Zeel B Patel
patel_zeel@iitgn.ac.in
patel-zeel.github.io

JAX ML Textbook

My proposal is related to the “Develop JAX examples and demos for an ML upcoming textbook” idea in the Idea list. I propose to work on the following problems during the GSOC.

1) Moving from scripts to notebooks for book1 and book2

The “Probabilistic Machine Learning: An Introduction” book links the code in the captions to reproduce the figures. Currently, the links redirect a reader to a particular cell in a chapter-wise unique colab notebook. Then, the reader can execute the cell to clone the pyporbml repo, automatically install the imports with superimport and see the output generated by the script. There are several issues with this approach:

1. It is difficult for the readers to edit the scripts and experiment with the code.
2. The whole repo is by default cloned (due to the automated process) before running any script.
3. There are no saved visuals from the last run of the script such as figures. So, without executing the script, readers can not see the output.
4. It is hard for the maintainers to manage the redirection from the book to a particular cell of a remote notebook that can have a random name.

Considering these challenges, I propose to solve them with the following procedure:

1. Convert all scripts to stand-alone Jupyter notebooks hosted on GitHub. readers can easily open them in Google colab and experiment with the code.
2. Readers can by default read the code and see the pre-executed results without connecting them to a live backend (such as colab).
3. Each figure caption redirects the users to the notebook hosted on GitHub so the maintenance process is simplified.

Additionally, I will add the automatic code style and correctness verification via the workflow. At the end of this objective, we will have pyprobml repository at a similar level to other software repositories in terms of maintenance, structure and code quality.

2) Latexify the figures

By default, the matplotlib library generates a fixed size figure which naturally does not suit for all the figures. So, a maintainer has to carefully set the figure size for each figure and then adjust it again in the latex. Without this, the visibility and quality of the figures can be degraded drastically. I have been using a set of tools to automate this process as introduced by my Professor over here. I propose to develop the process to latexify all the figures in the book1 and book2 which can happen at its own pace even after the GSOC period. Deliverable for this Idea would be a process that can be taken by almost anyone (via crowdsourcing) by following a simple set of instructions and several notebooks that I will latexify as examples.

3) Interactive demos for several figures in book1 and book2
Machine Learning is easier to understand when illustrated with visualizations. Interactive visualizations can add additional benefits by giving a virtual playground. I propose to add such demos to several notebooks. I understand that I will not be able to do this for all the figures but I propose to provide several examples on which others can build upon to take this work further at a certain pace.

To easily manage projects 1), 2) and 3) I have created dashboards on pyprobml repo that gets updated automatically based on the GitHub Actions workflow.

4) Develop new code in JAX for several figures/algorithms in book2
I am working on the application of Gaussian processes in my Ph.D., so I have a fair exposure to the basic concepts that form the basis of probabilistic machine learning. I have recently started using the JAX library and immediately became an admirer of it due to its flexibility, simple design and tools like Pytrees. Thus, I am motivated to implement new algorithms in JAX to modify several scripts in the current codebase. Deliverable for this idea would be a good quality code in JAX that makes it easy to understand how exactly the mathematical concepts can be implemented in code. I have started working on ADVI (Automatic Differentiation Variational Inference) as part of this project.

5) Contributing to the Gaussian Processes (GPs) chapter
I have been working on GPs since two years now so I can help in making the GP section better in the book. At places, I can solve the current errors, propose new sections, and if possible, contribute some content to several sections. I have already created several figure notebooks for this project on pyprobml repo.

Prospective mentor(s)
- Kevin Murphy
- Mahmoud Soliman
- Scott Linderman
- Nisa Ilhan

Why are you the right person to work on this project?
I am the right person to work on this project due to the following reasons:
- I have a year and a half of experience in writing and maintaining Python code for my own research projects in the GitHub ecosystem.
- I have personal expertise with several proposed ideas e.g. latexify
- I am personally interested in probabilistic machine learning in general and Gaussian processes in specific.
- I get satisfaction in making knowledge accessible with ease to a large number of people

What is your experience with open-source code, if any? Please include links to any open-source projects or contributions (if available).

- I have been regularly contributing to the open-source community for the last 6 months. Following is the selected list of PRs I have worked on in the past: https://gist.github.com/patel-zeel/e0d3547cebecf8cb19acaa97f9d55a13
- I have co-developed a Python library for spatial interpolation named Polire

Have you worked on other related projects? What knowledge have you gained from working on them?

- latexifying and figure making: As a PhD student I do these things on a regular basis and thus I have gained experience in them.
- Writing a code from scratch for ML concepts: I have implemented the FITC method from scratch in Stheno. I mainly learned how to convert math into efficient code by using the computations tricks.

Please describe your relevant technical background and experience.

I have recently completed my Ph.D. coursework in subjects such as Machine Learning and Advanced Machine Learning. In coursework, I have worked on the minitorch project to implement the backpropagation from scratch in PyTorch. Apart from this, I am working on Gaussian processes for two years and thus have touched upon the probabilistic machine learning concepts. I have published a AAAI paper in 2021 where we have developed a domain-inspired Gaussian process model to model air quality.

What are the challenges you expect when working on your project? How will you mitigate them?

- The volume of the codebase I am going to dive into is huge so I may get lost in several places. I am sure mentors will help me in overcoming this challenge.
- For several ideas, I am proposing something that changes the entire relationship between the book and the redirection of URLs to pyprobml repo. Thus, there could be unforeseen problems that may arise in the future, To be able to resolve this, I am following the GitHub workflow and keep documenting the things I do so that I can take help from the past whenever required in the future.
- As a practitioner, I understand that implementing ML concepts in code is not an easy task and has challenges such as difficulty in understanding the topics and being able to convert math into code. I am sure that the mentors will support me at places where I have full roadblocks.

Please provide a project timeline with clear tasks and how long (in days) you estimate them to take to complete. Highlight important milestones and deliverables.

1) I have started initial work in converting scripts to notebooks, based on that, I anticipate 15 days of work in the conversion of scripts to notebooks.
2) The latexify project should take at max 7 days as I am delivering a process instead of aiming at finishing the end goal itself.
3) I anticipate the interactive demo creation work to take around 14 to 21 days to complete based on the number of demos I am able to contribute.

4) I anticipate 30 days of time solely for writing new code from scratch for concepts presented in book2
5) Contribution to the GP chapter can span over the GSOC period because the deliverables are not fixed in this case.

Is there anything else you'd like us to know?
I have linked my website on the top in case you would like to look at my GitHub profile or any other details.