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

Forums / Data Analysis Assignment 2

Yet another report - 89

Subscribe for email updates.

Sort replies by: Oldest first Newest first Most popular

No tags yet.+ Add Tag

Anonymous · 5 days ago

My analysis wasn't terribly sophisticated, just a single, lonely tree. However, I tried to say as much about it as I could.

My overall strategy was to keep things simple so I could understand what was happening so I'd have plenty to say, reviewers wouldn't be blinded by stats and to allow the reviewers to easily award marks. Therefore, people looking for tips should consider this as an example of tailoring content to the markers and rubric, rather than a good and exciting analysis.

I did post-processing on the figure in Inkscape. I changed the labels on the graph to avoid any chance of a marker confusing legitimate, but automatically generated labels with R variable names. I also put boarders around the decision criteria on the tree and moved them to avoid overlaps.

I drew the in-line figure (Figure 2) in Inkscape, in case anyone is interested. I tried to do it in Dia, but it was both difficult and looked awful. I might have tried it in LibreOffice Draw, but I was running short of time by then and knew I'd be able to do a reasonable job in Inkscape so stopped faffing around!

I don't use dropbox much, so I hope I've done the links right!

Main text Figure Figure text

Comments welcome.

^ 3 ˅

Anne Paulson COMMUNITY TA · 5 days ago

Of all the analyses I've read, yours did the best at explaining what the various variables meant, and why certain variables separated the activities. Nice job.

^ 1 ˅

Anonymous · 4 days ago

Thanks Anne. I read yours - nicer job ;-)

I liked how you got some good error rates and thought "right, why are they wrong then?"

I look forward to giving some of those other methods a go.

⌃ 0 ⌄

+ Add New Comment

Henry George Bottomley · 5 days ago

Very nice - especially about subject 25 "laying", and then possible interpretations of the tree.

The one thing I might have been slightly unhappy about was the lack of a mention that the variables had been pre-processed to be in the range [-1,1].

⌃ 0 ⌄

Anonymous · 4 days ago

Hi Henry. Cheers for the pointer. I was totally blinkered on the pre-processing issue, it didn't cross my mind.

In case you're interested, I spotted subject 25 by mooching through scatter plots of all the variables, I think it was David's suggestion. Here's my code which borrowed heavily from Jeff's example when clustering:

```
for (i in 1:561) {

png(paste("./plotdump/scatter",i,".png", sep = ""), width = 800, height = 48
0)
plot(actsamsungData[,i], ylab = colnames(actsamsungData[i]), main = colnames
(actsamsungData[i]), col = actsamsungData$actnumber)
legend("bottomright",legend=unique(actsamsungData$activity),col=unique(actsa
msungData$actnumber),pch=19)
dev.off()

}
```

Anne found more awkward subjects in her analysis. They didn't jump out at me as wrong, so either I missed them or they were legitimate, but odd ways of carrying out an activity (or to be technical "silly walks").

⌃ 1 ⌄

Anonymous · 4 days ago

I just realised that won't work! I added activity as a number so I could use it for colour, did some re-ordering to group by subject and saved it as a new df before plotting. I could dig out the rmd if it is of interest.

▲ 0 ▼

Anne Paulson COMMUNITY TA · 4 days ago

I didn't spend any time trying to figure out if there were outliers that should be removed from the data, because I figured that as far as we know, all these numbers are actually the results of actual people wearing actual phones, so in the field, we would see data equally anomalous. If someone is wearing their phone upside down for part of the test-- well, people are going to do that in the real world, and our predictor better be able to handle it. So although my data found that some subjects were "difficult," I think that dealing with difficult subjects is our problem and we can't justify throwing them out to make the problem easier.

▲ 0 ▼

Anonymous · 4 days ago

I thought the opposite :D

This was a lab experiment where they've controlled for things like how and where people wear their phones in the experimental design. They've compromised ecological validity for a stronger signal. By choosing this design the team have to admit that you cannot apply the results of the experiment outside of the lab environment because the real world is a nasty, complicated place full of extraneous variables. However, it means that they can say with confidence that if you do these activities, on this apparatus with your phone here, we can tell what you're doing.

It makes no sense to allow a degradation of your experiment by allowing data to be used which has been collected outside your defined parameters. You can say, "that's how they roll in the real world", but we're not in the real world and you will never be able to make any claims about the real world from this experiment. All you are doing is complicating life and throwing noise at your signal.

▲ 0 ▼

Anne Paulson COMMUNITY TA · 4 days ago

Since this was a lab experiment with controlled conditions, how can we justify throwing out data collected in that lab experiment with controlled conditions? Our test data is more data from that same experiment. There might be people with upside-down phones in the test data set, and we'd like to be able to predict for them.

An upside-down phone is not noise; it's signal.

⌃ 0 ⌄

Anonymous · 4 days ago

I'd argue that you have no choice BUT to throw them out. Admitting data that breaches your experimental controls is going to cause you all sorts of trouble. It's like shrugging your shoulders after sneezing into a petri-dish in a biology experiment.

The data-munging takes place before the splits into training and test sets so if you are able to detect upside-down activities, you should be able to exclude them all.

I think of a lab experiment as creating a perfect world where you are giving the attribute you're studying the best possible chance to shine out. In the world they've / we've constructed to look for differences in how people move we are holding the phone position and orientation constant. We've created a world where everyone has their phone in the same place, the same way up. Upside-down man is an alien who resembles us in most ways, but on occasion suddenly starts walking while upside-down. If you transform his upside-down data and all the derivatives, that's fine, it is signal because he's not upside-down man any more. If you don't do this then you are building a model to predict the activity of people who will spontaneously break the laws of physics.

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

The data-munging takes place before the splits into training and test sets

No. No. No, it doesn't. That's the way to heartache.

I do agree that if you discover, in your training set, a way to identify and remove anomalous data, then you are justified in doing the same for your test set when you get to it. But you absolutely, positively must not do munging on the test data.

In Learning from Data, Prof. Abu-Mostafa describes a case he was involved in where violating that rule meant disaster. This was a financial application: they were trying to figure out how to arbitrage the British pound and the American dollar, I think, or something like that. Before they divided their data into training and test, they mean-centered and standardized it. Then they separated into training and test, and locked away the test data.

They developed a model to predict how to arbitrage. It looked good. They tested it on the test data. It still looked good. They deployed it in the field; it lost money. Oops. It turned out that in mean-centering all the data, instead of just the training data, they had biased their results.

You need to keep your hands off test data.

⌃ 0 ⌄

Anonymous · 4 days ago

You're so right. That kind of selective cleansing is out of order in the test set. There's a lesson I'm going to remember :-)

I still think that you shouldn't include his dodgy data in the training set. They've messed up the experiment by having the phone that way up, if you are going to control for orientation you need to control for orientation! It shouldn't be possible to have upside-down readings in the training set or the test set.

I think it's legitimate to say that subject 25 isn't just lying down or her phone is malfunctioning or ringing during the experiment and upside-down man is breaching experimental controls and the data should be withdrawn. We should be looking at how well we can predict activity under experimental conditions and bad data collected due to our ineptitude as experimenters should be excluded rather than weakening / generalising our model which we know is going to be useless outside of the lab.

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

It depends on whether you think upside-down man counts as a breach of experimental controls. If you do, and if you have a way to find upside-down people and eliminate them from the data, that's defensible.

But you might also take a different point of view. I assumed that the people in the experiment were actually wearing the phones (upside down or not), and doing the activities they were described as doing. Therefore, in my view, they were contributing valid, if messy, data and I should train on it. I would only eliminate someone from the data if I had reason to believe that the activity was wrong (the activity said they were walking, but they were actually lying down) or they weren't wearing the phone at all (they were walking, but their phone was over there on the chair). I wouldn't throw out data from a "malfunctioning" phone if phones often "malfunction" that way, because I want to build a predictor that predicts from the data we actually get from actual phones. I take a more expansive view of what is clean data.

⌃ 0 ⌄

David Hood · 4 days ago

The original experimenter's have already standardised (-1,1) our test set (and I think this was a general change to the entire data including their set). We are doing what we can with the data.

Knowing what I know now about the subjects, if I was building a basic model I would probably see if it was better (on the training set) to leave out particular subjects. The basic idea is that this person is such an outlier that they are making the model worse for everyone else. And if you leave them out is the cost from encountering people like them made up for by the improvement in accuracy for everyone else (so it better overall results). As an example subject 16 seems to be both an outlier and have a disproportionate effect on the way the model for sitting/standing is constructed.

You figure it out by cross validation, and try different models, and estimating how common (how much of an outlier) people like the subject are.

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

For our particular data, that's a grim task, though: You'd have to do 17 x 16 cross validations and build (17 x 16)/2 models. That is, you'd have to build a model for every set of fifteen subjects, then test it on the other two subjects, to see if leaving out subject A worked better for predicting subject B, for every combination of A and B.

At some point, we're becoming too familiar with our data. For me, building 136 models for every choice of model/parameters is that point.

⌃ 0 ⌄

Anonymous · 4 days ago

For a lab experiment I would take the narrow view. The pooch has already been screwed from an ecological validity perspective. If you're going to be artificial environment you might as well make the most of it and be clinical in your execution to get the most clean results possible. If you can't manage that, then do your best to clear up your mess!

On the other hand, in a natural setting, I'm 100% on your side. I'd expect to have to calibrate for phones at different angles, worn in different places or carried in a bag, ringing, not ringing, malfunctioning (if common or useful) - there is no such thing as "wrong" any more, there just is what there is and you need to model what you've captured however inconvenient your participants have been.

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

Here's the difference: you are thinking this is a lab experiment. I am thinking I want to build the best predictor I can. I don't care about lab experiments. I don't care about the physics except as it informs me to make a better model. I want to build the best predictor of activity for people wearing phones on their waists.

The only measure of success I care about is: does my predictor work for people wearing phones on their waists. I want to build the best predictor I can, and I want to know as much as I can about how good my predictor is. The only reason I'd have for throwing out data would be if I came to believe it would make my model worse, and I'm deeply skeptical that with only 17 subjects, I can make that judgement.

I'm not saying you're wrong. I'm saying you and I are solving different problems.

⌃ 0 ⌄

David Hood · 4 days ago

Anne, it is a grim task, but it is a thoroughly susceptible to automation.

To me the bigger issue, in this area, is that different subjects have unequal weights in the construction of the predictive model (depending on the kind of model) since they have different numbers of observations per activity. This can make some kinds of models very vulnerable to overfitting on an outlier subject (in terms of behaviours) that did a lot of a particular action. If that subject is enough of an outlier to move the model with their observations, we are arguably removing overfitting by removing that subject.

Of course, the difficulty is in the step "Is this having enough of an effect to justify removal".

⌃ 1 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

Of course, the difficulty is in the step "Is this having enough of an effect to justify removal".

Yes, exactly. That's what I meant by becoming too familiar with our training data. With such a small amount of training data (only 17 subjects) we are vulnerable to overfitting, and throwing out subjects because you suspect they are worsening the model to me looks like overfitting. You use validation to stop you from overfitting, but if you make too many choices with validation, that safeguard disappears. And our test data (four measly subjects) isn't helping us much. In fact, we're probably better off putting that test data into the training data. It's useless for testing anyway; we'd be better off training with it. Then at least we'd get some use.

⌃ 0 ⌄

Anonymous · 4 days ago

It's interesting how different angles of attack lead to very different choices.

Also, a flawed underlying assumption of my work is that there are no real differences between subjects. That gave me the liberty to remove only the dodgy activity data for a participant rather than all their data.

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

Also, a flawed underlying assumption of my work is that there are no real differences between subjects.

That was what informed my entire analysis: the realization that rather than having 5600 independent observations in our training set, we have observations from only 17 subjects. To me, that changes the task and the way to think about it in a big way.

⌃ 0 ⌄

Anonymous · 4 days ago

One of many reasons your analysis is better than mine :-)

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

I thought your analysis was very good. And while I got 81, you got 89, so the evaluators also thought your analysis was very good.

⌃ 0 ⌄

David Hood · 4 days ago

To some extent, I'm dealing with related issues right now as I put together the subject model. In the sitting/standing activities there is that clear evidence that the decision tree construction is being heavily influenced by 16 and 5 (and neither are representative of the major of subjects). It is just rather than excluding 16 and 5 from the model construction I have gone the route of overwork and have a decision pathway for the majority of subjects (12 of training set) and a pathway for the minority (5). It just happens that there is a really reliable decision tree that can be built about "should an observation take path A or B".

Now, it could be seen that I am optimising a set of decisions for a readily identifiable group of observations. But while working on it I was looking at it from the converse perspective of "removing the overfitting effect caused by basing a single decision tree of non-representative subjects". Now I tend to see both of those issues as the same problem, arrived at from different directions (at least with this data).

⌃ 0 ⌄

Anne Paulson COMMUNITY TA · 4 days ago

Some of the stuff you're doing has the whiff of overfitting to me, though.

⌃ 0 ⌄

David Hood · 4 days ago

It might, we will all know in a day or two (I hope) :)

I tend to imagine a lot of what I am doing is identifying structural (nature of data to the way the models are built/ interrelations in the data form useful patterns) sources of over fitting or underfitting and structurally compensating, but I'll also happily acknowledge that if I my structural interpretations wrong, it should go spectacularly wrong on the test data :)

⌃ 0 ⌄

+ Add New Comment

Priyanka Deshmukh · 5 days ago 🔗

Wow! that is really awesome. I think u r right Anne it was the best explanation of all for me also. congo!!!

⌃ 0 ⌄

+ Add New Comment

David Hood · 5 days ago 🔗

Very nicely done. Nothing fancy, but well researched and explained. I would say the marks were well deserved.

⌃ 0 ⌄

Anonymous · 4 days ago 🔗

I leave the fancy stuff to you! I've been thoroughly enjoying your exploration of subject. Thank you for taking the time. I find that I ask the same questions as you, but then you have the skill to answer them! I'm learning a lot from the thread and it'll be one I save to look at later before the bulldozers come in.

Thanks for your generous comments.

⌃ 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