JOHANNES KEPLER
UNIVERSITY LINZ

Exercise Part 2 of 3 - KV Computational Data Analytics (351.008)

Summer term 2020 - Prof. Dr. Johannes Fürnkranz, Florian Beck

---

Q&A Zoom Session:	**May 20, 10:15-11:45**
Deadline:	**May 31, 23:59**

The aim of the exercises is to gain practical experience in machine learning. For this we will use the machine learning frameworks `Weka`[1]. The exercises are split in three parts which will be uploaded when the corresponding lecture notes have also been uploaded. In total, up to 15 points are awarded for the successful completion of the exercises (usually two points per exercise), which are included in the overall grade with 15%. The tasks can be completed alone or in a small group (maximum three students, name and matriculation number must be visible in the submission). The submission should be in the form of a self-explanatory presentation (e.g. PDF, OpenOffice or PowerPoint) with a focus on interpretation or analysis. Results are only to be recorded to the extent that they are necessary to verify the statements. Further files (e.g. in TXT/ARFF/CVS) are not considered. The submissions are made in Moodle (one per group) and must be made by 23:59 on the deadline day at the latest.

4 Pre-processing: Discretization (2 P.)

Choose among the datasets one with at least five numeric attributes. Create a discretized version of the dataset by applying the filter `weka.filters.supervised.attribute.Discretize` with standard parameters. Use 10-fold cross-validation with the same random seed for all experiments.

1. To how many different bins are the previously numeric variables distributed?
2. Compare the accuracy of `J48` on the original and the discretized dataset.
3. The meta-classifier `FilteredClassifier` combines a pre-processing method with a classifier to a new classifier which executes the pre-processing during the training process. Compare the accuracy of the combination `Discretize` and `J48` on the original dataset with the previous results. How do you interpret the quality of the accuracies and the size of the learned trees?
4. To obtain ordered bins, change the configuration of `Discretize` by setting makeBinary to `True`. Repeat the experiments and compare the results.

[1] `https://waikato.github.io/weka-wiki/downloading_weka/`

5 Association Rule Learning (2 P.)

Open the dataset `weather.nominal` in `Weka`, go to the Associate tab and choose the `Apriori` algorithm[2]. Customize `Apriori` by setting `outputItemSets` to `True` and `numRules` to a sufficiently high number (e.g. 999999) to obtain all itemsets and rules in the output. For the first run of the `Apriori` algorithm, set `lowerBoundMinSupport` (minimum support) to 0.1 and `minMetric` (minimum confidence) to 1. The output first lists all frequent itemsets with their frequency and then all built associated rules sorted by confidence.

1. How many frequent k-tuples exist for each k? In this special case, when a minimal support of just one instance is needed, how do you interpret these sets with minimum and maximum k and their size?

2. Test different parameter combinations by increasing the support and decreasing the confidence. Compare the total number of itemsets and association rules and justify which configuration(s) is/are the most suitable. What are the best rules with respect to confidence, lift and leverage? Which three rules do you find the most interesting and surprising? Also list three rules that receive a high evaluation, but do not seem to be interesting.

3. Repeat the previous subtask with the dataset `supermarket` and collect again the best and/or most interesting rules you obtained. Compare and interpret the settings that you think are the most suitable for the two datasets.

[2]The implementation is slightly different from the algorithm presented in the lecture.