CIS 520: Machine Learning **Spring 2019**

Project Guidelines

1 Project Goals and Overall Process

The goal of the project is for you to demonstrate that you can formulate real-life problems as machine learning problems and can design machine learning solutions to them. To this end, you are expected to propose a data-driven problem that you are excited about (involving real data), to develop and implement a machine learning solution for it, and to write a report explaining what you did. In most cases, the project will involve understanding the problem/data involved and applying some standard machine learning algorithms that are suited to solving the problem; in some rare cases, it may require designing a novel algorithm or a variant of an existing algorithm. Since the project is expected to involve a significant amount of work, the recommended team size for the project is 3-4 students.

Some guidelines on how to go about this:

1. **Pick a supervised or unsupervised learning problem that you would like to solve, together with a suitable data set for it.** E.g. the problem could be to develop a face recognition system, or a document classification system, or a speech recognition system, or a system to diagnose medical images, or a system to identify artists from images of paintings, or anything else. Be creative and imaginative! There is no limit to what type of supervised or unsupervised learning problem you might want to solve; the only two things you need to ensure are that you can find suitable data sets for the problem, and that you can realistically complete the project in the time-frame of the course. As far as possible, it is recommended that you use publicly available data sets (with proper citation); if a data set you use is not publicly available, you must have permission to use it.
2. **Find out what has been done before (and make sure your project is distinct).** Have people tried to solve a similar problem before? Do some online searches to find relevant references (these could be articles on the Web or they could be research publications). What data sets and methods have people used? What difficulties have they faced? What is the current state of the art? It's okay if others have solved a similar problem before; but you should be aware of it and should mention it in your report, and you should make sure your work is not doing exactly the same. (Note that in some cases, there might be more modern, massive-scale solutions available today than what you develop in your project. That's okay; this is likely your first project in machine learning, and we are not expecting you to compete with industry-scale methods!)
3. **Formulate the problem as a machine learning problem (and identify any data processing that needs to be done).** Is your problem a supervised learning problem or an unsupervised learning problem (or does it have components of both)? Is it a classification problem, a regression problem, a clustering problem, or something else? What is the precise form of input (data), and what is the desired form of output (model)? Does the data have any special properties (we encourage you to perform some exploratory analysis on the data to understand its properties)? Do you need to pre-process the data before you can use it, and if so, what type of processing will be needed? How will you measure the performance of the final learned model? Define the performance measures that will be used and make sure you can justify why they make sense for your problem.

4. **Decide on which machine learning methods/algorithms you will use.** In most cases, we encourage you to compare at least 2–3 possible solution methods for your problem. If your problem fits into one of the standard types of problems, you can choose from some standard algorithms for the problem type. (E.g. if you are building a system to recognize artists from paintings and are formulating the problem as a multiclass classification problem, then you might use multiclass logistic regression, multiclass/one-vs-all SVMs, and multiclass neural network classifiers – or you might use multiclass versions of nearest neighbor and decision trees – or you might come up with some base rules and decide to use a multiclass boosting approach – there are many possibilities!) If your problem doesn't quite fit into a standard category, you may need to come up with some new methods/variants of existing methods. In both cases, think about what properties are important for your problem (performance measure, model representation, computation speed in training and testing phases, memory, etc), and choose your solution methods accordingly.
5. **Decide on who will do what, and what your timelines are.** Now that you know what problem your project will tackle (including how it's related to/different from previous work), what types of data processing will be needed, and what solution methods you plan to implement, you should be able to draw up a plan for which parts of the project each team member will do, how long they will take, and how everything will fit together; you should also be able to come up with a clear plan of what you will accomplish by the time of your milestone meeting (more details below). Don't forget to factor in some buffer time for contingencies, and the time needed to prepare your final project report!

At this point, you will be ready to **submit your project proposal**. Your proposal will give us a summary of each of the above steps, so that we know what to expect from your project. Your proposal should also include a thorough description of your data set and summary statistics (e.g. number of data points; number and types of features; means, ranges, and standard deviations of some or all of the features; possible fraction of missing values; etc). We include this requirement to ensure you've loaded your data set into memory before committing yourself to using it.

Once you have submitted your proposal to us, you will be ready to get started on the actual implementation of your project. We will assign one of the TAs as a project mentor for your team; you can reach out to him/her (or to any of the other TAs) if you have questions about your project.[1]

The final project deliverables will be a **project report** (accompanied by your code) and a **short spotlight presentation**.

2 Project Checkpoints/Deliverables

The project has four checkpoints/deliverables:[2]

1. **Project proposal.** This is due as a PDF submission on Canvas by **Tue March 12** (by 11:59 PM).
2. **Milestone meeting.** On **Fri April 5**, at least one member of your team must attend recitation to meet with your project mentor and describe the progress you have made so far. This will also give your team an opportunity to discuss any concerns you may have about your project or any difficulties or challenges you might encounter.
3. **Project spotlight presentation.** All project teams will give short spotlight presentations on their projects on the last day of class, Tue April 30. Slides for the spotlight presentation are due as a PDF submission on Canvas by **Fri April 26** (by 11:59 PM).

[1] Note, however, that each TA will be overseeing roughly 6–8 projects, and is expected to spend only about 1 hour overall for each team being mentored, so please be considerate in reaching out to the TAs; they can give you some suggestions, but ultimately, it is your project and you are responsible for its completion.

[2] No late submissions will be accepted for any of the project deliverables, so be sure to plan well in advance.

4. **Project report.** This is due as a PDF submission on Canvas by **Tue April 30** (by 11:59 PM). At the same time, you also need to submit your **code**, as a zip file and in a form that can be run by our team, also on Canvas (or if you prefer, you can provide a link to a github repository in your PDF submission).

Below we give some details of what will be expected in each of these deliverables; further details will be provided in due course as needed.

2.1 Project Proposal

This should be at most 2 pages, plus cover page and references. It should be submitted as a single PDF file containing the following components:

- **Cover page (1 page).** Include the following:
  - **Project title**
  - **Team member details** (Include names, Pennkeys, and email addresses.)
- **Main document (at most 2 pages).** Summarize the outcomes of the 5 steps discussed above. Specifically, we recommend including the following sections:
  - **Introduction** (Describe the problem and data set, including a pointer to the data set source. Also report summary statistics for the data set here, as disussed at the end of Section 1 above, preferably in a well-formatted table.)
  - **Related work** (Describe related previous work, including citations to references.)
  - **Problem formulation** (Describe your problem formulation, any data processing that will be needed, the performance measures that will be used and why, etc.)
  - **Solution methods** (Describe the machine learning algorithms you plan to use and why.)
  - **Plan of work** (Describe what each team member will do and what your timelines will be. Include a detailed description of what you aim to achieve before the milestone meeting.)
  - **Acknowledgments** (If anyone has given you any inputs on the project, mention this here.)
- **References (no limit).** List all the references that you consulted and that are relevant to your project, including references that describe previous work related to your project (e.g. these could be Web articles or research publications). We encourage you to use bibtex.

2.2 Milestone Meeting

At this meeting, you will present the work you have completed so far to your project mentor. You will explain what each team member has contributed and what your group has achieved. Be prepared to show your mentor the code you've written, the cleaned outputs of data pre-processing steps, and results from a baseline model. If the goals you laid out for the milestone meeting in your project proposal were sufficiently ambitious yet tractable, then your mentor will be checking that you've completed everything you set out to. (If the goals were too ambitious or trivial, then your mentor may ask you to adapt them accordingly.)

2.3 Project Spotlight Presentation

This will be a very short presentation, to be given during the last class. The slides for the presentation should be submitted in advance as PDF files. The exact duration of each presentation and the number of slides allowed will depend on the total number of projects; we will give more details in due course.

2.4 Project Report

This is the main deliverable for the project. It should be at most 5 pages, plus cover page and references, and in most cases should expand on the project proposal. It should be submitted as a single PDF file containing the following components:

- **Cover page (1 page).** Include the following:
  - **Project title**
  - **Team member details** (Include names, Pennkeys, and email addresses.)
  - **Project mentor** (Include TA name.)
  - **Team member contributions** (Include a table highlighting the contributions made by each team member to the project; e.g. these can include problem/data set identification, literature review, problem formulation/definition, writing proposal, data processing, algorithm implementation, model evaluation/visualizing results, preparing presentation, writing project report, etc.)
  - **Code submission option being chosen** (Mention whether you are submitting your code as a zip file on Canvas, or whether you are providing a link to a github repository; in the latter case, provide the link here.)
- **Main document (at most 5 pages).** The report should give a clear, self-contained description of your project together with associated results and observations. We recommend including the following sections (but feel free to adapt as needed):
  - **Abstract** (Give a brief 1-paragraph summary of your project and findings.)
  - **Introduction** (Describe the problem; give some motivation and explain why it is interesting.)
  - **Related work** (Describe related previous work, including citations to references.)
  - **Data set** (Describe the data set you used, including a pointer to the data set source; provide any interesting insights into the data or any special properties it may have, e.g. through summary statistics or some suitable visualizations.)
  - **Problem formulation** (Describe your problem formulation, any data processing that was needed, the performance measures that you decided to use and why, etc.)
  - **Algorithms** (Describe the machine learning algorithms you used and explain why you chose these methods; describe how they were implemented, including programming language used etc.)
  - **Experimental design and results** (Describe how you set up your experiments, including choice of training/testing data, choice of parameters, etc, in a way that your results can be reproduced; then describe your results, including performance measurements as well as any other insights, e.g. training/testing times, possible visualizations of results including both examples of good performance and examples of what types of errors are made, etc.)
  - **Conclusion and discussion** (Summarize your findings and note any observations/additional thoughts you may have, as well as notes on any lessons you learned from the project as a whole.)
  - **Acknowledgments** (If anyone has given you any inputs on the project, mention this here.)
- **References (no limit).** List all the references that you consulted and that are relevant to your project, including references that describe previous work related to your project (e.g. these could be Web articles or research publications). There is a good chance that during your project, you will find additional references relevant to your project beyond those you included in your proposal; include these as well and discuss them in your 'related work' section. We encourage you to use bibtex.

Remember: At the same time as your report, you also need to submit your code for the project, in a form that can be run by our team, either in a zip file on Canvas or as a link to a github repository (as noted above, you should specify on the cover page which option you are using; in the latter case, provide a link to the github repository on the cover page).

3 Grading/Evaluation

Grading for the project will be based largely on the project report, although the proposal, milestone meeting, and presentation (and possibly code) will also count toward the project grade. We will not be disclosing the exact breakdown of the project grade, but the following criteria will be used in evaluating the projects:

- **Technical quality** (Is the problem well-formulated? Are the algorithms chosen well-suited to the problem? Is the experimental design sound? Do the results make sense? Are the principles of machine learning applied in a correct manner? Are there any major gaps in the authors' approach/understanding?)
- **Significance** (Is the problem important or meaningful?)
- **Novelty** (Is this a 'new' problem? Did it involve some interesting data or solution approach?)
- **Clarity of presentation** (Is the report well-written? Is it easy to read? Are the results presented and/or visualized in a clear and helpful manner?)

Good luck with your projects. We look forward to reading about your work!