Assignment 1 – CSC4007 Advanced Machine Learning

Version 2.0: 7th, Feb 2019

Note: This assignment amounts to 30% of the module mark.
Deadline: 24:00 Friday, 1st March 2019

Summary: In this assignment, you will use linear regression to solve a real-world problem that builds a model to predict house prices. Your tasks consist of:

- Using linear regression to fit a linear model to a given training data of house prices
- Using different features: linear, quadratic, cubic, radial basis function (RBF)
- Performing regularization to tackle over-fitting
- Using cross-validation to select hyper-parameters
- Interpreting, explaining, and reporting experiment results.

All codes are required to be written using Python on Jupyter Notebook. Only basic Python code, e.g. numpy and matplotlib, is allowed.

Dataset: This assignment uses a dataset that contains information collected by the U.S Census Service for houses in the area of Boston Mass (see http://lib.stat.cmu.edu/datasets/). The dataset contains 506 cases. For your reference, the file "*boston.data*" contains a dataset of 506 rows, 14 columns. The meaning of each variable in order is:

Column 1: CRIM per capita crime rate by town
Column 2: ZN proportion of residential land zoned for lots over 25,000 sq.ft.
Column 3: INDUS proportion of non-retail business acres per town
Column 4: CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
Column 5: NOX nitric oxides concentration (parts per 10 million)
Column 6: RM average number of rooms per dwelling
Column 7: AGE proportion of owner-occupied units built prior to 1940
Column 8: DIS weighted distances to five Boston employment centres
Column 9: RAD index of accessibility to radial highways
Column 10: TAX full-value property-tax rate per $10,000
Column 11: PTRATIO pupil-teacher ratio by town
Column 12: B background of neighborhoods
Column 13: LSTAT % lower status of the population
Column 14 (this is the output): MEDV Median value of owner-occupied homes in $1000's

Task 1 (18%): In this task, you will make use of the dataset given above to fit a linear model for house price prediction.

1.1 (2%): Load the data in file "*boston.data*" and divide this dataset into two sets: the training (80%) and the test (20%) sets.

1.2 (4%): Using linear regression (without regularization) with linear features. Fit a linear model for the house price data using only training data set, and report the accuracy of this model over both the training and testing sets.

1.3 (4%): Using linear regression (without regularization) with quadratic and simplified cubic features, fit two corresponding non-linear models for the house price data using only training data set, and report the accuracy of these models over both the training and testing sets. Interpret and explain these results when compared to the result in sub-task 1.2. Note that the quadratic (in a

complete form: consists of a bias term 1, all linear terms x_i , and all cross product terms $x_i x_j$) and cubic (in a simplified form: consists of all term in $\phi_{quadratic}$, and cubic terms x_i^3) features should be defined as:

$$\phi_{quadratic}(x)=[1, x_1,\ldots, x_d, x_1^2, x_1 x_2, x_1 x_3,\ldots, x_i x_j,\ldots, x_d^2]$$

$$\phi_{cubic}(x)=[1, x_1,\ldots, x_d, x_1^2, x_1 x_2, x_1 x_3,\ldots, x_i x_j,\ldots, x_d^2, x_1^3, x_2^3,\ldots, x_d^3]$$

note that: x_i here means the dimension *i* of an instance *x,* and *d=13* (the number of input dimensions/variables).

1.4 (4%): Using ridge regression with a regularization term $\lambda=0.5$. Repeat all sub-tasks 1.2 and 1.3 (replace linear regression by ridge regression), and report the same results. Give an explanation and interpretation of the results for the models fitted in 1.2, 1.3, 1.4, and in particular interpret and explain the difference on their accuracies.

1.5 (4%): Using 5-fold-cross-validation for ridge regression with $\lambda=2^k$ where $k=\{-15,-14,\ldots,40\}$ on training data. Repeat all tasks in 1.4 (ridge regression with linear, quadratic and cubic features). Report the best λ and the testing error of the associated optimum models (evaluate this best model using the testing dataset received in sub-task 1.1). Visualize the curves of training and validation error estimations (over 5 folds) together with theirs variances. Interpret and explain the difference of the results of the cross-validated accuracies for the models received in this sub-task and previous models in sub-tasks 1.2, 1.3, and 1.4.

Task 2 (12%): In this task, you will use ridge regression with radial basis function features (RBF) fit a non-linear model for house price prediction.

2.1 (4%): Using all input instances in the training set (created in 1.1) and assigning them as the set of centers $C=\{c_j\}_{j=1}^n$, where *n* is the number of instances in the training set. Construct the set of RBF features based on $C=\{c_j\}_{j=1}^n$ as:

$$\phi_j(x)=\exp\left(\frac{-\|x-c_j\|^2}{2\sigma^2}\right), \forall i=[1,2,\ldots,n]$$

where we denote $\|x-c_j\|$ as the length/magnitude of the vector $(x-c_j)$. Use ridge regression (set $\lambda=0.5$) with the above constructed RBF features (with $\sigma^2=1.0$) to fit a non-linear model on the training set, and report the accuracy on both the training and testing sets. In particular, interpret and explain this result in comparisons with the ones in sub-task 1.4.

2.2 (4%): Using 5-fold-cross-validation to select a better σ^2 on the range $\sigma^2=0.2*k$ where $k=1,2,\ldots,100$ (with a similar setting to sub-task 2.1). Report the best σ^2 and the testing error of its associated optimum model (evaluate this best model using the testing dataset received in sub-task 1.1). Visualize the curves of training and validation error estimations (over 5 folds) together with their variances. Discuss and explain the difference of the results of the cross-validated accuracies on both training and testing data for the models received in this sub-task and the one in 2.1.

2.3 (4%): Randomly select 100 instances in the training set to assign to centers $C=\{c_j\}_{j=1}^{100}$. Repeat the steps in sub-task 2.2, and discuss and explain the difference of the results of the cross-validated accuracies (the model on the best σ^2) on both training and testing data for the models received in this sub-task and the one in 2.2. Visualize the curves of training and validation error estimations (over 5 folds) together with their variances.

Assessment Criteria

The report will be weighted 70%. The evaluation of the report will be based on its quality, clarity, accuracy, and completeness. Use your own words to explain and interpret your methods used for implementation and the results received in each sub-task. You should also demonstrate your understanding and reasoning in every sub-task. For example, what are the results you receive? what is the meaning and interpretation of each result? Why and what are the results different from sub-task to sub-task (as asked explicitly in some sub-task)?

Code (submitted in a jupyter notebook file) will be weighted 30%. Your code should be efficient, concise and has good structure and organization, e.g. try to use functions for repetitive codes, avoid hand-coded settings for variables, etc.. Use comments and markdown cells to explain your code's functionality (can be concise like our code samples for practical labs). All vector and matrix operations must use Numpy. All visualization tasks must use Matplotlib. You are not allowed to use black-box or off-the-self machine learning libraries to solve these tasks.

Guidelines and Submission

Submission including your code (jupyter notebook file) and report must be submitted online, using the QOL webpage of the Advanced Machine Learning module, by **24:00 Friday, 1st March 2019**.

Note that your report should not contain details of code implementation, use code comments and markdown cells in Jupyter Notebook for code explanation. The report should have information (student name, student number, time, data, and assignment number) clearly written in the first page.

All files must be in a ZIP file called "Assignment1_StudentNumber.zip". It's noted that your submission is your own work and you are aware of the University policies regarding plagiarism and collusion.

Late submissions will be treated according to standard university penalties.

Summary of change to version 1.0:
- Correcting the list of k in task 1.5 to: $k = \{-15, -14, \ldots, 40\}$