ML4VA: Machine Learning for Virginia

by N. Rich Nguyen, Ph.D.

Department of Computer Science

University of Virginia

September 13, 2019

1 Objectives

My main objective is to prepare you to apply what you learn in this course to a real-world scenario, especially one that exists from the UVA local community to the state of Virginia at large. Through this 12-week project, you will be working in a team of three students to use machine learning to make meaningful contribution to the **well-being of the state of Virginia and its residents**. The project will provide you with a unique opportunity for exploring one or more areas of machine learning that we covered in the course. You should choose a data set, apply machine learning techniques to it and compare their performance with an well-known solution. If you want to tie the class project to your research project, you are strongly encouraged to do so. However, you should be able to demonstrate the following criteria:

1. **Technicality:** Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the team members convey novel insight about the problem or algorithms?

2. **Significance:** Did the team members choose an interesting or a "real" problem to work on, or only a small "toy" problem? Is this work likely to be useful and have an impact on UVA or the state of Virginia?

3. **Novelty:** Is this project applying a common technique to a well-studied problem, or is the problem or method relatively unexplored?

In order to highlight these components, it is important that you present a solid discussion regarding the learning from the development of your method, and summarizing how your team will plan to complete its deliverable on time according to the following time-line:

- Week 04: Team Registration and Project Introduction
- Week 06: Proposal Due (20 pts)
- Week 10: Checkpoint Due (30 pts)

- Week 14: Video Due (50 pts)
- Week 15: Final Report Due (100 pts)

2 Team, Topic, Dataset, and Platform

Many good class projects come from students picking either an area that they're interested in, or picking some sub-field of machine learning that they want to explore more. There, you should pick something that you can get excited and passionate about! Be brave rather than timid, and do feel free to propose ambitious things that you're excited about. However, be sure to ask us for help if you're uncertain how to best get started.

Pick a Team

Your first task is to pick a project team. We recommend teams of 3 students. The reason we encourage students to form teams of 3 is that, in our experience, this size usually fits best the expectations for the ML4VA projects. In particular, we expect the team to submit a completed project, so keep in mind that all projects require to spend a decent minimum effort towards gathering data, and setting up the infrastructure to reach some form of result. In a three-person team this can be shared much better, allowing the team to focus a lot more on the interesting results and discussion. Each team must complete **a team registration form.**

Select a Topic

If you're looking for project ideas, please come to the office hours of the instructor or the TAs, and we'd be happy to brainstorm and suggest some project ideas. Most teams would pick the application where a machine learning algorithm could help answer some question or solve some problem that result in a positive contribution to the well-being of the state of Virginia or its residents. Some teams would pick an algorithm, improve some of its aspects, and design an experiment to demonstrate the impact to the community. Some project might combine elements of applications, algorithms, and theories.

Search for a Dataset

For this project, we will use the Google Dataset Search Engine to search for a dataset which is applicable to your topic. Because the scope of the project is about solving problems within Virginia, your data set must be related to the state of Virginia. There are many datasets on the state of Virginia which is **available here.** Note that you can look for the available datasets of the left column. Your team is encouraged to explore different search keyword, given that it is related to Virginia, to select the most interesting dataset for your purpose.

Use Google Colab Platform

Since this is a collaborative project, we will use an online Google Colab platform in which team members can write and share code simultaneously. Google Colab uses Jupiter notebook format in simple interface. You may use your UVA computing ID to sign in and share code with other team members. You can sign in Google Colab at **colab.research.google.com.**

3 The Proposal

In the project proposal, you'll pick a project idea to work on early and receive feedback from the instructor and the TAs. If your proposed project will be done jointly with a different class' project, you should obtain approval from the other instructor and approval from us. Please come to the project office hours to discuss with us if you would like to do a joint project.

Mentors

Based on the topic you choose in your proposal, we'll suggest a project mentor given the areas of expertise of the TAs. We will dedicate some office hours to answer your questions and help you out with the project. This is just a recommendation; feel free to speak with other TAs as well. You may choose other mentors (ie. your friend or other professor), but please use their time **sparingly**!

Format

For this project, Overleaf is strongly recommended to collaborate and share your latex document throughout the semester. UVA has an agreement with Overleaf and as UVA student you can sign up for free. You can start at www.overleaf.com. After you have signed up for an account, log in and click on "New Project", then "Academic Journal" from the templates. In the template gallery, pick **"Style and Template for Preprints (arXiv, bio-arXiv)"** template, click "Open Template" and start editing the document template. Note that the template already has some existing text which you may overwrite for your purpose of the project, but keep some LaTeXstructures block (ie. subsection, figure, reference, ect) just in case you need them later. Again, you may share this document with your teammates so that everyone can edit using the same source.

Your proposal should be submitted in PDF document (**two** page maximum), giving the title of the project, the project category, the full names of all of your team members, the UVA ID of your team members, and a 300-500 word description of what you plan to do. You will continue to work on this document for the checkpoint (Section 4) and final report (Secion 6), so make sure to get familiar with the template and how to write research manuscript. Your project proposal should include the following information:

1. **Motivation:** What problem are you tackling? Is this an application or a theoretical result?

2. **Dataset:** Presenting an URL to a dataset you found in Section 2.

3. **Related work:** At least one example of prior methodology on the topic are a valuable addition.
4. **Intended experiments:** What experiments are you planning to run? How do you plan to evaluate your machine learning algorithm?

Grading (30 points)

Each team needs only to submit **one** proposal document PDF on Collab before **the specific deadline**. The proposal is mainly intended to make sure you decide on a project topic and get feedback from TAs early. As long as your proposal follows the instructions above and the project seems to have been thought out with a reasonable plan, you should do well on the proposal.

4 The Checkpoint

The checkpoint will help you make sure you're on track, and should describe what you've accomplished so far, and very briefly say what else you plan to do. You should continue to use the same proposal document and write it as if it's an "early draft" of what will turn into your final project. You can write it as if you're writing the first few pages of your final project report, so that you can re-use most of the text in your final report. Please keep in mind that the intended audience is Prof. Nguyen and the TAs. Thus, for example, you should not spend two pages explaining what logistic regression is. Also, we will expect your final write up to be on the same topic as your checkpoint.

Contribution

Your milestone should include the full names of all your team members and state the full title of your project. Please include a section that describes what each team member worked on and contributed to the project. This is to make sure team members are carrying a fair share of the work for project. If you have any concerns working with one of your project teammates, please fill out **this optional teammate evaluation form** (only seen by the teaching staff).

Format

Your milestone should be at most 3 pages, excluding references. Similar to the proposal, it should include:

1. **Abstract:** A short summary describing your project.
2. **Motivation:** What problem are you tackling, and what's the setting you're considering?
3. **Method:** What machine learning techniques have you tried and why?
4. **Preliminary Experiments:** Describe the experiments that you've run, the outcomes, and any error analysis that you've done. You should have tried at least one baseline.
5. **Next steps:** Given your preliminary results, what are the next steps that you're considering?

Grading (30 points)

Each team needs to submit **one** PDF on Collab. It is mostly intended to get feedback from TAs to make sure you're making reasonable progress. As long as your milestone follows the instructions above and you seem to have tested any assumptions which might prevent your team from completing the project, you should do well on the milestone.

5 Video Presentation

The class project will be presented as a 3-5 minute video presentation. Each team should create a video, and be prepared to give a very short story or animation, in front of the camera, about their work. At the video session, you'll also have an opportunity to see what everyone else did for their projects. Google has an **A.I. Experiment** which have many examples of these videos. Here are some good project videos from which you can draw some inspiration:

- Visualizing High Dimensional Space by Smikov [Video Link]
- Identifying Bird Sounds by Manny Tan & Kyle McDonald [Video Link]
- Thing Translator by Dan Motzenbecker [Video Link]

Format

If you choose to do animation, you can also look at an example done by some of my former students. Similar to the examples, videos with nice, illustrative figures are preferred over video with lots of text. Here's a generic guideline for creating this video:

1. **Write a script and create a storyboard:** Based on your proposal and checkpoint, write what and how you are going to say or perform specifically in your video. Also, creating a storyboard helps you plan each step of the process and allows you to seeks and gathers necessary footage and resources.
2. **Film your story/media:** Typically a smart phone camera (in landscape mode) held in a stable manner will be fine. Make sure you done filming in advance and with multiple takes, it might come handy in the editing process.
3. **Edit your media:** You may use any digital media editing software to add soundtracks, voice overs, special effects, captions, and titles. Many free software are available here. Please do **not** spend your money or purchase any software or tools.
4. **Give credits:** Include references, citations, and acknowledgement. If you use photo, graphic, or sound file, be sure to give proper credit to your sources (either original or permission for its use by the owner). All information sources, including 3rd-party media, must be cited in credits at the end of your video. Furthermore, make sure you state the following at the end of your credit: **This work has been a part of the "Machine Learning for Virginia" project at the University of Virginia.**

5. **Upload to YouTube:** Your team must upload your video on YouTube under a public setting or access only by URL. Note that your video must be accessible and viewed by TAs in order for your team to receive any credit.

Grading (50 points)

Each team needs only to submit **one** URL of the video on Collab before **the specified deadline**. The top videos will also be played during the Project Expo. We will be grading videos on the video quality and clarity, the technical content of the video, as well as the knowledge demonstrated by the team when discussing their work with teaching staff at the video session. The video will be evaluated based on the following criteria:

- **Content**: The content includes a clear statement of purpose or theme and is creative, compelling and clearly conveyed. Show strong understanding and in-depth analysis about the topic.
- **Organization**: Media shows high degree of attention to logic and reasoning of points. Unity clearly leads the audience to conclusion and stirs thought regarding the topic.
- **Discussion and Analysis**: Clearly discuss what you have learned from doing the project. Main points are well developed with high quality and quantity support. Reveals high degree of critical thinking.
- **WOW factors**: Amazing story with lots of creativity. The approach is innovative and visually appealing.

Important Note: Your digital media must reflect your own understanding of the topic and be communicated in your own words. Make sure copyright infringement has not occurred. Please refer to this link for more information on what copyright infringement is and how to avoid it. Academic integrity will be strictly observed for the project showcase.

6 Final Report

Because the teaching staff will have only a few hours to see a large number of videos at the video session, we'll only be able to get an overview of the work you did at the session. We know that most students work very hard on the final projects, and so we are extremely careful to give each write-up ample attention, and read and try very hard to understand everything you describe in it.

Format

Final project write-ups can be at most 6 pages long (including appendices and figures) expanding from the document in the Section 4. If you did this work in collaboration with someone else, or if someone else (such as another professor) had advised you on this work, your write-up must fully acknowledge their contributions. For shared projects, we also require that you submit the final report from the class you're sharing the project with. Here's more detailed guidelines of what we expect to see in the final report as a academic paper:

1. **Abstract:** A short summary describing your project.
2. **Introduction:** The introduction begins by introducing the broad overall topic and providing basic background information. It then narrows down to the specific research question relating to this topic. It should be a quick review of the previous methods and the new idea being researched.
3. **Method:** What machine learning techniques have you tried and why? The methods section will describe the design and methodology used to complete to the study. The general rule of thumb is that readers should be provided with enough detail to replicate the study.
4. **Experiments:** Describe the experiments that you've run, the outcomes, and any error analysis that you've done. You should have tried at least one baseline.
5. **Results:** This section should focus only on results that are directly related to the problem. Leaving out your personal opinion, detail your results and provide solid evidence. Graphs and tables should only be used when there is too much data to efficiently include it within the text. *If you have code repository and Jupyter Notebook, make sure you include the URL or supplemented files.*
6. **Conclusion:** This section should be a discussion of the results and the implications on the project to the well being of Virginia and its residents. The hypothesis should be answered and validated by the interpretation of the results. This section should also discuss how the results relate any shortcoming of the findings, and potential for future work.
7. **References:** The paper is not complete without the list of references. This section should be an alphabetized list of all the academic sources of information utilized in the paper.

Contribution

Please include a section that describes what each team member worked on and contributed to the project. If you have any concerns working with one of your project teammates, you can also fill out this optional teammate evaluation form (only seen by the teaching staff). We may reach out and factor in contributions and evaluations when assigning project grades. **Note:** that we expect you to do your part in this team-based project. Up to 100 points may be deducted based on your contribution and partner evaluation.

Codes

All of your code must be in Jupyter Notebook format. Zip all files into as single .zip file and submit the file to Collab along with the final report. If you use Github, please include a zip file or preferably a link to a Github repository with the code for your final project. You do not have to include the data or additional libraries (so if you submit a zip file, it should not exceed 10MB).

Grading (100 points)

Each team needs only to submit **one** copy of your final report on Collab. Making your code submission available to the instructor and TAs is **your** responsibility. If we cannot access your file then you will not get any credit. The final report will be judged based off of the clarity of the report, the relevance of the project to topics taught in CS4774, the novelty of the problem, and the technical quality and significance of the work as stated in Section 1.

FAQs

Should final project use only methods taught in classroom?

No, we don't restrict you to only use methods/topics/problems taught in class. That said, you can always consult TA if you are unsure about any method or problem statement.

Is it okay to combine the CS4774 term project with that of another class ?

In general it is possible to combine your project for CS4501 and another class, but with the following caveats: You should make sure that you follow all the guidelines and requirements for the CS4501 project (in addition to the requirements of the other class). So, if you'd like to combine your CS4501 project with a class X but class X's policies don't allow for it, you cannot do it. You cannot turn in an identical project for both classes, but you can share common infrastructure/code base/datasets across the two classes. Clearly indicate in your milestone and final report, which part of the project is done for CS4501 and which part is done for a class other than CS4501. For shared projects, we also require that you submit the final report from the class you're sharing the project with.

Do all team members need to be enrolled in CS4774?

Yes, and please explicitly state the work which was done by team members enrolled in CS4501 in your milestone and final report. This extends to projects that were done in collaboration with research groups as well.

Can my team have more then 3 members?

In exceptional cases, we might allow a team of 4 people but it must be declared before the proposal deadline. The team size will be taken under consideration when evaluating the scope of the project, meaning that a four-person team is expected to accomplish more than a three-person team would.

Do I have turn in any deliverable in person?

No, all project documents will be submitted via Collab.

Is it okay for team member to miss the project expo?

Part of your project grade part depends on your presentation at the video session, so we really urge you not to miss it. That said, if (and only if) you have a final exam conflict there are a few possibilities If you are working on the project as a team, the rest of the team could present the video without you there. If none of above options work for you, come talk to one of the TAs.

What fraction of the final grade is the project?

The class project is 200 points or 20% of the final grade.

Can I continue work on the project after the course?

After CS4501, you can choose to continue to pursue your project at your pace. Of course, depend on what you like to accomplish with your project, we will able to provide further guidance (ie. additional dataset, more advanced methods) should you require it.

References

[1] Ng, Andrew. "CS229: Machine Learning." CS229: Machine Learning, Stanford University, 2017, cs229.stanford.edu/.

[2] Qi, Yanjun. "2018 Fall CS4501 - Machine Learning." 2018 Fall CS4501 - Machine Learning · 2018 Fall, 2018, qiyanjun.github.io/2018fUVA-CS4501MachineLearning/.

[3] Praphamontripong, Upsorn. "Software Testing." CS 6501, 2018, www.cs.virginia.edu/ up3f/swtesting/syllabus.html.

[4] "AI Experiments — Experiments with Google." Google, experiments.withgoogle.com/collection/ai.

[5] "Dataset Search." Google, toolbox.google.com/datasetsearch.

[6] "Manuscript Templates for Conference Proceedings." IEEE - Advancing Technology for Humanity, www.ieee.org/conferences/publishing/templates.html.