作业要求：

- 每人完成总分值为 100 分的候选题目
- 使用的方法模型不限、编程语言不限
- **要求提交物：**
  - 实验报告：应包括任务定义、输入输出、方法描述、结果分析（性能评价）、源码运行环境
  - 代码：除开源工具以外的其它源码或可执行文件
- **提交方式**：每次作业布置后一周内，通过电子邮件发送至课程邮箱 yuansassignment@163.com，邮件主题为：学号-姓名-作业名称

其它说明：

- 关于分组：
  - 不采用多人分组，每人独立完成至少 100 分值的作业
- 关于加分：
  - 如果对于一个题目提供了不同的解决方案，或在一个解决方案之上提供了改进方案，则可额外加最多 10 分，具体根据完成情况确定
  - 最后一次课为作业演示时间，演示者通过 PPT 向大家介绍自己的某一个或几个作业，演示者则可额外加最多 10 分，具体根据演示情况确定
- 诚信说明：经鉴定为抄袭或被抄袭，两种情况均得 0 分
- 每人最高得分为 100 分

Problem assignment

Task 1: Boston Housing Problem

Boston Housing dataset from the CMU StatLib Library that concerns prices of housing in Boston suburbs. A data sample consists of 13 attribute values (indicating parameters like crime rate, accessibility to major highways etc.) and the median value of housing in thousands we would like to predict. The data are in the file housing.txt, the description of the data is in the file housing desc.txt.

Method 1. Linear regression (20 points)

Our goal is to predict the median value of housing based on the values of 13 attributes.

Assume that we choose a linear regression model to predict the target attribute:

- (a) Write a function LR_learn that takes *x* and *y* components of the data and returns a vector of coefficients *w* with the minimal mean square fit.

- (b) Write a function LRs_predict predict that takes input components of the test data (*x*) and a fixed set of weights (*w*), and computes vector of linear predictions *y*.

- (c) Write and submit the program that loads the train and test set, learns the weights for the training set, and computes the mean squared error of your predictor on both the training and testing data set. Which one is better?

Method 2. Bayes classifier (20 points)

Our goal is to predict the level of housing cost (for example, high, medium, low) based on the values of 13 attributes. You should design a method to transform the original continuous values of y into several discrete levels.

Assume that we choose a naive Bayes model to predict the target class:

- (a) Write a function Bys_learn that takes *x* and *y* components of the data and returns a learnt model.

- (b) Write a function Bys_predict predict that takes input components of the test data (*x*) and the learnt model, and assigns a class label for the input *x*.

- (c) Write and submit the program that loads the train and test dataset, learns the weights for the training set, and computes the accuracy of the classifier.

Method 3: SVM for classification or regression (30 points)

Support vector machines have been used in a variety of classification and regression applications. In this method, you are required to use SVM for predicting the values (regression) or the class label (classification) for the input x.

Method 4: Neural Network for classification or regression (20 points)

Also for the Boston housing problem, you can design a neural network to predict the values (regression) or the class label (classification) for the input x. A possible network could be:

1. Regression: linear outputs & sum-of-squares error
2. Binary classifications: logistic sigmoid outputs & cross-entropy error

Task 2: Handwritten Digits Recognition

In this problem set we use the MNIST data set to build a handwritten digits recognizer. The digits have been size-normalized and centered in a fixed-size image.

Method 1: Feedforward Neural Network (30 points)

In this assignment you will practice putting together a simple image classification pipeline, based on the Softmax and the fully-connected classifier. The goals of this assignment are as follows:

- understand the basic Image Classification pipeline and the data-driven approach (train/predict stages)
- understand the train/val/test splits and the use of validation data for hyperparameter tuning
- implement and apply a Softmax classifier
- implement and apply a Fully-connected neural network classifier
- understand the differences and tradeoffs between these classifiers
- implement various update rules used to optimize Neural Networks

Do something extra!

- Maybe you can experiment with a different loss function and regularization? **(BONUS: 10 points)**
- Or maybe you can experiment with different optimization algorithm (e.g., batch GD, mini-batch GD, SGD or other optimization algorithms) **(BONUS: 10 points)**

Method 2: Convolutional Neural Network (30 points)

In this assignment you will practice training Neural Networks and Convolutional Networks. The goals of this assignment are as follows:

- understand **Neural Networks** and how they are arranged in layered architectures
- implement **dropout** to regularize networks
- effectively **cross-validate** and find the best hyperparameters for Neural Network architecture
- understand the architecture of **Convolutional Neural Networks** and train gain experience with training these models on data

Do something extra!

- In the process of training your network, you should feel free to implement anything that you want to get better performance. You can modify the solver, implement additional layers, use different types of regularization, use an ensemble of models, or anything else that comes to mind. (**BONUS: 10 points)**

Task 3: NBA Prediction

This data contains 2004-2005 NBA and ABA stats for:

-Player regular season stats, Player regular season career totals

-Player playoff stats, Player playoff career totals

-Player all-star game stats

-Team regular season stats

-Complete draft history

-coaches_season.txt - nba coaching records by season

-coaches_career.txt - nba career coaching records

-Currently all of the regular season

Project task:

- Outlier detection on the players; find out who are the outstanding players. **(20 points)**
- Apply one classification model (or more for comparison) to model such task **(10 points)**
- Predict the game outcome. **(20 points)**

Hints: Strategies about data normalization may be needed.

Task 4: Bike-sharing Rental Prediction

In this assignment you will implement a machine learning model to:

- Predict bike rental count hourly or daily based on the environmental and seasonal settings. **(20 points)**
- Detect event and anomaly. Count of rented bikes are also correlated to some events in the town which easily are traceable via search engines. For instance, query like "2012-10-30 washington d.c." in Google returns related results to Hurricane Sandy. Some of the important events are identified in bike rental data. Therefore the data can be used for validation of anomaly or event detection algorithms as well. **(20 points)**

Data: You can access the data from Bike Sharing Dataset Data Set.

Task 5: Character Recognition

Optical character recognition, and the simpler digit recognition task, has been the focus of much ML research. We have two datasets on this topic (see attached file). The first tackles the more general OCR task, on a small vocabulary of words: (Note that the first letter of each word was removed, since these were capital letters that would make the task harder for you.)

Project suggestion:

Use a sequential model to exploit correlations between neighboring letters in the general OCR case to improve accuracy. (Since ZIP codes don't have such constraints between neighboring digits, HMMs will probably not help in the digit case.) **(30 points)**

Task6: Image Captioning

In this assignment you will implement recurrent networks, and apply them to image captioning on Microsoft COCO. **(40 points)**

The goals of this assignment are as follows:

  - Understand the architecture of *recurrent neural networks* (*RNNs*) and how they operate on sequences by sharing weights over time
  - Understand the difference between vanilla RNNs and Long-Short Term Memory (LSTM) RNNs
  - Understand how to sample from an RNN at test-time
  - Understand how to combine convolutional neural nets and recurrent nets to implement an image captioning system
- Hints: You can implement Vanilla RNNs or LSTMs to model the problem
- Do something extra! **(+20 points)**
  - Given the components of the assignment, try to do something cool. Maybe there is some way to generate images that we did not implement in the assignment?