Methods in Computational Linguistics
Master of Science *Computational Linguistics*
Exercise: Classification

apl. Prof. Dr. Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart

January 17, 2020

In this exercise, you demonstrate that you understood the concepts of "classification" and "clustering", you perform some essential calculations related to classification tasks, and you apply *Weka* to run and evaluate automatic classifications.

Classification

1. Provide an example of a classification task in Computational Linguistics that is not listed in the handout from the lecture.

2. Justify the role of a similarity measure in automatic classification. What is the purpose of the measure as a classification parameter?

3. Many English words are ambiguous between verbs and nouns (such as *run, plan, fight*). Assume that your task is to automatically distinguish between English verbs and nouns in an English corpus. Which features might be useful?

4. Calculate the Euclidean distance for the English verb pair *cook–sleep* according to the feature vectors in the handout, and compare the distance value with that of *cook–bake*. Does the difference in distance correspond to your intuition?

5. In the following table you find four meanings of the German verb *abnehmen*, accompanied by the total frequencies of the senses in an annotated corpus, and co-occurrences of the annotated verb senses with five German nouns in their contexts.

   Calculate the most probable sense (1, 2, 3 or 4) for a specific corpus instance of the verb *abnehmen* in the context of the nouns *Diät 'diet'*, *Kilo 'kilo'* and *Reparatur 'repair'* (as provided by the **binary** values in the last row of the table). Use a *Naive Bayes Classifier* and *Maximum Likelihood Estimation*, as introduced in the lecture.

Verb	Meaning	Freq	*Buch 'book'*	*Diät 'diet'*	*Geschichte 'story'*	*Kilo 'kilo'*	*Reparatur 'repair'*
abnehmen$_1$	TAKE	1500	8	1	1	1	1
abnehmen$_2$	QUALITY	800	3	1	25	12	85
abnehmen$_3$	BELIEVE	200	28	8	110	1	3
abnehmen$_4$	WEIGHT	1000	18	69	2	95	1
abnehmen$_i$	?	n.a.	0	1	0	1	1

6. Calculate the three entropy values (0.940/0.811/0.048) for the decision tree splitting example (*Wind*) in the handout.

7. Describe *n-fold cross-validation* in your own words.

8. Perform pair-wise evaluation for the example classification of blue vs. red objects in the handout.

   (a) How many and which pairs exist in the correct (gold standard) and in the automatic classification? Assume that $\langle lemma_1, lemma_2 \rangle$ and $\langle lemma_2, lemma_1 \rangle$ represent the same pair.

   (b) Calculate precision, recall and f-score.

Classification and Clustering in Weka

Information on Weka is available from the web site `http://www.cs.waikato.ac.nz/ml/weka/`.

1. Access Weka in `/mount/studenten/weka/`.

2. Use an editor of your choice to check out the two `arff` files `weather.nominal.arff` and `verb.subcat.arff` in the directory `/mount/studenten/weka/data_ssiw`. What information do the `arff` files contain?

3. Use the latest Weka version in the directory `/mount/studenten/weka/`. Go into the respective sub-directory.

   Start Weka: `java -jar weka.jar`

4. Use the Weka-Tool `ArffViewer` to once more inspect the two `arff` files `weather.nominal.arff` and `verb.subcat.arff`. How many objects, how many attributes and how many and which classes are provided by the two `arff` files?

5. Use the Weka `Explorer` and `Classify/Cluster` to classify/cluster the objects in the files `weather.nominal.arff` and `verb.subcat.arff`: apply and compare

   (a) the algorithms `J48` (the Decision Tree Classifier) and `SimpleKMeans`, and

   (b) the evaluation method "Use training set" in comparison to "Cross-validation" (for `J48`) and "Classes to clusters" (for `SimpleKMeans`).

   Click on the command line next to `Choose` to check out further parameters.

   Answer the following questions:

   - Why are the "Use training set" results better than the "Cross-validation"/"Classes to clusters" results?
   - How many folds are reasonable for the data set?
   - What are the precision values for the classes in a selected J48 run?
   - For the classifier J48, check in the `Confusion Matrix` whether some classes are modelled explicitely well. In the same way, check the `Classes to Clusters` assignments for SimpleKMeans.
   - For the classifier J48, draw the best decision tree for the data set.