CMP 464 & 788 Mid-Term Project – Stage 2

During the first stage of the midterm project, we explored the dataset from Kaggle Titanic project and handled missing values. For the second stage, we will apply Machine Learning models that we have learned to build classifiers on the dataset, and then evaluate their performances. This stage is due on **Monday, March 19 at 11:59pm**. You are expected to submit a complete Jupyter notebook with python code (with detailed comments), numerical results, and graphs (with proper labeling) to Blackboard.

1. Data preparation

1) Sex in the original dataset is categorial, and thus cannot be directly fed into Machine Learning models. Please convert the categories into 0's and 1's.

2) Embarked is also categorical. However, it is not ideal to simply convert the three categories into 0, 1, and 2, since it creates a bias by imposing an order to the values. Instead, we can add an indicator variable (also known as a dummy variable) for each category separately. Please replace the Embarked variable with three variables – Embark_S, Embark_C, and Embark_Q – to indicate the place a passenger boarded the ship. For example, if a passenger boarded from Southamptons, Embark_S should be set to 1, and Embark_C and Embark_Q should be zero. (pandas.get_dummies() can help you with this)

3) Feature Scaling: Machine learning models usually work best on datasets whose variables have similar ranges. Please scale each variable so that it has zero mean and unit variance (sklearn.preprocessing.StandardScaler() can help you with this).

4) Because the test dataset has no information on survival, it cannot be used to evaluate the performance of the model. As a result, we need to further split the data in train.csv into training set (80%) and validation set (20%). We will use training set to build the models, and use validation set to evaluate them. Split the dataset into training set, validation set, and test set. (sklearn.model_selection.train_test_split() can help you with this)

(Optional: Feature Engineering) Previously we noticed that some passengers' names have rare titles, which may indicate their unusual social status. Extracting a new feature from existing data is called feature engineering. Another interesting thing you may observed is that some passenger got on board for free! This could also be a potential indicator of the social status of the passenger. Please create an indicator variable on whether a person paid for the ticket.

2. Building models

So far, we have learned four models: linear regression, polynomial regression, logistic regression, and k-nearest neighbors method. Choose models that are suitable for this problem and use corresponding sklearn classes to fit the data. Note that each model class has several parameters, you are welcome to change the default parameter values and try to improve the performance of the model (in next stage we will fine-tune the models in a systematic manner).

3. Evaluate the models

For each model you have trained, perform the following evaluations:

1) cross validation (show precisions)

2) confusion matrix, precision, and recall

3) precision and recall tradeoff

4) ROC curve and show AUC (area under curve)