Introduction

The first assignment of this course is a competition. The challenge is to build a classification model for a given dataset, and all submitted predictions will be ranked (and graded) on achieved accuracy. You are given a labeled dataset on which you can try out several algorithms, and an unlabeled dataset for which you are asked to provide the labels (predictions) using your model.

Experimenting

The labeled dataset can be found in the attachments of this assignment (challenge-labeled.arff). Use it to preprocess the data, select algorithms, optimize parameters and build models, using the Weka Explorer or Experimenter. Note that this a rather large dataset, and some classifiers may require a lot of memory. Therefore, it is good to start Weka with additional memory, e.g., using 'java -Xmx1000M -jar weka.jar'.

Submitting your predictions

The unlabeled dataset can be found in the attachments of this assignment (challenge-unlabeled.arff) . When you have done all preprocessing and have selected your classifier and parameter settings, you should use the generated model to generate predictions for this unlabeled dataset. For instance, you can do the following:

- Use the Explorer to build a model on the labeled dataset. You will get performance results as usual.
- Left, under 'Supplied test set', set the unlabeled dataset.
- Under 'More Options', make sure that 'Output predictions' is checked. If you use Weka 3.7, select CSV as output format.
- In the result list (bottom left), right-click your model and choose 'Re-evaluate model on current test set'.

- In the output window (right), you will find the predictions under the header '=== Predictions on user test set ==='

Using this method, you will get an output that looks like this:

=== Predictions on test set ===

inst#	actual	predicted	error	probability distribution
1	?	2:+	+	0.104 *0.896
2	?	1:-	+	*0.741 0.259

These are the instance number, actual label (unknown), the prediction (pos or neg), the error (unknown) and the probability of each prediction. If you use WEKA 3.7, the output can be slightly different.

Finally, send in the entire prediction list.

Timing

- Deadline: November 19, 2023

What to hand in

You should hand in a predictions file named xxxxxxx-prediction.csv, where xxxxxxx is your student number (no leading 's'). Do not compress your file.

The prediction file should start with the header line 'inst# actual predicted error probability distribution'. The number of instances should be exactly equal to the number of records in the unlabeled dataset file. (If it is not, you are probably about to hand in a result on the training data.) Both comma and tab delimited prediction files are accepted. Please **make sure** that the submission is in the right format!

A small (one page) report describing the choices and steps you took to achieve the final predictions.

Both files can be submitted on brightspace.