Subject: Re: Getting ready for Nov 1 submission
From: Philip Michael Van Every <pvanever@purdue.edu>
Date: 11/04/2016 09:38 AM
To: Priyank Jeevanlal Jain <jain206@purdue.edu>

Hi Priyank,

Any changes I had are in. Really, all I did since we spoke last was fix a little accuracy calculation error I had during the k-fold cross validation in the ParameterTuner class.

SVDD can be used directly through SVC by setting on of the configuration parameters. It can be done with quadprog as well. The second page of this paper (https://www.csie.ntu.edu.tw/~cjlin/papers/svdd.pdf) give the dual solution with upper and lower bound constraints etc... I never actually implemented this one with quadprog because when I implemented ocsvm with quadprog it was extremely slow (in matlab anyway...).

One of the things returned by svmtrain in libsvm in matlab is a vector of decision_values. I believe these are just the inner product of each data instance with the hyperplane. The prediction labels we get are just the sign of the decision_values. To generate an ROC curve, I would sweep across all decision_values, setting a threshold for each one and predicting samples with a smaller decision value to have a negative label and samples with a higher decision value to have a positive label. The means, basically, that we are ignoring the actual predicted_label output for now and just creating our own (really, the predicted_label output is the same as what we get when we set the threshold equal to zero). Sweeping across all decision values this way is equivalent to shifting the classifying hyperplane laterally across all data points. At each threshold, we calculate the True Positive Rate (The number of correctly classified negative data points divided by the number negative data points) and the False Positive Rate (the number positive data points to which we incorrectly assigned a negative label divided by number of positive data points). Plotting the FPR vs. TPR gives an ROC curve. Essentially this tells us how well the classifier separates data in the hyperspace.

I know that that is a very wordy paragraph. The wikipedia article on ROC curves is actually pretty helpful I think. It may not be worth the trouble of doing these. I thought it might be a good idea initially because I have used them before but I think just using the predicted_label given by your classifier and calculating accuracy is pretty solid.

I am not sure how to get decision values with SVC, but I bet there is a way.

Cheers,
Phil

A Revisit to Support Vector Data Description - csie.ntu.edu.tw

www.csie.ntu.edu.tw

Machine Learning manuscript No. (willbeinsertedbytheeditor) A Revisit to Support Vector

Data Description Wei-ChengChang Ching-PeiLee Chih-JenLin

From: Priyank Jeevanlal Jain
Sent: Thursday, November 3, 2016 8:43:41 PM
To: Philip Michael Van Every
Subject: RE: Getting ready for Nov 1 submission

Hi Phil,

I am so glad you wrote to me. Yes, things were a bit heavy so far. But should be easy here on J
Hope your work-related issue is resolved and you are killing it for the two courses.

I just wanted to understand the SVDD part of your past ML project so that I can apply it for the course project too. But I understand company policies might restrict sharing of that information, so any blogpost/link regarding SVDD application/ROC curve would be fine too. It would be awesome if you could commit your changes to the repository, there may be something I would include in the final project. Thanks for all your work, let's hope we can collaborate in the future!

Regards,

//
Priyank Jain
Graduate Student,
Masters in Computer Science,
Class of 2016
Purdue University
Phone: +1-765-476-3000
Email: jain206@purdue.edu
Skype: live:priyank.purdue
About me: priyankjain.net
\\

From: Philip Michael Van Every [mailto:pvanever@purdue.edu]
Sent: Thursday, November 3, 2016 4:24 PM
To: Priyank Jeevanlal Jain <jain206@purdue.edu>
Subject: Re: Getting ready for Nov 1 submission

Hello Priyank,

I just wanted to check in and make sure everything is ok. I didn't see you in Algorithms last night and I haven't heard from you for a little while. I know you are busy with some deadlines and applications,

so I hope everything is going well for you.

If you need anything in regards to the project or anything else, just let me know.

Cheers,
Phil

From: Priyank Jeevanlal Jain
Sent: Friday, October 28, 2016 5:58:51 PM
To: Philip Michael Van Every
Subject: Re: Getting ready for Nov 1 submission

Hi Phil,

Sorry to hear that you dropped the ML course. Hope the work-related issues are sorted out. I greatly appreciate your support over the weekend, much needed since I am pretty squeezed for time because of upcoming deadlines.

Yup, hold on to the changes, I'll make a commit tomorrow/today and let's talk after that.

Thanks again for your understanding!

Regards,
Priyank Jain
Graduate Student,
Masters in Computer Science,
Purdue University
+1-765-476-3000

From: Philip Michael Every
Sent: Friday, October 28, 5:37 PM
Subject: Re: Getting ready for Nov 1 submission
To: Priyank Jeevanlal Jain

Hi Priyank,

That sounds fine to me. I did make some changes in DataGen. If just reading a few lines makes things easier, that is fine.

I am glad you are doing the feature selection. I started a pipeline script using SVC feature selection, but I haven't gotten it totally going yet, so I will defer to you implementation of that...

I started making minor changes to stuff while I was building this quick experiment pipeline, but it sounds like you have stuff going and that you have modified it how you need to make it work, so I'll hold off on the stuff I am doing.

I have some unfortunate news. For some kind of complicated work-related reasons, I had to drop the machine learning course. However, I will not just disappear. I will definitely continue working with you on the project this weekend- don't worry. I will make sure we get some initial results and a decent report completed by the deadline. I am not just going to leave you hanging.

In light of this, I think it is best to let you take control of the pipeline- so again, I will hold off on what I was making because it sounds like you have a plan. We should let your changes to any modules override mine. I did push some changes in my folder this morning but we should disregard those if it serves what you are working on better.

When the feature selection stuff is complete (or before that), just let me know what you need me to do... The pipeline script I was building basically tested for accuracy over a range of feature subset sizes and sample sizes. I thought those would be good plots to show to start of with. If you have something else in mind that is fine too.

Cheers,

Phil

From: Priyank Jeevanlal Jain
Sent: Friday, October 28, 2016 1:53:45 PM
To: Philip Michael Van Every
Subject: Getting ready for Nov 1 submission

Hi Phil,

I was making changes to your DataGen.py file in the code folder, also you made some changes to that file in phil folder. To avoid any merge issues, can you please make sure that your final working code is always there in the code folder? The phil and priyank folder are simply playgrounds for our development. Also, can you make sure you editor is using 4 spaces for tabs? It will result in indentation error otherwise. I always use a tab for python identation, saves me lot of syntax errors, moreover it's the standard.

I just edited the readData function in DataGen.py since we don't want to read the complete data. I modified to simply read a few lines from the file and then closing it, this is much more efficient I guess and won't create trouble with my SSD. Does that work for you? I am creating Univariate feature selection module and then apply different feature selection techniques to the SVC with and have few plots ready for the submission.

Also, let's stick to analyzing just 5000 or 10000 examples, what do you say? This would be much faster.

I'll talk to you soon after the crux of the work has been done.

Cheers,
--

//
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
\\