Final Project Instructions
CMPSCI 589 - Fall 2016

Project Proposal: Due Th. Nov 10, 2016 by 1:29pm. 5% of final grade.
Final Report: Due Mon. Dec 12, 2016 by 1:29pm. 30% of final grade.

Late days cannot be used for either the project proposal or the final report.

(Both the proposal and report should be submitted through moodle)

General Guidelines:

We expect to see a creative and well-executed project where you select one or more data sets, frame a machine learning problem, build a data processing pipeline to solve the problem by combining existing components or implementing new components, conduct experiments to evaluate properties of your pipeline or compare multiple alternatives, and write up your results in conference paper format. Your grade will depend on how creative and well executed your project is, and how much effort you put in to it. We do not expect all projects to be successful at matching or improving on state-of-the-art results in established problem domains.

Projects can be completed in groups of up to three students. The expectation is that groups of 2 or 3 will do 2 or 3 times more work than a single student could accomplish on their own. Students that are involved in research can tie this project to their existing work, but all of the experimentation done for the course project must be new, and whatever you submit for the course project must be entirely your (or your group's) own work.

Specific Requirements for Project Proposals:

Your project proposal is a short description of what you plan to work on for your final course project. You should read the specific requirements for the final project report before starting to draft your proposal. You may change your mind about some of the details as you go, but you should contact us if you decide to completely change your project topic. The maximum page limit for the proposal is **1 page** plus an additional page for references, and a **1 page** collaboration plan for groups of two or more. Students working in groups should each submit a copy of the project proposal.

Your proposal should include:

1. **Title:** Select an informative title for your project
2. **Author(s):** List the names of all group members (your name if working alone).
3. **Data Sets:** A brief description of the data sets you plan to use including a link to the data sets if possible or a clear description of how you will collect or get access to the data. Links to data resources will be provided.

4. **Problem:** A clear description of the machine learning problem you will formulate over your data set. In most cases this will be a classification, regression, dimensionality reduction, or clustering problem.

5. **Methods:** A sketch of the pipeline you plan to implement including a description of the methods you intend to apply to the data set and what code libraries you will use or implement. Your project should integrate methods from both the supervised and unsupervised learning sections of the course.

6. **Experiments:** A description of what experiments you intend to perform to validate your pipeline, study its properties, or compare to other existing alternatives.

7. **Related work and Novelty:** A statement of what others have done with the same or similar data/task/problem (including citations), and how your project will do something different or novel.

8. **Collaboration Plan:** For groups of 2 or 3, provide an additional summary of which group members will do what work on the project (up to one additional page). There must be a clear separation of tasks among group members and the work must be equally distributed across the group.

9. **References:** Provide a list of references to support your assessment of related work. You can provide up to one additional page of references.

Specific Requirements for Final Projects:

Your final project should follow standard machine learning paper structure including the following sections. The number of pages per section listed below should be taken as a rough guide, but there is a firm upper limit of **5/8/12 pages** excluding references for groups of 1/2/3 students. Your report should be prepared in NIPS conference format. All group members should submit an identical copy of the final project report.

1. **Title:** Select an informative title for your project

2. **Author(s):** List the names of all group members (your name if working alone).

3. **Introduction (0.5-1 pages):** An introduction describing the problem you are solving, a discussion of why you think it's important or interesting, and a summary of your solution and findings.

4. **Related Work (1-2 pages):** A related work section summarizing 3-5 pieces of prior research related to the problem your project addresses. You can use Google Scholar http://scholar.google.com/ to help identify relevant papers. Look for papers in good quality venues such as JMLR, NIPS, ICML, UAI, AISTATS, AAAI, IJCAI, KDD, etc.

5. **Data Set(s) (0.5-1 Pages):** Describe your data set(s) including where it was obtained from or how you collected it. Describe the number of data cases, the num-

ber of features, what the features represent and what their data types are, etc. For data sets with large numbers of features, you should provide a summary of the features and not an exhaustive listing (include a reference to a published paper or website that describe the data in more detail for large data sets if possible).

6. **Proposed Solution (1-4 pages):** This section will describe the pipeline you have built in detail. Your pipeline may include pre-processing components (missing data imputation, feature selection, feature learning, dimensionality reduction, etc.), one or more core models (classification, regression, dimensionality reduction, clustering), and hyper-parameter or model selection methods (cross validation etc.). You should include mathematical descriptions of the models used and indicate the considerations you took into account when selecting and evaluating components.

7. **Experiments and Results (1.5-4 Pages):** You must perform experiments to explore some aspect of your pipeline and compare it to one or more alternatives. Typical experiments may involve the use of cross validation to optimize hyperparameters, and testing several different models to determine which works best on your data. You must carefully describe the experiments you perform and report the results using suitable figures (your report must contain at least 2/4/6 results figures or tables for groups of 1/2/3).

8. **Discussion and Conclusions (0.25-1 Pages):** Discuss the results of your experiments. How do your results relate to what has been reported in the literature previously? What seemed to work well and what didn't? Did you run into any particular problems? What else would you have done if you had more time?

9. **References:** Provide a list of references to support your assessment of related work. You can provide up to one additional page of references.

Final Project Marking Scheme:

The following criteria will be taken into account when marking final project reports:

1. **Creativity:** Did you use an existing data set to pose a new problem? Did you engineer or learn new features or representation for the data? Did you leverage multiple data sets in a novel way? Did you collect a new data set? Did you combine methods in a novel way? Did you investigate methods not covered in assignments or in class?

2. **Clarity/Relevance:** Did you clearly describe the problem you are trying to solve and the proposed solution? Is the problem a machine learning problem?

3. **Related Work:** Did you give an appropriate discussion of the relationship between your problem and previous work? Did you include references and use citations appropriately?\l

4. **Experiments and Results:** Are your experiments well designed? Did you select

hyperparameters in a valid way? Did you compare methods in a valid way? Did you present results using appropriately selected graphs? Do your experiments support the conclusions you draw from them?

5. **Figures/Tables/Writing:** Is your report easily readable? Are the figures and tables properly labeled?

6. **Reproducibility and Code Quality:** Did you design your code so that your results are reproducible? Is your code well documented?

Data Sets:

There are lots of interesting data sets available on the web. One way to make your project more interesting is to think about new and creative ways to formulate machine learning problems over existing data sets that may have been used for other purposes in the past.

You can also collect a lot of interesting data on the web, but if you do, make sure you follow the acceptable use policy of any web sites you collect data from.

You can use data from existing machine learning competitions like Kaggle, but basing your project on participating in a Kaggle competition is generally too limiting and will result in a poor creativity score if the data set is already highly curated and well-described, with a single problem of interest.

Resources for other data sets will be posted to the project section of Piazza.

General Project Advice:

• Be selective! Don't choose a project that has nothing to do with machine learning. Don't attempt to address a problem that is irrelevant, ill-defined or unsolvable.

• Be creative. Try to pick a project that looks at existing data in a new way or uses it to solve a different problem. Don't be afraid to use messier data that has not already been curated for use by machine learning algorithms. Projects that solve previously well defined problems using existing data sets from the UCI archive or Kaggle are unlikely to score well in terms of creativity unless you're doing something very novel on the methods side.

• Be honest. You are not being marked on how good the results are. What matters is that you try something sensible, clearly describe the problem, your method, what you did, and what the results were.

• Be modest. Don't pick a project that is too hard or that will require more computation than you have access to. Usually, if you select the simplest thing you can think of to try, and do it carefully, it will take much longer than you think.

- Be careful. Don’t make basic mistakes like test on your training data, set parameters by cheating, compare unfairly against other methods, include plots with unlabeled axes, use undefined symbols in equations, etc. Do sensible crosschecks like running your algorithms several times.

- The final report is due on the second to last day of class, Dec. 12. This is only four days before the final exam! It is strongly suggested to get it done early so that you have time to prepare for the final (and other classes).

- Have fun! If you pick something you think is cool, that will make getting it to work less painful and writing up your results less boring.