**CS2220 Introduction to Computational Biology**

**Assignment #1**

Please submit this assignment to Prof Wong Limsoon by 21/9/2017

This assignment contributes 10% to the final course grade

The objective of this assignment is to familiarize you with a basic gene feature recognition and prediction technique.

Data:

We work on the translation initiation site recognition problem from Pedersen & Nielsen (Proc ISMB 1997, pp 226-233). Please download the data from

<http://www.comp.nus.edu.sg/~wongls/courses/cs2220/2017/Assignment1.zip>.

Tool:

The WEKA machine learning package is used in this assignment. Please download and install it from <http://www.cs.waikato.ac.nz/ml/weka/>. Note that the default memory limit may be too small; please set it to 500MB.

Q1 [3 marks]: Describe in pseudo codes how the positive ATG segments in the file pos.fasta are extracted from the raw data. You can also provide an actual program instead of pseudo codes. The pseudo codes or program should handle the situation where an ATG segment does not have enough up-stream or down-stream flanking nucleotides.

Written in Java and the results are the same as the given pos.fasta for all the sequences.

Q2 [3 marks]: Describe in pseudo codes how the negative ATG segments in the file neg.fasta are extracted from the raw data. You can also provide an actual program instead of pseudo codes. The pseudo codes or program should handle the situation where an ATG segment does not have enough up-stream or down-stream flanking nucleotides.

Written in Java and the results are the same as the given neg.fasta for all the sequences.

Q3 [3 marks]: Given an ATG segment in pos.fasta or neg.fasta, how do you generate a feature vector based on frequencies on up-stream in-frame 3-grams and down-stream in-frame 3-grams. Please describe using pseudo codes. You can also provide an actual program instead of pseudo codes.

Written in Java, a feature vector is generated for every segment.

Q4 [2 marks]: Refer to the Inframe\_3\_Gram.arff file.

* What is the count of each up-stream/down-stream in-frame 3 gram in the first sample of the file?

Checked this up by clicking the edit… button in Weka. For the first sample:

upstream:

AAA:1,AAC:0,AAG:0,AAT:0,ACA:1,ACC:0,ACG:0,ACT:0,AGA:1,AGC:2,AGG:0,AGT:0,ATA:0,ATC:0,ATG:0,ATT:0,CAA:0,CAC:0,CAG:0,CAT:0,CCA:1,CCC:0,CCG:2,CCT:0,CGA:2,CGC:1,CGG:0,CGT:0,CTA:0,CTC:1,CTG:0,CTT:0,GAA:0,GAC:0,GAG:0,GAT:0,GCA:0,GCC:0,GCG:2,GCT:0,GGA:0,GGC:0,GGG:0,GGT:0,GTA:0,GTC:1,GTG:1,GTT:0,TAA:0,TAC:0,TAG:0,TAT:0,TCA:0,TCC:0,TCG:0,TCT:1,TGA:0,TGC:0,TGG:0,TGT:0,TTA:0,TTC:1,TTG:0,TTT:0

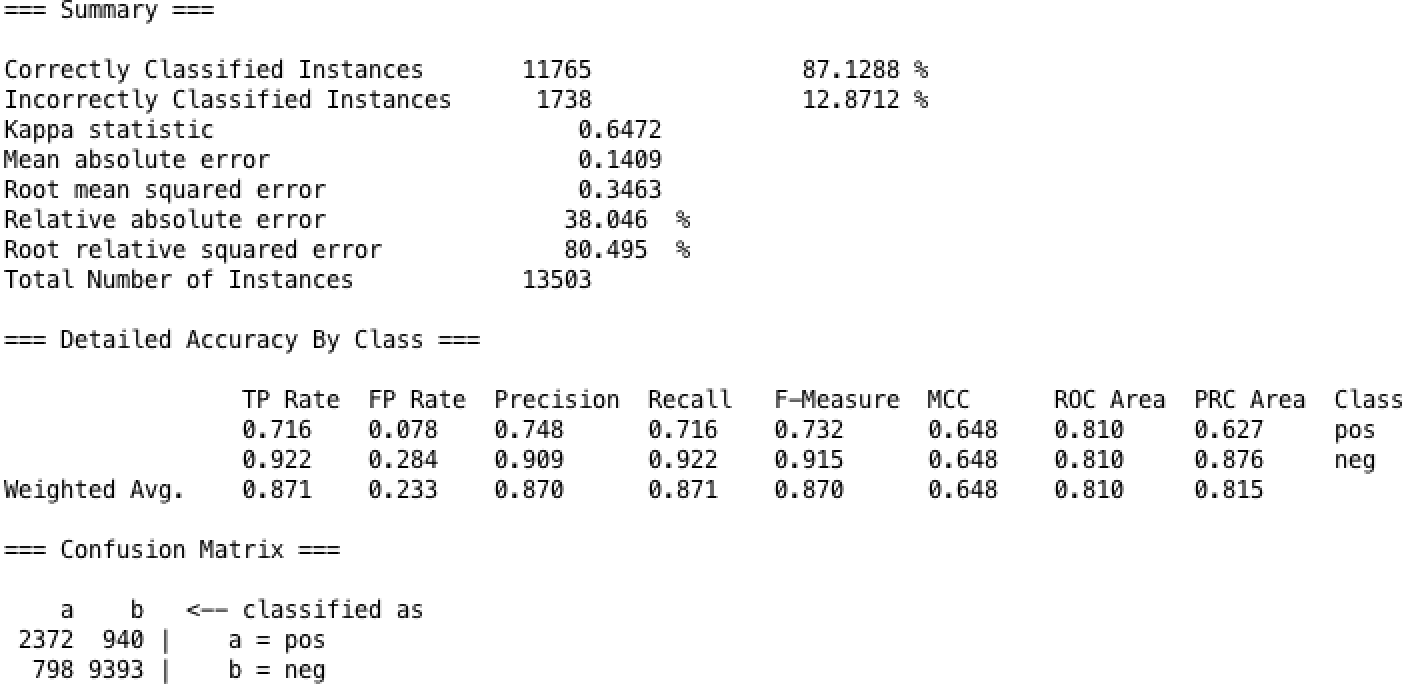
downstream:

AAA:1,AAC:0,AAG:3,AAT:2,ACA:0,ACC:0,ACG:0,ACT:0,AGA:1,AGC:0,AGG:1,AGT:1,ATA:0,ATC:0,ATG:1,ATT:1,CAA:0,CAC:1,CAG:0,CAT:0,CCA:0,CCC:0,CCG:0,CCT:1,CGA:0,CGC:0,CGG:0,CGT:0,CTA:1,CTC:1,CTG:0,CTT:0,GAA:2,GAC:1,GAG:1,GAT:3,GCA:2,GCC:0,GCG:0,GCT:0,GGA:2,GGC:0,GGG:0,GGT:0,GTA:0,GTC:0,GTG:0,GTT:0,TAA:0,TAC:0,TAG:0,TAT:1,TCA:1,TCC:0,TCG:0,TCT:0,TGA:0,TGC:1,TGG:0,TGT:0,TTA:0,TTC:0,TTG:2,TTT:2

* Is the first sample a true translation initiation site of the file?

Yes, it is. A ‘pos’ sign has been found for it.

Q5 [3 marks]: Build a C4.5 classifier for predicting translation initiation sites using the data in Inframe\_3\_Gram.arff file and the WEKA package.



* What is its 10-fold cross validation sensitivity, precision, and accuracy?

sensitivity:71.6%, precision: 74.8%, accuracy:87.1288%

* What is the size of decision tree of the classifier?

1037

* What does this implies?

1037 nodes, 519 leaves (decision rules), so there are 518 features used for classification. The tree is quite large relative to the total number of instances. For such a large tree, overfitting and fragmentation problem may occur. The decision rules generated may over fit this data set and lose some generalization capacity. Some rules may be locally reliable but globally insignificant, thus missing many globally significant rules and misleading the system. Some leaves are too small; the tree is too specialized relative to the size of the sample. If adjust the parameter to specify the lower bound for leaves- then the more generalized tree has an improved accuracy- the tree becomes more apt at generalization.

Q6 [2 marks]: Use the “chi-square” filter in WEKA to select the 10 most discriminative features based on the Inframe\_3\_Gram.arff file.

* List the features that are selected and their chi-square values.

evaluate on all training data

chi-square feature

1672.97447 15 INFRAME\_UPSTREAM\_ATG

1242.41924 121 INFRAME\_DOWNSTREAM\_TGA

984.44107 95 INFRAME\_DOWNSTREAM\_CTG

887.80022 99 INFRAME\_DOWNSTREAM\_GAG

878.70452 102 INFRAME\_DOWNSTREAM\_GCC

732.70099 98 INFRAME\_DOWNSTREAM\_GAC

714.42593 111 INFRAME\_DOWNSTREAM\_GTG

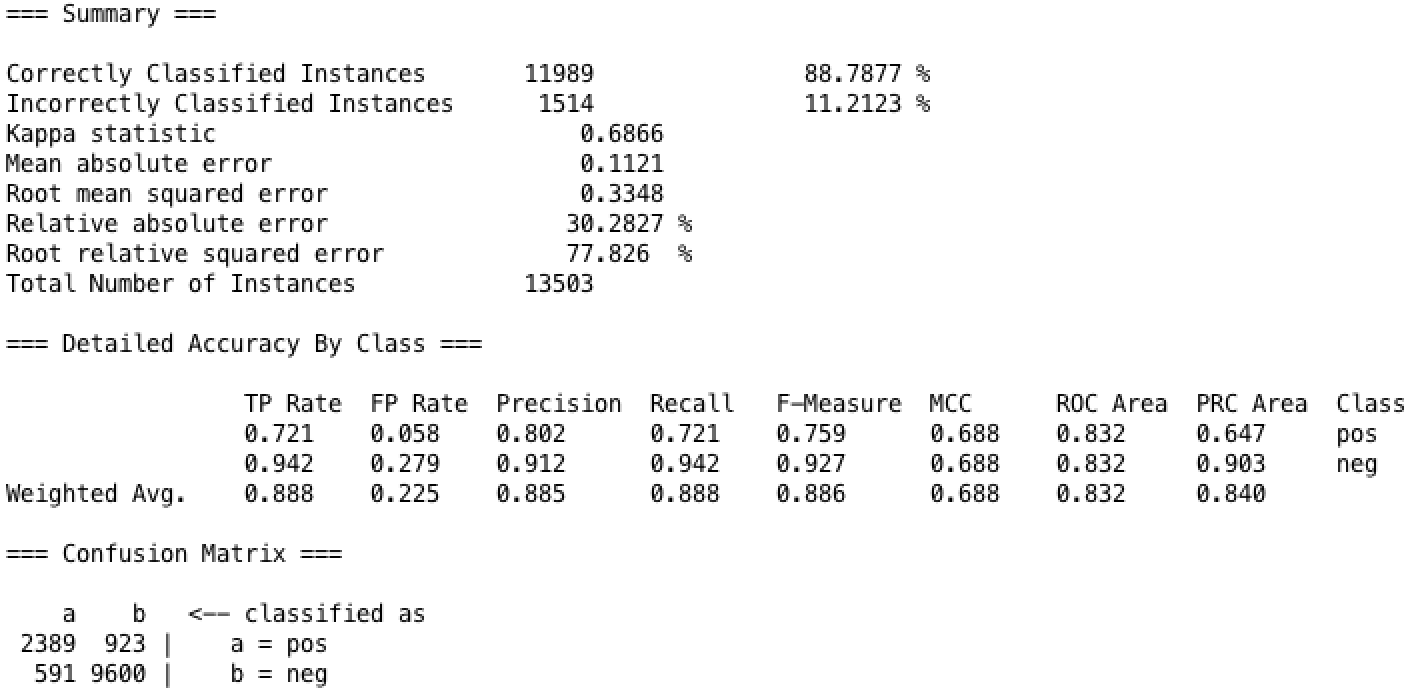
565.2399 100 INFRAME\_DOWNSTREAM\_GAT

556.80568 113 INFRAME\_DOWNSTREAM\_TAA

538.27158 115 INFRAME\_DOWNSTREAM\_TAG

Q7 [4 marks]: Build a SVM classifier for predicting translation initiation sites using the data in Inframe\_3\_Gram.arff file and the WEKA package. The classifier should consider only the top 10 “chi-square” features in the data. Be careful that the top 10 features should be selected fresh in each f­­­­old when you are doing 10-fold cross validation.

* What is its 10-fold cross validation sensitivity, precision, and accuracy?



sensitivity: 72.1% precision: 80.2%, accuracy: 88.7877%

* Show the WEKA 10-fold cross validation workflow diagram for this classifier.

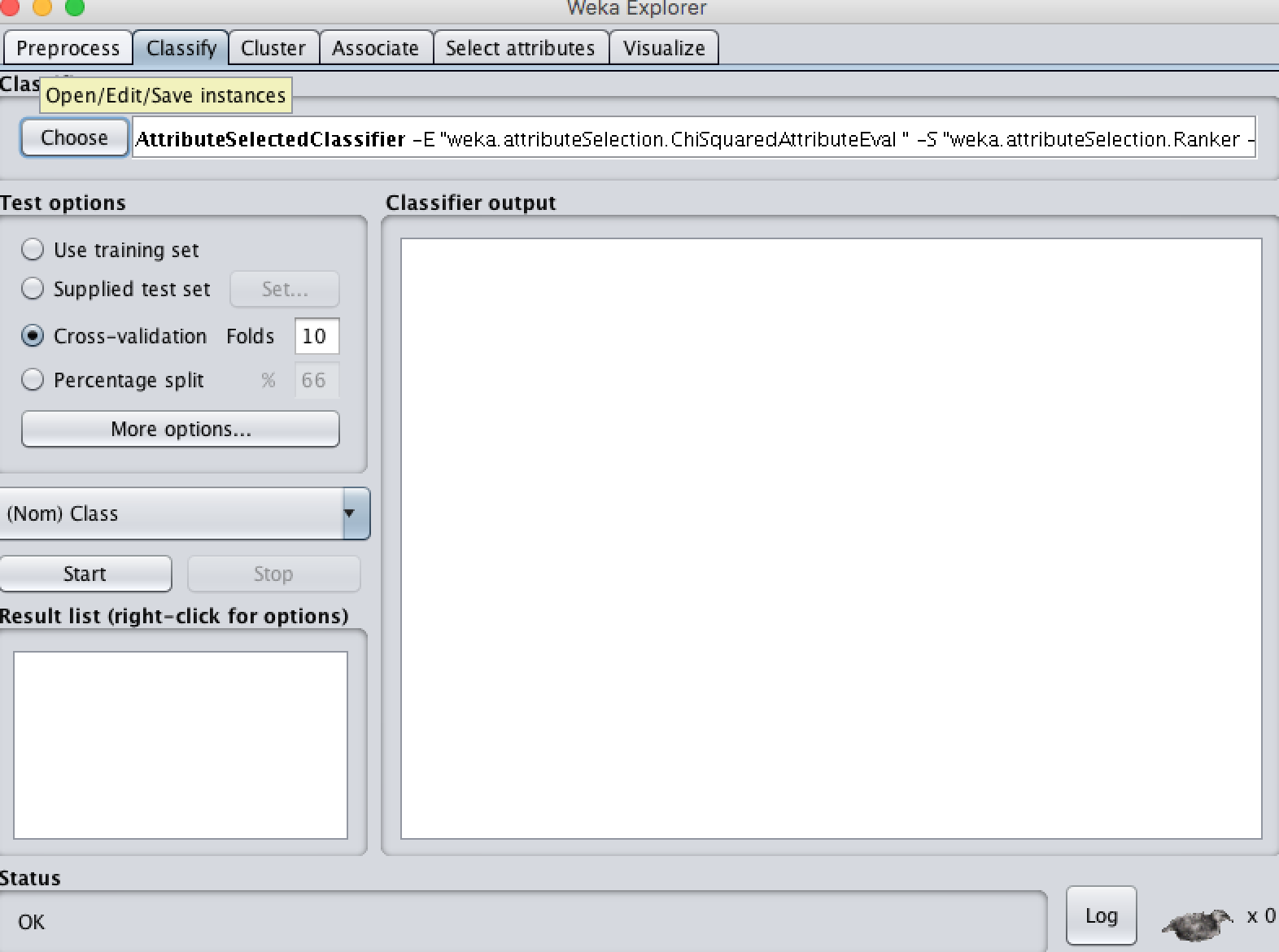
log:

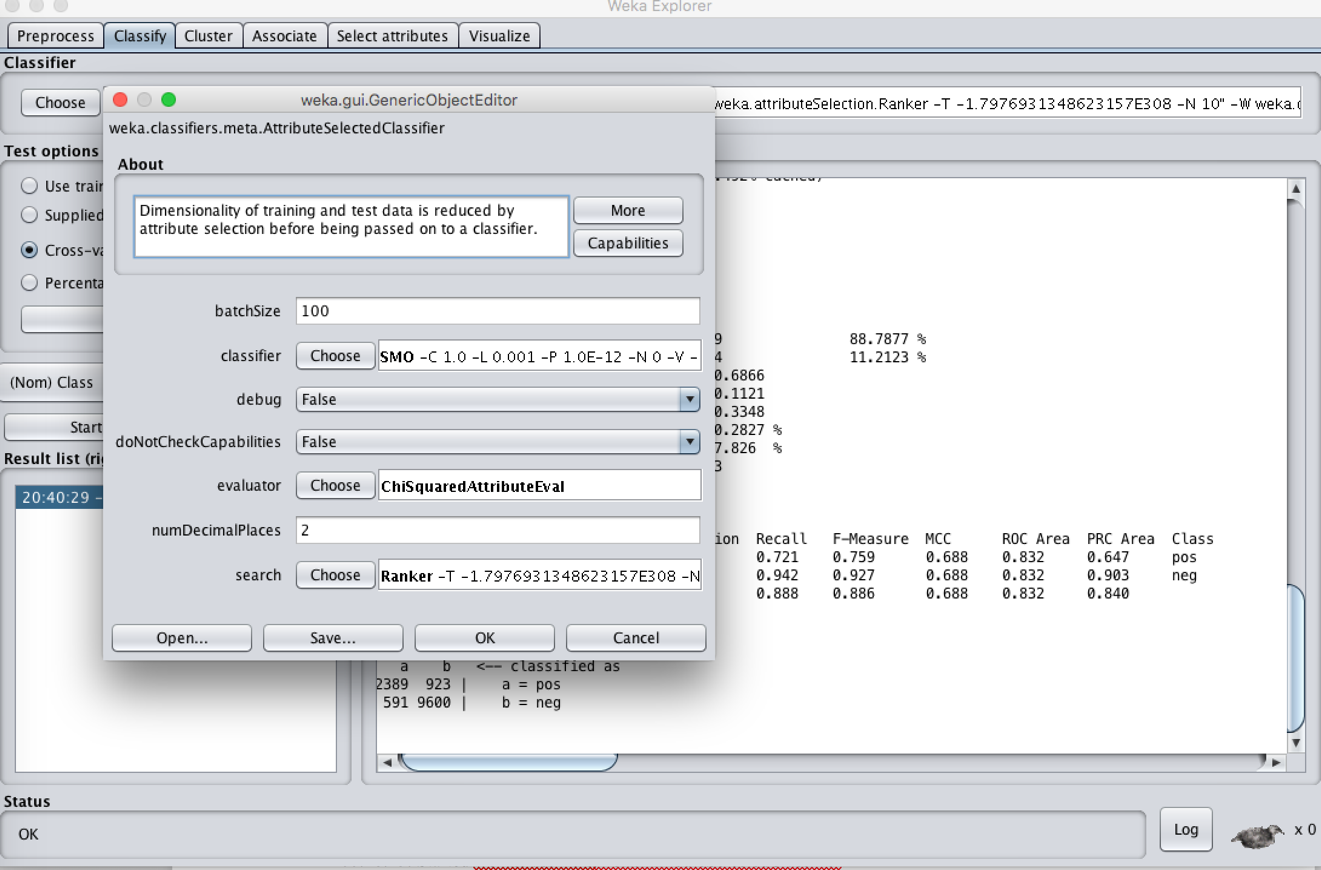
20:39:05: Base relation is now In-frame 3 gram, 3312 pos and 10191 neg (13503 instances)

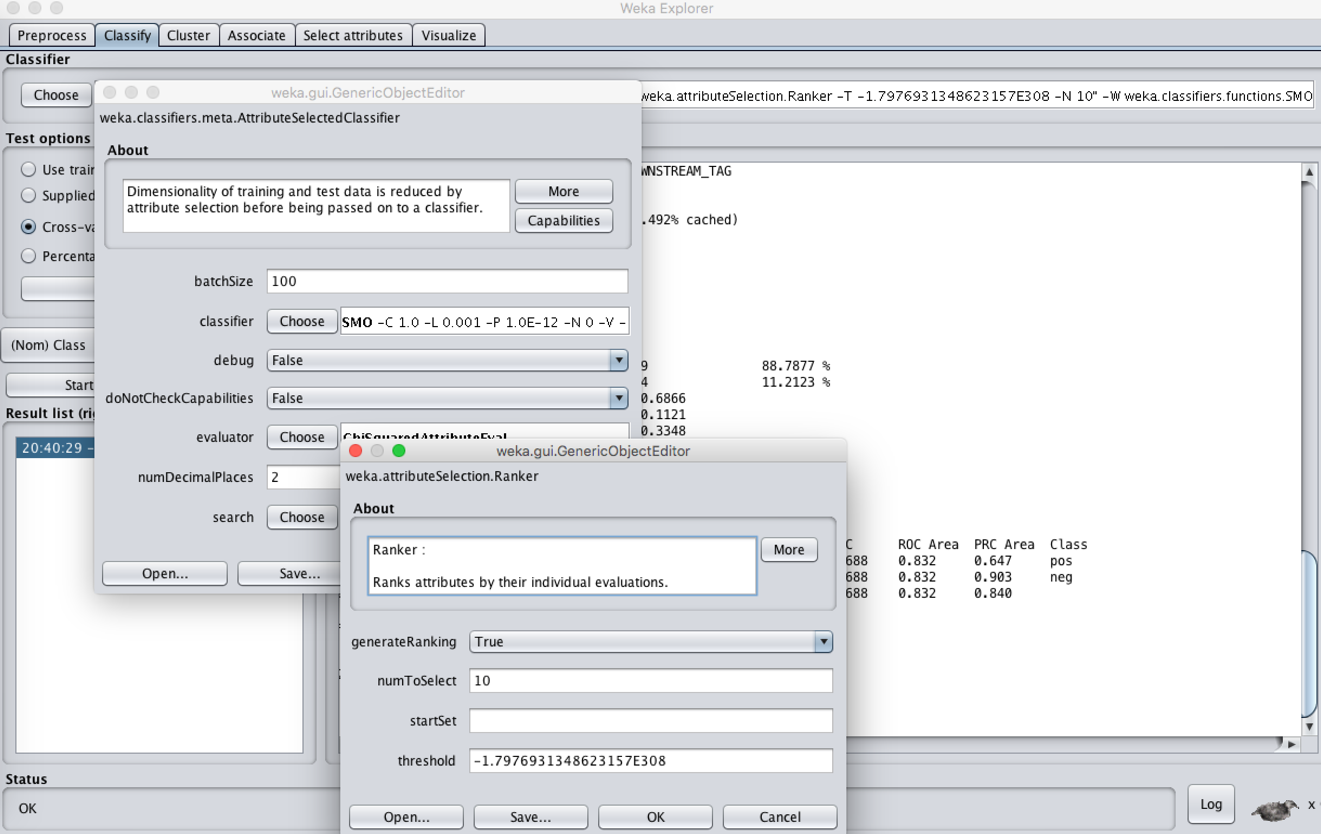
20:40:29: Started weka.classifiers.meta.AttributeSelectedClassifier

20:40:29: Command: weka.classifiers.meta.AttributeSelectedClassifier -E "weka.attributeSelection.ChiSquaredAttributeEval " -S "weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N 10" -W weka.classifiers.functions.SMO -- -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K "weka.classifiers.functions.supportVector.PolyKernel -E 1.0 -C 250007" -calibrator "weka.classifiers.functions.Logistic -R 1.0E-8 -M -1 -num-decimal-places 4"

20:40:44: Finished weka.classifiers.meta.AttributeSelectedClassifier







--------------------end of assignment #1--------------------