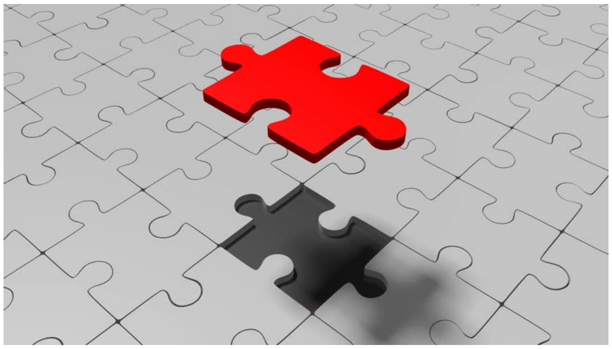
Project 3 : Imagine

Due 4/19/2019 @ 12pm (noon)

TOPICS:

- Python/IPython
- K-Cross Validation
- Handling Missing Values
- Categorize continuous data



BACKGROUND:
During class, we discussed how to create a validation set to generate a model that is general for testing on a data set for which we do not know the target values. We also discussed how to handle situations where the data is missing.

DIRECTIONS:
In this project, you will use 1) K-cross validation on a set of digit images, 2) imputing for filling in the missing data on Pima Indian Diabetes data, 3) categorize continuous data and fill in missing data in the Titanic data.

I have provided three Jupyter notebooks as prototypes for each part of this project. First, you will implement K-cross validation using several different values of "K" and graph the accuracy of the predictions of the digits. Second, you will replace each missing data in the Pima Indian Diabetes data that I provide using the imputing method we described in class. You will calculate the accuracy of the predictions of the outcomes (last column) based on the other features. Finally, you will categorize the Age and Fare features in the Titanic database I provide and apply a machine learning algorithm we learned in class to make a prediction of survival by optimizing for the accuracy of the test set. You must find a way to deal with the missing data (Age and Embarked). Aside from these constraints, you are encouraged to try different things.

IMPLEMENTATION NOTES:
Any program that does not execute completely without errors will not be graded.

COMMENTS AND STYLE:
Although there will be no formal policy on commenting and style, the reader should able to easily follow the main purpose of the code. Each set of code that does something significant must be commented. The variable names should be easily recognizable and acronyms should be avoided if possible.

Do not be surprised if help is not forthcoming if your code is poorly commented and/or difficult to follow. You have been warned.

PROJECT SUBMISSION:
You will turn in the modified IPython notebooks.

The programs and graphs should be in a single directory named "Imagine". The contents of the directories must be archived in a tarball that is gzipped called Proj3.tar.gz.

Place the gzipped tarball in your Drop Box on Sakai before it is due.

PLEDGED WORK POLICY:
Assignments in Computer Science courses may be specified as "pledged work" assignments by the professor of the course. When an assignment is specified as "pledged work" the only aid that the student may seek is from either the course professor or TAs (including CS Center tutors) that the professor has explicitly specified. On "pledged work" assignments the student may not use the services of a tutor.

For this project, you and your partner will develop code together into shared repositories, so you will see and share work. In addition, you may discuss only **basic programming language syntax** and general computer science concepts with everyone else. Any other communications of the project (e.g., giving your code to someone else or seeing someone else's code) are strictly prohibited except with the professor and TAs of the course. Your code and your implementation of the project must be the product of your own work and that of your partner.