Assignment 1: Decision Trees & k-Nearest Neighbours

CHANGELOG: 4 NOV 2016, 30 NOV; 12 DEC 2016

Individual Home Assignment (Undergraduate Students in Language Technology)
SUBMISSION DEADLINE: SUNDAY 18 DECEMBER 2016, 23:59
Make sure that your answers are thorough and comprehensive.

Assignments' Deadlines
18 Dec 2016: Ass1 and Ass2
15 Jan 2017: Ass 3
24 Feb 2017: Final submission date for all assignments.

Learning objectives

In this assignment we will explore the use of decision trees and nearest neighbor classifiers to learn morphological classes. In this assignment you are going to:

- use Decision Trees and k-Nearest Neighbours algorithms as implemented in Weka;
- explore how attributes/features affect classification results.
- work with the morphology of English verbs;
- classify verb inflections;

Data: The English Past-Tense Dataset

In this assignment, you will use the English dataset that we have already used in class. The English dataset consists of 4330 verb lemmas. The class refers to the past tense information rule. The phonological representation covers the last three syllables. The list of verbs and their morphological representation is in past-tense.dat. The weka native file format of the dataset is in past-tense.arff.
Explicit links:
past-tense.dat < http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/past-tense.dat >
past-tense.arff < http://stp.lingfil.uu.se/~santinim/ml/2016/Datasets/english_past_tense.arff >

G tasks

Task 1: Theoretical question

Q1: Define the concept of entropy in your own words. Give an example. Briefly explain the difference between Information Gain and Gain Ratio.

Task 2:

Data Exploration
Download the dataset past-tense.arff on to your computer. Start weka and choose the Explorer Interface. Open the dataset in Weka (Preprocess --> Open File).
Once you understand how the data sets are composed, you should analyze the informativeness of the different features using information gain and gain ratio. To do this, choose *Select attributes* in Weka and then choose *InfoGainAttributeEval* as the

Attribute Evaluator (which also requires you to choose *Ranker as the Search Method*). This will give you a ranking of the features in terms of information gain and gain ratio.

Q2: Which features are most informative for the dataset? Try to explain why some features are more informative than others. Also try to explain any differences that you observe between results with information gain and gain ratio.

Decision trees

Induce a decision tree for predicting the past tense form of an English verb. Open the **Classify** tab and choose J48 in the folder **trees** as the classifier (J48 is Weka's implementation of the C4.5 algorithm). Build a decision tree for the data set and analyze its performance. Compare training error (Test option: **Training set**) to test error (Test option: Cross-validation) and see whether there are signs of overfitting.

Q3: How accurate are the decision tree classifiers for the data set? Look at overall accuracy as well as precision and recall for specific classes. How does training error relate to test error? Does the model overfit?

K-Nearest Neighbor

Use k-nearest neighbour to predict the past tense form of an English verb. Open the **Classify** tab and choose **IBk** in the folder **lazy** as the classifier. Apply the classifier to the data set and analyze its performance.
Compare training error (Test option: **Training set**) to test error (Test option: **Cross-validation**). Vary the number of neighbors used to predict the class (click on options next to the **Classifier** choice to change the value of the parameter KNN) and see how this affects training and test error.
One of the properties of (simple) nearest neighbor classification is that all features are given equal weight, which means that irrelevant features could hurt classification accuracy. Check whether you can improve accuracy by removing features. Compare the best accuracy to that obtained with decision trees.

Q4: How accurate are the nearest neighbor classifier for the data set? Look at overall accuracy as well as precision and recall for specific classes. What is the effect of varying the k parameter? Can you improve accuracy by removing less informative features? Does k-nearest neighbor perform better or worse than decision trees?

VG tasks

Decision trees

Compare tree induction with and without pruning. Click on the settings next to the Classifier choice to switch the parameter **unpruned** from false to true and see how this affects the size of the tree as well as the relation between training and test error.
Q5: What is the effect of pruning?

k-Nearest Neighbours

Q6: Can you force the nearest neighbour classifier to behave like the *pruned* decision tree on the English past tense data? How? What results do you get?

To be submitted

A written report (at least 2 pages) containing the **reasoned** answers to the tasks and questions above and a short section, that you can call "*Conclusions*", where you summarize your experience and your reflections.
Warning: Cutting and pasting Weka's results page into the report without commenting or explaining the whys and wherefores is not enough to get a pass on the assignment.

Submit the report in **PDF** format to santinim@stp.lingfil.uu.se no later than **18 December 2016, 23:59**. **Please, write this phrase in the subject line of your email: "ML4LT 2016 – Ass1: *your name*".** Attach any additional material that you think is important to fully understand your report.

Naming conventions

Please, name your pdf report in this way (it will be easier for me to organize and archive them): surname_name_*ass1report*.pdf (ex: *santini_marina_ ass1report.pdf*).

--the-end--