Projects

Project Details

The course project will consist of groups of three students working together. There are **two** options for you to pick your project; one is more useful if you are interested in learning machine learning, but not necessarily pursue it as a research career option, and the other is more suitable for students who are already familiar with machine learning and want to handle something more advanced. The two options are:

1. **Datasets**: Implement and analyze an application of machine learning to a given dataset, amongst the ones listed below. (more details: **here**)
2. **Independent project**: Submit a project abstract of your interest before **November 5**. *For advanced students only!* (more details: **here**)

Forming a Team

Deadline: November 7

Once you have identified your teammates and decided on a project, coordinate, and do the following:

1. Come up with a short and simple team name
2. Add yourself to a **project group** on Canvas (see detailed instructions below), and rename the group to be the team name following the format "[team_name]([topic])" where "[]" is to be replaced.
3. Once everyone has been added, one of you add the topic you agreed upon after your team name, before the deadline at Canvas
   - **Option 1:** Topic should be the name of the dataset
   - **Option 2:** Topic should be the title of your project.
4. If you're doing Option 2, we will inform you of the appropriateness of your proposal after the proposal deadline.

Joining a Group on Canvas

1. Go to **Canvas (https://canvas.eee.uci.edu/courses/20234)** , and navigate to 'People' on the left navigation bar.
2. Click the 'Groups' tab. You should see a list of potential groups to join.
3. If your team has not already claimed one, join one of the *empty* "Project Group: X" groups. Otherwise, join the same group as your teammates. **Warning:** Do NOT use the "+GROUP" button to create a new group!
4. Change the group name from "Project Group X" to your team name. To do this, navigate to your group's homepage and click the "Edit Group" button. Note: You will only be able to do this if you are the first member of your team to join the group.

Option 1: Datasets

You have the choice of the following six datasets:

Dataset	Type	#Instances	#Labels	Each Instance	URL
Diabetes 130-US hospitals	Tabular	100K	3	49 features	UCI ML Repo (https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008)
Dota2 games results	Tabular	103K	2	116 features	UCI ML Repo (https://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results)
20 Newsgroups	NLP	20K	20	Newsgroup document	Website (http://qwone.com/~jason/20Newsgroups/)
IMDB Reviews	NLP	50K	2	Movie review	Website (http://ai.stanford.edu/~amaas/data/sentiment/)
Fashion-MNIST	Vision	70k	10	28x28	Github (https://github.com/zalandoresearch/fashion-mnist)
CIFAR 10	Vision	60K	10	32x32	Website (https://www.cs.toronto.edu/~kriz/cifar.html)

Project Requirements

I am looking for several elements to be present in any good project. These are:

1. Explore the various aspects of the data, visualizing and analyzing it in different ways. It is really important that you are familiar with it. You should describe how you made various design choices, based on the dataset exploration.
2. Exploration of at least one or two techniques on which we did not spend significant time in class. For example, using neural networks, support vector machines, or random forests are great ideas; but if you do this, you should explore in some depth the various options available to you for parameterize the model, controlling complexity, etc. (This should involve more than simply varying a parameter and showing a plot of results.)
3. Other options might include feature design, or optimizing your models to deal with special aspects of the data (missing features, too many features, large numbers of zeros in the data; possible outlier data; etc.). Your report should describe what aspects you chose to focus on.
4. Performance validation. You should practice good form and use validation or cross-validation to assess your models' performance, do model selection, combine models, etc. You should not simply try a few variations and assume you are done.

5. Adaptation to under- and over-fitting. Machine learning is not very “one size fits all”; it is impossible to know for sure what model to choose, what features to give it, or how to set the parameters until you see how it does on the data. Therefore, much of machine learning revolves around assessing performance (e.g., is my poor performance due to underfitting, or overfitting?) and deciding how to modify your techniques in response. Your report should describe how, during your process, you decided how to adapt your models and why.

Your team will produce a single write-up document, **approximately 4 pages long**, describing the problem you chose (with dataset analysis) and the methods you used to address it, including which model(s) you tried, how you trained them, how you selected any parameters they might require, and how they performed in on the test data. Consider including tables of performance of different approaches, or plots of performance used to perform model selection (i.e., parameters that control complexity). Within your document, please try to describe to the best of your ability who was responsible for which aspects (which learners, etc.), and how the team as a whole put the ideas together.

You are free to collaborate with other teams, including sharing ideas and even code, but please document where your predictions came from. (This also relaxes the proscription from posting code on Piazza, at least for project purposes.) For example, for any code you use, please say in your report who wrote the code and how it was applied (who determined the parameter settings and how, etc.) Collaboration is particularly true for learning ensembles of predictors: your teams may each supply a set of predictors, and then collaborate to learn an ensemble from the set. To encourage such discussion, you can open up posts in Piazza by allowing individual students and instructors to see or use Canvas/Discussions to discuss.

Some possible components of a successful project include:

- **Semi-supervised methods**: investigate how your knowledge of the test features can be used to improve prediction. As examples, see e.g., label propagation (**http://www.cs.cmu.edu/~zhuxj/pub/CMU-CALD-02-107.pdf (http://www.cs.cmu.edu/~zhuxj/pub/CMU-CALD-02-107.pdf)**), or using EM (within e.g. naive Bayes or a Gaussian mixture model, e.g., **http://www.kamalnigam.com/papers/emcat-mlj99.pdf (http://www.kamalnigam.com/papers/emcat-mlj99.pdf)**).
- **Kernel learning**, or similarity/metric learning of the measure of dissimilarity used in, for example, nearest neighbors or SVMs, to improve their performance. See for example Weinberger and Saul 2008, **http://www.cse.wustl.edu/~kilian/papers/jmlr08_lmnn.pdf (http://www.cse.wustl.edu/~kilian/papers/jmlr08_lmnn.pdf)**.
- **Neural networks and deep learning**; using existing packages like PyTorch, Keras, MxNet, and PyLearn2.
- **Support vector machines.** For example, you could investigate the effect of different kernel choices, regularization, etc.). The implementation libsvm is pretty good.
- **Go in-depth with ensembles.** Use lots of learners, stacking, and information from your leaderboard performance to try to improve your prediction quality.
- **Feature selection methods**, such as stepwise regression (or in this case, classification); e.g. **http://en.wikipedia.org/wiki/Stepwise_regression (http://en.wikipedia.org/wiki/Stepwise_regression)**. (Note: if you use feature selection, you should use a predictor that is

sufficiently complex to need feature selection!)

- **New Features** Techniques for creating new features, including “kitchen sink” features (**http://books.nips.cc/papers/files/nips21/NIPS2008_0885.pdf (http://books.nips.cc/papers/files/nips21/NIPS2008_0885.pdf)**), clustering-based features, etc. Once you have many features, of course, you may also have to explore feature selection (see above) or regularization to control complexity.
- **Sophisticated decision tree structures**, e.g., **http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.1587 (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.1587)**.
- etc.

Option 2: Independent project

You and your teammates are supposed to **submit** an abstract within 100 words to illustrate your idea(s). In the abstract, be sure to include a concrete dataset resource you are going to use, what problem(s) you are going to solve, and your expectations of the experiments. The deadline of the abstract is **November 5**. All submissions after it will not be considered. We will give a short feedback regarding the appropriateness of the projects.