# Project 2: Building a Classifier

*MATH 456, Spring 2016*

[Schedule]

## Overview

You will choose a data set from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.html) and build a model to predict / correctly classify a response variable. You will write up your analyses in the form of a journal article.

## Instructions

1. Visit the UCI ML Repository and browse through the data sets available. Use the filters on the right hand side to apply the following filters, or use this link to go directly to the resulting 94 data sets.

- Default Task: Classification
- Data Type: Multivariate
- Attributes: 10 to 100
- Format Type: Matrix

You can use the filters on the right hand side to further narrow down the data sets to find one that interests you.

2. Write which data set you are wanting to use on the Class Google Document titled "Machine Learning Project List". One data set per person.
3. Get your project approved by the specified deadline.
4. Start analyses and writing.

## Report details.

Your paper should have the following sections. Use Markdown formatting to clearly define each section. - Introduction: What question is being addressed and why? - Data: A description of the data and how it was collected. (Don't say UCI) - Include proper citations. What is the response variable and generally describe the predictor variables. How many records? - Methods: A **description** of the analyses and algorithms used. (No results here!) - Use a 70/30 testing/training sample split. - Use variable selection and/or data reduction tools as part of your model building process. - Use at least 2 different types of classifying algorithms to build your predictive model. - Specify what criteria you are using to assess model performance (at least 2). - The results: The results of all analyses performed (including model selection). - Produce a table that compares model performance on the training sample. - Create at least 2 figures that graphically convey results from the model building, selection, or testing phase. - Conclusion: How well does the chosen algorithm fit on the testing sample?

Examples of journal articles using predictive algorithms (some are full texts, some are just abstracts)

- http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144572
- http://bfg.oxfordjournals.org/content/early/2016/04/03/bfgp.elw005.short?rss=1

- http://dst.sagepub.com/content/9/1/86.full.pdf+html
- http://www.ijcaonline.org/archives/volume107/number1/18717-9939
- http://online-journals.org/index.php/i-jet/article/view/4189

## Submission items

### Project Approval

On the specified deadline submit via BBLearn your proposed project for approval. You can write the answers directly in the submission text, or submit a document that contains this information.
a. What data set are you wanting to use? Provide the direct link to the UCI information page for this data set. b. Describe *in paragraph form* what you will be predicting and with what type of information. Do not simply list all attributes in the data set. c. What types of data management problems are you likely to encounter? Do you feel confident that you can successfully address these problems?

### Report version 1

This should not be considered a *rough draft* or something that you slap together the night before it's due. This should be a real first version of your paper, as complete and clean as you can make it prior to submission. Show all R code (this is so your peers can help out debug problems). This will be submitted as a turn-it in assignment on BBlearn.

### Peer Review

Using the rubric provided (will be adapted from this Cornell College Rubric, you will evaluate and comment on three of your peer's papers. Consider the following advice from an article from an instructor at Westmont

> "...students to be habitually more affirming and less critical of each other's work than they should be. As a reviewer, your job is not just to find nice things to say. Your job is to test arguments for their strength and identify problems to correct. In your reviewing, consider encouragement optional and specific correction required."

Please see me if you have any comments or concerns about the reviews or suggestions you receive in your peer feedback.

### Final Version

Revise your paper to address the comments and suggestions provided by your peers. Turn this version into something that could be published by hiding all R code and output (with the obvious exception of graphics and tables). This report should be knitted as a PDF document. If you are having problems with your LaTeX installation then it is advised that you

## Grading

The final project grade is out of 100 points and will be based on your project approval (4 pts), the first version of the report (20pts), your peer evaluation of 3 classmates (6pts), and the final version (70pts).