|  |
| --- |
| **SNU ECE.** Machine Learning Due : 2017.06.13  **Final Project**  Lecturer : Kyomin Jung |

**1. House Prices Prediction (Coding Problem, 100p)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Instructions (please read carefully)**   * This assignment should be done with your teammate * **Hardcopy** (1 printout or manuscript) : submit to TAs with your solutions   + Your write-up on solutions/explanations.   + DO NOT attach your code to the write-up * **Softcopy** (1 zipped file) : submit to ETL   + Your source code and results   + The zipped file should be named as ‘teamID\_project.zip’ (e.g. 1\_project.zip)   + The zipped file should contain two folders(code, data)   + The code folder should contain 2 folders(regression, classification) and each folder should contain executable files (+ extra compiled files)   + Your program should be coded with one of C/C++(gcc,g++ version 4.8.4)/MATLAB/Python2.7   + Code of each languages should run with just one command as follows. We will examine and score your code in  |  |  | | --- | --- | | Language | Command(e.g. regression) | | Python | Python regression.py | | C/C++ | ./regression |  * MATLAB code should have executable files named as ‘folder\_name.m’ and it should work with the ‘run’ button in the GUI. * Related outputs should be printed on the screen if possible, or capture as image files and archive to the zip file with your codes. * **We will run your code in Ubuntu 14.04 LTS** * **Due to the capacity of TAs, we will only check executable files with correct directory and specified command as above.** * **You will get 0 point if files have errors or not in correct directory.** * **Write down your executable file names at the top of the project report(hardcopy)**   + For example:   + A. regression 1 : regression.py   + A. regression 2 : regression\_opt.py * **Write down your environment in detail(libraries, packages, etc)** |

The given data is a part of Ames Housing Data, consisting of 80 variables directly related to the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property(e.g. When was it built? How big is the lot? How many square feet of living space is in the dwelling?)

The data consists of 20 continuous variables(include house price), 14 discrete variables, 23 nominal variables and 23 ordinal variables. **The minimum requirements** of the project is to use continuous, discrete, ordinal variables all together. You will get extra point if using nominal data additionally. Download the attached data(1260 train data) and design your algorithm. **Try to find out how to check your model performance without additional test data.**

We have another 200 test data for checking your result and **we will check whether you used the given partial data but not the whole dataset.** **You should describe how to run model with additional test data on your project report.**

**-TIP-**

**You have to consider how to properly treat the discrete, nominal, ordinal variables as input variables along with continuous variables** (e.g. for ordinal case, the most likely the sensible thing to do is assigning consecutive integers, beginning with 0 for example, to each of the successive categories of the ordinal variable. You would be assuming, in that case, the increments between levels of the categorical variables.

**A. [Price prediction]**

1 .Program your price prediction model that can predict the house price.

* For example, use linear regression model, neural networks, etc
* Use a proper loss function to train your model.
* Describe the procedure and results in detail(at least, you should evaluate your model with following metrics)

\* Bias :

\* Maximum Deviation :

\* Mean Absolute Deviation :

\* Mean Square Error :

\* Y = actual home price

\* = predicted value

2. Show which variables are comparatively more related to the house price

* For example, use dimensionality reduction algorithms
* Without those comparatively unrelated variables, does the model works better?

**B. [Price range classification]**

1. Program 2 classification models that can classify the house price into 2 classes (price<160000 and price160000).

* Freely develop features as many numbers as you want, and model complexity as your favor.
* Explain your algorithms and the reason for choosing it.
* What is in-sample error rate of your algorithm?

2. Show which variables are comparatively more related to the house price class.

3. Improve the model performance of one model.

* For example, use dimensionality reduction algorithm, Ensemble
* Explain what method you applied and why you chose that method.
* You can pick one of your model or make an ensemble of models.