This semester you have learned about many aspects of machine learning – fundamental algorithms for classification and clustering, theoretical foundations, evaluation methodology, societal implications, etc. Over the next few weeks you will have the opportunity to craft personalized assignments that will allow you to explore topics of particular interest to you. You may do this multi-week assignment on your own, or you may work with a partner. It will be convenient if your project partner is your tutorial partner, but it is not necessary. You may choose to work with someone else in the class.

- **Part 1**: Your deliverable for the week of November 28 will be a project proposal written in the form of a CSCI 374 assignment. It should include an introduction, including motivation (i.e., why is this a valuable topic? why do we care?). It should also include a list of readings. Finally, it should include a description of what you will do beyond any reading. Complete a set of problems? Write a program and demonstrate it? Perform an empirical analysis and write it up? Give a presentation? Note that on some level each of you will do the latter. That is, Part 2 of this assignment requires that you present your project.

- **Part 2**: This is where you will present (and turn in) the work you have done. I hope that we will be able to meet as a class for these presentations, as I think it will be informative and fun to share your work with each other. If we can’t find a time to get everyone together, a good option might be meeting in smaller groups (of, say, five or six). If neither of these works out, we can always fall back on meeting in pairs, as usual. I will send out an email getting your availability and preferences. Note that if we go with one of the “large group” options, we can consider going into Reading Period or even later, if that’s what you all prefer. No matter what, I’ll aim to get us together for a celebratory lunch or dinner.

In the following sections, I set out some ideas for projects. These are merely intended to get you thinking about the range of possibilities. By providing them here, I do not mean to imply that you must select one of them. **Note that you are more than welcome to drop by to discuss your ideas. I am happy to provide pointers to papers and other resources.**

## 1 Idea 1: Apply Machine Learning to a Real Data Set

You now have a solid set of machine learning tools, so why not try them out on some real data. There are many great sources of data, including

- The UCI Machine Learning Repository
- Kaggle – a data science website that includes data sets, competitions, etc.
- Miscellaneous data made public by various institutions (such as the US Government) and companies.

I’ve identified a few data sets that are rich and interesting in different ways, just as inspiration. Below you’ll find descriptions of the data sets as well as a set of special rules, should you decide to pursue a project along these lines.

### 1.1 Sentence Classification

http://archive.ics.uci.edu/ml/datasets/Sentence+Classification

From the “README” for the data set: This corpus contains sentences from the abstract and introduction of 30 scientific articles that have been annotated (i.e. labeled or tagged) according to a modified version of the Argumentative Zones annotation scheme. These 30 scientific articles come from three different domains: (1) PLoS Computational Biology (PLOS), (2) The machine learning repository on arXiv (ARXIV), (3) The psychology journal Judgment and Decision Making (JDM).

Also from the “README”: Argumentative Zones (AZ) is a scheme for classifying (i.e. annotating, labeling, or tagging) sentences according to function. There are seven labels in the original AZ scheme:
1. AIM: “A specific research goal of the current paper”
2. TEXTUAL: “Statements about section structure”
3. OWN: “(Neutral) description of own work presented in current paper”
4. BACKGROUND: “Generally accepted scientific background”
5. CONTRAST: “Statements of comparison with or contrast to other work; weaknesses of other work”
6. BASIS: “Statements of agreement with other work or continuation of other work”
7. OTHER: “(Neutral) description of other researchers’ work”

There is a lot of data here: labeled articles, unlabeled articles, lists of words that might be associated with the various classes, etc. But the data are in quite raw form. That is, the examples (i.e., the sentences) are not expressed as vectors of attribute values. It would be up to you to decide what the attributes should be. If you wanted to pursue a project involving text, you’d want to do some reading on extracting features from text.

1.2 LSVT Voice Rehabilitation
https://archive.ics.uci.edu/ml/datasets/LSVT+Voice+Rehabilitation

As described by the contributor of the data set, the aim here is to assess whether voice rehabilitation treatment leads to phonations considered ‘acceptable’ or ‘unacceptable’. (So this is a binary classification problem). The paper associated with the data set shows that it’s possible to achieve 90% accuracy, so your goal will be to see whether you can do better. One of the challenges here is that there are 309 attributes but only 126 examples.

1.3 Urban Land Cover

This problem involves “classification of urban land cover using high resolution aerial imagery.” As the contributor says, there are a low number of training samples for each class (14-30) and a high number of classification variables (148), so it would be an interesting data set for testing feature selection methods.

1.4 Credit Card Fraud Detection
https://www.kaggle.com/dalpozz/creditcardfraud

This data set contains 284,807 credit card transactions that occurred over two days, 492 of which were fraudulent. Clearly the data set is highly unbalanced, so if you did something like this, you would want to do some reading on dealing with highly unbalanced data.

1.5 Special Rules for Projects that Apply Machine Learning to Real Data

If you plan to try your hand at something like this:

- Once you’ve decided on a data set, let me know. Some data sets are quite large to begin with, and as you generate training sets, test sets, and results, you’ll be using both a great deal of disk space and processing time. We’ll want to coordinate with Mary to be sure you have the resources you need while not causing problems for other students or for the lab in general.

2 Idea 2: Explore a New Learning Algorithm

We’ve explored classification algorithms quite extensively, but there are many more algorithms out there. For instance, you might want to learn more about clustering algorithms by implementing a hierarchical clustering algorithm. Or you might want to explore reinforcement learning by implementing Q-learning. (Note: Don’t do the latter if you’re planning to take AI.)

You might also consider lists of "top algorithms for data mining", such as those found here:

http://www.datasciencecentral.com/profiles/blogs/top-10-machine-learning-algorithms
3 Idea 3: Explore an “Issue” in Machine Learning

Over the course of the semester we’ve discussed situations that are problematic for our algorithms, such as the “curse of dimensionality” or the difficulty of learning from unbalanced training data. You might want to explore techniques for handling these problems. For example, you might read up on feature selection techniques, implement one or two, and then apply them to a variety of data sets and analyze the results.

This semester we’ve also focused primarily on nominal-valued data that are static in time. You could explore techniques for dealing with time series data.

4 Idea 4: Explore a Machine Learning Toolkit

This is a very practical idea, but it might also allow you to explore a class of algorithms that we haven’t explored through implementation. For instance, you might choose to learn about TensorFlow (https://www.tensorflow.org/) a machine learning library that is especially well suited for learning with deep networks. This would involve downloading and installing quite a bit of software, so you’d want to think very carefully before choosing this or anything similar. (Do you really want all this on your laptop?)

4.1 Special Rules for Projects that Involve Exploring a Toolkit

If you plan to try your hand at something like this:

- Do not assume that you can install the code for such a project on the lab machines. If there is a toolkit that involves minimal installation, we can explore the possibility with Mary, but do not download anything without first discussing it with me.

5 So many other ideas...

All of the above suggestions are, in a sense, application-oriented. I have focused on those, as we have done less this semester in the way of applications and much more on theory and evaluation. However, there are many more ideas to explore in the non-application realm.

For inspiration, you might skim articles in high quality machine learning conferences and journals. You can find pointers to those from the course “Resources” page.