Re: Project Progress

Subject: Re: Project Progress

From: Priyank Jain <jain206@purdue.edu>

Date: 10/15/2016 01:58 AM

To: Philip Michael Van Every <pvanever@purdue.edu>

Hey Phil,

Thanks for updating me, I am a bit lagging on this. See my comments below. I would spend more than half day tomorrow on this and we can talk Sunday afternoon if that's fine with you? I should be able to address your queries too after I have spend time on this tomorrow. Thanks for the great work so far, talk soon.

Regards,

Priyank Jain

Developer Team, Purdue IronHack

Research Center for Open Digital Innovation (RCODI)

Purdue University, Discovery Park

West-Lafayette, 47907-IN Phone: +1-765-476-3000 Email: jain206@purdue.edu Skype: live:priyank.purdue

About me: www.purdue.edu/opendigital/about/students

On 10/14/2016 05:11 PM, Philip Michael Van Every wrote:

Hello Priyank,

I just wanted to touch base on project progress.

I have been going through the Python tutorial on lynda.com and I am checking out how to use numpy arrays (which is what we'll probably use for all of our data right?) Yup, numpy would be best, most python-stats libraries would use it.

I have looked into normalization also. I should have classes that do both the Gaussian and 0-1 normalization implemented by tomorrow afternoon pretty easily. I am also looking into scipy stuff. They have their own normalization modules, so I may use some of their stuff in other classes. I am not sure if simply using existing libraries for normalization is ok though. In the project specs, Dr. Honorio says that we can use existing libraries for learning algorithms, but that we must do other stuff on our own- so I am not exactly sure where the line is drawn here. I would say use the existing libraries for now (while at the same time understand their internal working through documentation). If time permits, we could do the implementation ourselves too, would be a good learning point.

I also thought a little more about the lasso feature selection. I think we can do it just fine as

1 of 2 10/15/2016 06:41 PM

Re: Project Progress

long as we use a different loss function: namely- use y -sign(theta*phi) instead of just j - theta*phi. The problem we might run into, however, is calculating phi. I'm still not so sure I'm thinking about this the right way. I totally get how it works with regression but I'm not sure how to implement it with classification. I have read about other people using it for this, however, so I trust that it will work and that you know what you're doing. In either case, I do see how we can ignore the theta from this stage farther down the pipeleine if we just get the subset of features, which is what I believe you were trying to tell me the last time we spoke. I recently reviewed the information analysis approach introduced in the same lecture and I see how this can easily be implemented, so I may try to build a sub-module that does this. It can't hurt to try more than one thing here and report on the results right? Yup, doesn't hurt, infact looks good, since it's what true ML work in the project is, rather than implementation of code. I am not sure if I got your query right for the first part of this paragraph, but I'll read this again tomorrow (Saturday) once I have done some research.

I can't remember when we planned on talking next. Is it tomorrow or Sunday? I'm free either day and the afternoon would be better for me in either case. I'm also free all day Monday. Whatever is best for you. Let's talk Sunday, so I have sufficient time to catch up on the backlog and so that we both can discuss what we did.

Cheers, Phil

2 of 2 10/15/2016 06:41 PM