Ex4: Numerical Attributes & Linear Regression, etc.

Tokyo Tech.
Intro. to Comp. & Data
Exercise&hw week4ex

- How to take care of numerical attributes, and the mixed case.
- Introduction to linear regression and a neural network type classification model.

1. Homework assignment #1.

please send one pdf file, e.g., "o4_19M12345.pdf" via email to
Suzukakedai: watanabe.o.aa-cd18s@ml.m.titech.ac.jp
Ookayama: watanabe.o.aa-cd18o@ml.m.titech.ac.jp
before week5lect *of each campus*

2. Some explanation on Weka.

*1 Weka is constructed and provided by the University of Waikato
*2 For materials, see http://tcs.c.titech.ac.jp/DataMining/index.html

1. Homework assignment #4: Task

Your task #1:
How to handle numerical attributes in the classification

(a) By using several example data sets (at least three sample balance.scale.arff, cmc.arff, hepatitis.arff), try various methods (including, Weka Preprocess: "Discritize", "NumericToNominal", etc.) to obtain several classification rules,

and compare these methods from the results.

* Use "Percentage split" with default 66%.

* You might want to use "MultilayerPerceptron" that can be chosen from the choice of classifier under "function." In this case, use it under the no hidden layer option (see also the next page).

Your task #2: Use the standard linear regression with nominal attributes

regression 回帰分析 = a process for obtaining a *"numerical estimator"*

(b) Use breast-Tumor.arff (see below for more specific instruction) to experience basic learning algorithms for computing numerical value estimators. Also investigate several ways to treat nominal attributes.

+ Use linear reguression methods, i.e., "LinearRegression" and "MultilayerPerceptron" (under the single layer option) to obtain numerical value estimators.

  * Use files breast-TumorR*k*-229.arff as training sets and breast-TumorR*k*-57.arff as test sets, where $k = 1,2,3$ (These files are given in the web page: http://tcs.c.titech.ac.jp/DataMining/index.html)

  * Execute learning algorithms under "Using test set" mode and re-evaluate the obtained estimators by using a test set.

  * For using "MultilayerPerceptoron", set the "hidden layer" option to 0, which means to create a single layer neural network.

Your task #2: Use standard learning algorithms for "numerical value estimators" with nominal attr. (Cont.)

(b) Use breast-Tumor.arff (see below for more specific instruction) to experience basic learning algorithms for computing numerical value estimators. Also ...

+ Use linear "LinearRegression" and "MultilayerPerceptoron" ..... to obtain numerical value estimators.

+ Try several ways to change nominal attributes to numerical ones.

Difference may be minor.

* Our learning algorithms can automatically take care of nominal attributes. But besides using this feature, try to use "NominalToBinary" filter ("Preprocess" → "choose" "filters" → "unsupervised" → "attribute" I also would like to encourage you to transform data sets by your own method (e.g., by using your own program).

+ Compare the obtained estimators.

* Again I encourage you to develop your own way to compare the performance of the obtained estimators.

page 6

1. Homework assignment #4: Report

submit via email *before* week5lect

Required items that you need to explain:

From Task #1 and Task #2:

(1) State the results of your comparison.

* I would omit specifying the required items one by one from this exercise because by now you can determine, I believe, from the previous assignments.

Optional:

(2) For Task #1, give some reasonable explanation why a better discretizing method varies depending on learning algorithms and/or data sets.

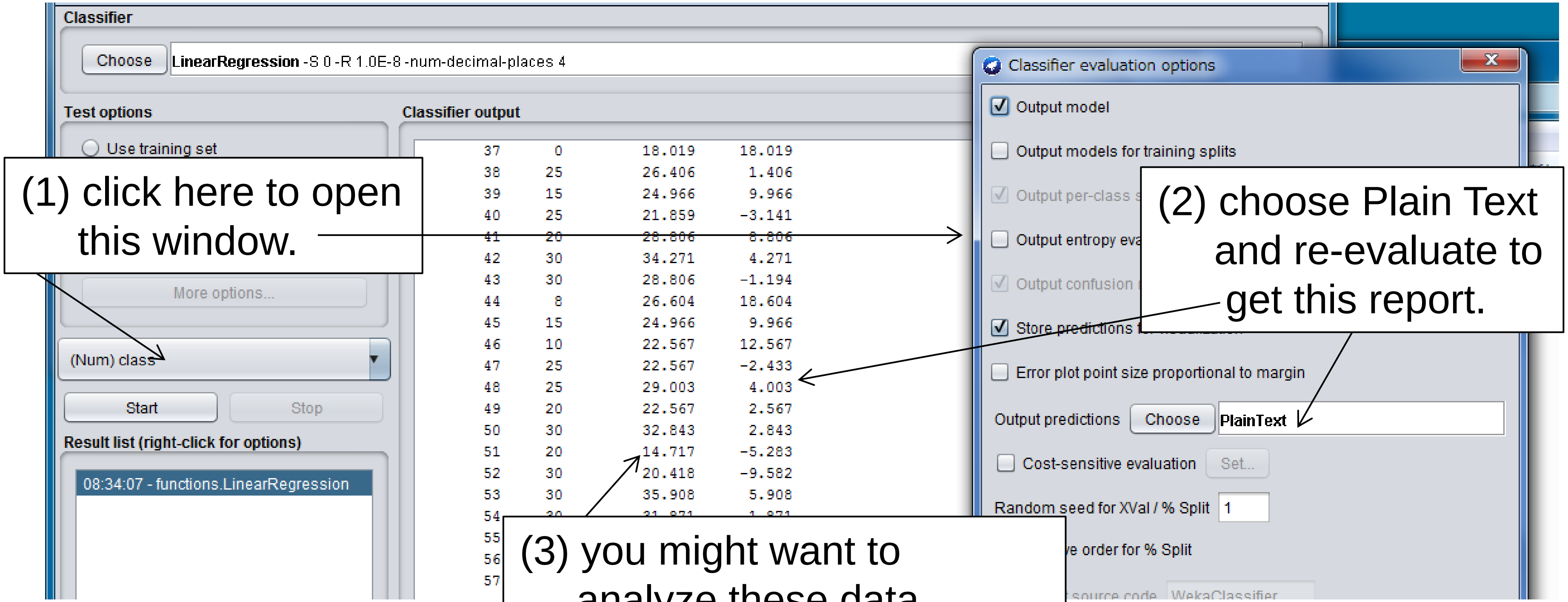
Honestly speaking,
I do not have a definite answer.

2. Tips for using Weka

demo. in the ex. session

How to see a detail report on your test:

- For getting a report on the result of the obtained numerical estimator on your test set, you can use "Output prediction" option when re-evaluating the estimator as follows.



Classifier

Choose LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4

Test options

○ Use training set

More options...

(Num) class

Start Stop

Result list (right-click for options)

08:34:07 - functions.LinearRegression

Classifier output

```
    37      0         18.019    18.019
    38     25         26.406     1.406
    39     15         24.966     9.966
    40     25         21.859    -3.141
    41     20         28.806     8.806
    42     30         34.271     4.271
    43     30         28.806    -1.194
    44      8         26.604    18.604
    45     15         24.966     9.966
    46     10         22.567    12.567
    47     25         22.567    -2.433
    48     25         29.003     4.003
    49     20         22.567     2.567
    50     30         32.843     2.843
    51     20         14.717    -5.283
    52     30         20.418    -9.582
    53     30         35.908     5.908
    54     30         31.871     1.871
    55
    56
    57
```

Classifier evaluation options

☑ Output model

□ Output models for training splits

☑ Output per-class s

□ Output entropy eva

☑ Output confusion

☑ Store predictions for visualization

□ Error plot point size proportional to margin

Output predictions Choose PlainText

□ Cost-sensitive evaluation Set...

Random seed for XVal / % Split 1

e order for % Split

source code WekaClassifier

(1) click here to open this window.

(2) choose Plain Text and re-evaluate to get this report.

(3) you might want to analyze these data by Excel or your program.