data? Maybe not I'm confused...

^ 0 ⌄

Mykola Dolgalov · 5 days ago

Stephanie, my understanding is that k-fold cross-validation cannot overfit the model because you actually retrain the model with chosen parameters (e.g. Random Forest with *n* trees and *m* variables) *k* times, and every time you get a newly trained model, and you apply it to a completely new dataset for this newly trained model. So during cross-validation you pick meta-parameters, but you instantiate the model (actually create and train a new one) *k* times, and you apply it to the part that you excluded from training to serve as a test set. Then you pick the best variant and retrain it on the entire training set. After that you apply to the test set.

I am still reading your report. It is very intense, interesting, you demonstrated deep understanding of the subject and many aspects of the models. You definitely deserve a higher score, 10 points higher than 76! I would try apealing to TAs with a request for Jeff to have a look at your work, or to ask to re-evaluate your work by TAs.

^ 0 ⌄

Stephanie Yeah · 5 days ago

Do you know how I can email/appeal to the TAs? I was looking how to do that but can't figure out a good way

Thanks for the compliment and your review

What is meta-parameters? How can you pick something, but create something else not including it? "pick meta-parameters, but you instantiate the model (actually create and train a new one) k times" Sorry, your paper is too complex for me, I don't understand randomforests that well, or how CV works with them

^ 0 ⌄

Mykola Dolgalov · 5 days ago

Stephanie, metaparameters is actually a fancy word for quite simple thing - parameters that you can choose for your models, like number of variables to include, or the number of trees to build in Rendom Forest. The goal is to choose the parameters that influence your models to gain the best accuracy and stability. For example, I tried 2 variants of Random Forest models - one with all 561 variables and one with 111 variables. Both times I left the number of trees to grow to be default 500 tress. Now I know (from other reports) that it would be better to grow 1000 or even more trees, and that it should be 1001 trees to break voting ties. So the number of trees is another parameter to play with when performing model selection and cross-validation. So the idea is to train different models with different parameters to see which model with which combination of settings produces the most stable results on validation data sets.

Random Forests are very powerful and are realy great, but the ideas behind them could be explained on higher level, and you do it in your paper. For me this is the first time I've encountered Random Forest. At the beginning of the year I knew nothing about R, and very little details about data analysis. These two courses by Prof. Peng and Prof. Leek were a great advance for me.

I must admit that the closure of your report is somewhat confusing for me. Your Table 5 shows clearly that Random Forest model performed much better on the test set, so the logical way would be to use RF model rather than Trees. Random Forests mitigate overfitting by voting as you describe in your paper, and my cross-validation confirms that.

I also made the same mistake as you did - I tested two models on the test set and I reported error rates for two models rather than on one final model. I realized this only after I saw the evaluation choices where reporting test results on one model scored 5 and reporting on several models scored 3. Now I realize that only one model and only one error rate must be reported on the final result set.

⌃ 1 ⌄

Stephanie Yeah · 5 days ago

That table is confusingly labeled, the randomforest and tree3 confusion matrices are applied to the training set (not test set), I put them next to the final test confusion matrix to compare the mistakes made.

I agree, Jeff's course is so good and I also have the same background as you (but you are a Data Warehousing Solution Architect so you are much more knowledgeable) and also graduated from undergrad in 2012. Your explanation of randomforest is really good, thanks I understand now. My paper is confusing because you can see from the citations, most comments are from other sources (not me, I just tried to compile them all and in the end did not resolve all the questions I had). The good stuff is just from research I did (not my own knowledge). Thank you for reviewing my paper in your thread.

⌃ 0 ⌄

Stephanie Yeah · 5 days ago

I don't think you can use a randomforest model without looking at confidence intervals and whether the variables that performed best were significantly better than each other. The source I linked above had some equations to calculate this (like mean +/-

sd * (t value for k=#) or something, I tried to decode that from an equation in the paper)

⌃ 0 ⌄

+ Add New Comment

Mykola Dolgalov · 5 days ago

Regarding TAs, you can search replies on the forums marked with Community TA yellow mark, open their profiles and search for their LinkedIn contacts. Some of them have other ways to contact. E.g.:

https://www.coursera.org/user/i/bc19796ba4f528c7b929ada0122715b7

Or write to them directly in some threads, where they are active and ask how to contact them re Assignment 2.

⌃ 0 ⌄

+ Add New Comment

New post

Bold	*Italic*	Bullets	Numbers	Link	Image	Math		<HTML>

□ Make this post anonymous to other students

☑ Subscribe to this thread at the same time

Add post