You got a comment on your Github project

EPFL Extension School <no-reply@extensionschool.ch>

Ven 27.09.2019 10:46

À : dujaj@hotmail.com <dujaj@hotmail.com>

Dear John,

Michael Notter just commented on your Github project.

Michael Notter (commented 10 minutes ago)

Hello John,

Thank you for your project solution. You completed all the tasks and did excellent work throughout. Well done!

The following is a list of thoughts and comments I had while reviewing your project:

- Your "most frequent" baseline shows you an accuracy of 22.8%. This is because of the randomness in the train/test split. If you use the `stratify=y` parameter during the splitting (i.e. `train_test_split(Xp, y, test_size=0.3, random_state=0, stratify=y)`) you can make sure that classes are balanced and the most frequent baseline is again at 25%. (I just saw that you use `stratify` in later notebooks, so never mind :-))
- In the KNN example, your grid search looks for PCA components between 210 and 335. The result of the grid search reveals that 210 is the best number of components. As this was also the lower bound of your grid search, it is recommended to increase the grid search range in this direction. Otherwise you cannot be sure if 210 was an optimal value or if 190 or 150 would have been even better. (The same point is true for the SVM with RFB kernel grid search.)
- Your KNN notebook is very well done. Very well structured, written and commented.

Bravo!

- As you observed, your random forest is overfitting. This is due to the `max_depth` parameter, which by default is `None` or full depth. This means it will use all the features to do the fitting and can therefore perfectly learn the training dataset. For this reason it is recommended to do a grid search on `max_depth` and `n_estimators` when working with random forest classifiers.
- Concerning your "computation speed up" question in the SVM notebook. Your machine seems to be already very strong. So, the speed up would need to come from reducing the number of computations per model fit. Which in this case would be a reduction of the features (which you already do with PCA). What you could do in the future is reducing the number of PCA component explorations. A good rule of thumb is to increase the values by one magnitude (e.g. 10, 100 and 1000). And as a second run you can then restrict the exploration a bit more.
- The SVM grid search take in particular longer, as we're also using a cross-validation, in this case of 5. In other words, we're performing the grid search 5 times, i.e. 24 * 5 = 120 times. So in the end, your 9min for the SVM-RBF grid search is 4.5 seconds per grid point. Which seems to be reasonable.
- To your second part of the "computation speed up" question: Your code looks correct and optimal. You did it the way we expected it and how it should be done.
- Congratulations on the convolutional neural network! The implementation is very well done and you chose a great architecture (size of convolution, number of layers, with a dropout to prevent overfitting,…) BRAVO! And the convolution kernels at the end look really beautiful. It's clear that the network learned something meaningful. Looking at the train and validation curve, you could have perhaps even run it for a few more epochs before overfitting would have taken place. It's nice to see that both of those two curves are evolving almost in parallel. Again, very well done!

Your predictions on the withheld test set were very good. You've reached a prediction accuracy of 84.3%, BRAVO!

```
              precision    recall  f1-score   support

       truck      0.830     0.876     0.852       250
         car      0.847     0.888     0.867       250
    airplane      0.891     0.752     0.816       250
        ship      0.814     0.856     0.834       250

   micro avg      0.843     0.843     0.843      1000
   macro avg      0.845     0.843     0.842      1000
weighted avg      0.845     0.843     0.842      1000
```

To your final comments in the last notebook:

- Your notebooks are very well written, well structured and well commented. Bravo! You solved all the tasks and performed them like we expected from our learners. Very well done! Your observations throughout are correct and you show a high skill set in data science.
- A good data scientist does not always get the "highest" accuracy predictions. The important part is that you are aware about the dataset, what does it represent. Do a good EDA to see particular characteristics and identify outliers and missing values. And then explore different models. And most importantly, be aware about their strengths and their weaknesses and comment, comment and comment your observations and decisions. All which you did very well!
- And don't worry about your progress speed. We have many learners who take the full

- And don't worry about your progress speed. We have many learners who take the full 18 months for our course. This is mostly due to family, job or thorough learning. Anybody can try an rush through the workshop content, but not everybody takes the time well enough to learn the things in more detail and to get a deeper understanding of what is learned.
- To your question about model performance. I always find it interesting to see that each learner identifies another of the different models in this project as the best one. They are all capable to perform rather well, if the parameters are set exactly right. So, the discrepancy is mostly due to the grid search. Having said so, what can also be done is combining the predictions of the different models in something like an ensemble classifier. For example, if you have 8 classifier, you could take the majority rule and predict the class that is predicted by the most classifiers. Such an approach is not guaranteed to get better results, but it can help to profit from the pros in all the classifiers and controlling the cons in them.
- I'm happy that you had joy going through this project. You clearly show a good data science mindset and skill set. I hope that you can get the opportunity to get more analytics task in your work.
- About the python group to learn with and meet, I unfortunately have no clear pointers.

You can now proceed with the capstone project. Don't hesitate to book a 1-1 video call with me if you have any questions about the course or want to discuss the project results in more details. You can do it directly via this link: https://calendly.com/s/fllFSLyh

Kind regards,
The EPFL Extension School team