Small group projects

We've concentrated on *classification* because among supervised learning problems, classification is most different from what you might learn in a statistics class (often *regression*) and the foundation for many methods of *recommendation* and *prediction*.

Big idea: if you can classify, you can often predict or recommend.

Now you can move to small-group projects and work intensively on one dataset with the help of our instructors, TAs, and industry consultants.

Suggested datasets: Here are some suggested datasets. I really think starting with one of these will allow you to get the furthest, because we've cleaned the data and thought through solution strategies already:

- The Titanic! Survival classification/prediction. I actually haven't cleaned this data, but it's so well-studied that it is not too hard. Data already provided.
- The Wisconsin breast cancer data set (diagnostic). It has the information about biopsies and your challenge would be to find the best model. `https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)`
- The City of Austin animal shelter outcomes. Can you classify dogs into quick-adoption versus slow-adoption versus transfer, for instance? I can provide the data – main problem is lots of categorical variables.
- Diabetes: I know of a dataset with de-identified information about real-world hospital admissions for diabetes. Can you predict who will be re-admitted soon? Dataset at `https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008`. It'll be hard due to some categorical values...
- Economic data: can you accurately classify which states voted for Trump and which for Clinton in the last presidential election based on economic data? We can help you find data.

You can choose your own dataset, too, but what I ask is that you pick a dataset in which you have data observations (features) and some specific outcome (a label) for each observation. This is required for a supervised classification problem.
There are tons of datasets in the world, and you have decades of life in which to grapple with them, so you might not be able to do your ideal machine learning problem this week! That's ok – start it next week and email me about it!

Process:

1. Pick a dataset and ask the classification question (usually, "What gives the most accurate classification on test data?").
2. Explore the data.
    - What are all the columns? What do they mean? Dig into this.
    - What are summary statistics on each variable? Find the average, variance, and standard deviation for each variable. Find the min and max of each variable. Find how many records/observations have *nothing* for that variable (just failed to record an answer).

- How do the variables co-vary? That is, what's the covariance of all the variables? Use techniques like the heat map and scatter matrix in Python to visualize this – these are really powerful. See the Titanic data analysis Jupyter notebook.

3. Once you've explored the data, divide your data into training and testing sets, or if you want to be very sophisticated, training, validation, and testing sets.

4. Think about which models might be best for your data: decision tree? random forest? linear support vector machine? support vector machine using the kernel trick? k-nearest neighbors? neural network? You can try them all or dig into a few!

5. Start modeling and comparing models. This is a great place to split up the work and assign certain people to certain models. Try all the models listed. Which is most accurate? Can you play with parameters to make one more accurate than another?

6. How are you going to present this? We heard from many speakers that the storytelling and communication aspect of work in machine learning was crucial. Start working on telling people about your data and justifying your decision-making regarding models.