CS 6364 Homework 4

September 14, 2019

Deadline for the first submission: **Sep-19-2019**.
All assignments **MUST** have your name, student ID, course name/number at the beginning of your documents.
Your homework **MUST** be submitted via Blackboard with file format and name convention as follows:

HW#_Name_writeup.**pdf** (for writing part)
HW#_Name_code.**zip** (for coding part)

If you have any questions, please contact me.

Q1 (Linear Regression): Use the python library (sklearn.linear_model) to train a linear regression model for the Boston housing dataset:

https://towardsdatascience.com/linear-regression-on-boston-housing-dataset-f409b7e4a155.

Split the dataset to a training set (70% samples) and a testing set (30% samples). Report the root mean squared errors (RMSE) on the training and testing sets.

Q2 (Linear Regression): Use the python library (sklearn.linear_model) to train a linear regression model for the Advertising dataset (the dataset "advertising-data.zip" is also available on the blackboard):

https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0.

Split the dataset to a training set (70% samples) and a testing set (30% samples). Report the root mean squared errors (RMSE) on the training and testing sets.

Q3 (Logistic Regression): Use the python library (sklearn.linear_model) to train a logistic regression model for the Titanic dataset:

https://blog.goodaudience.com/machine-learning-using-logistic-regression-in-python-with-code-ab3c7f5f3bed.

Split the dataset to a training set (80% samples) and a testing set (20% samples). Report the overall classification accuracies on the training and testing sets and report the precision, recall, and F-measure scores for each of the two classes on the training and testing sets.

Q4 (Logistic Regression): Use the python library (sklearn.linear_model) to train a logistic regression model for the Pima Indian Diabetes dataset:

https://blog.goodaudience.com/machine-learning-using-logistic-regression-in-python-with-code-ab3c7f5f3bed.

Split the dataset to a training set (80% samples) and a testing set (20% samples). Report the overall classification accuracies on the training and testing sets and report the precision, recall, and F-measure scores for each of the two classes on the training and testing sets.