# Exam

[Submit Assignment](https://webcourses.ucf.edu/courses/1268093/assignments/5375721)

* **Due** Wednesday by 11:59pm

* **Points** 45

* **Submitting** a file upload

The following exam is to be done on your own. **Any indication that you have worked with others will result in a grade of 0.**

### Question 1

Download 'regression\_data.csv' and 'regression\_data.readme.txt' from the files section of the course website and identify the best model for the data from the following set of possible models:

* Ordinary least squares
* Lasso Regression with a regularization coefficient of either 0.001, 0.01, 0.1, 1, or 2
* Kernalized Ridge Regression with a regularization coefficient of 1 and
  + a linear kernal
  + polynomial kernels with coefficients gamma = 1, r = 1, and M of either 2, 4, 7
  + isotoptic gaussian kernel with coefficient sigma of either 0.1, 0.5, 1, 2, 4

### Question 2

Download 'classification\_data.tsv' and 'classification\_data.readme.txt' from the files section of the course website and identify the best model for the data from the following set of possible models:

* Kernalized Ridge Regression with a regularization coefficient of 1 and
  + a linear kernal
  + polynomial kernels with coefficients gamma = 1, r = 1, and M of either 2, 3
  + isotoptic gaussian kernel with coefficient sigma of either 0.1, 0.5, 1, 2, 4
* Logistic Regression
* Hard margin SVM
* Soft margin SVM with a coefficient of either 0.1, 0.5, 1, 2, 5

### Question 3

Download 'classification\_data2.txt' and 'classification\_data2.readme.txt' from the files section of the course website and identify the best model for the data from the following set of possible models:

* Kernalized Ridge Regression with a regularization coefficient of 1 and
  + a linear kernal
  + polynomial kernels with coefficients gamma = 1, r = 1, and M of either 2, 3, 4
  + isotoptic gaussian kernel with coefficient sigma of either 0.1, 0.5, 1, 2, 4
* Logistic Regression
* Hard margin SVM
* Soft margin SVM with a coefficient of either 0.1, 0.5, 1, 2, 5

## Solution Submission:

Each solution shall be contained within a separate jupyter notebook (ending with**an .ipynb extension**).   
Each solution shall be coded in **Python 3**

Solutions that are not submitted as a jupyter notebook, or not written in Python 3, will only be worth **50%** credit.

## Point Distribution (15 points per question):

Split the data set into a training set (80%) and a testing set (20%) (1 point)

* Explore the data
  + Calculate and display the summary statistics of the data (1 point)
  + Create and display histograms and scatter plots of the attributes (1 point)
  + Calculate the correlation values between attribute columns (1 point)
* Preprocess the data
  + Drop attributes that provide no useful information, if any exist (1 point)
  + Transform/scale the data if required (1 point) – the data exploration will inform this decision
* Train a set of models and select the model (algorithm + hyperparameters) with the best performance
  + Use 5-fold cross-validation (1 point)
  + Use grid or random search (1 point)
    - For regression, measure performance with both the R^2 value and the mean squared error
    - For classification, measure performance with a confusion matrix
* Measure your final model on the test data (1 point)
  + For regression, measure performance with both the R^2 value and the mean squared error
  + For classification, measure performance with a confusion matrix
* Documentation (6 point)
  + Each stage of your solution must be accompanied by a description of what you are doing and why.
  + At a minimum, you must explain:
    - Why are you splitting your data into a training and testing set?
    - Why are you dropping certain attributes?
    - Why are you normalizing or scaling your data?
    - What is cross validation and why are you using it?
    - Why are you using a grid search?
    - Why is one model better than another, what does the performance measure actually tell you?

**-5 points if your solution is incorrect**, to a minimum score of 0.