Statistical Methods in Natural Language Processing (NLP)

Class 11: Machine Learning: LDA, QDA, FDA, C5.0, Model Comparison

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Machine Learning

- C5.0
- Creating and Evaluating a Decision Tree

C5.0

1. The C5.0 is a classification algorithm developed by Ross Quinlan (Quinlan, 1993).
2. It assesses class factors, such as the dialect, based on a defined set of predictors.
3. It evaluates recursively the data and employs the predictors that can provide the best splitting of the data into more refined categories.
4. The splitting criterion is the difference in information entropy (a.k.a., the normalized information gain).

C5.0

Task Classify whether a fricative consonant is stressed or unstressed.

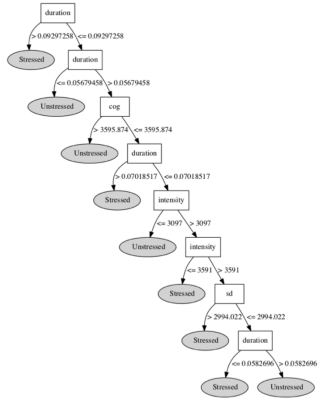
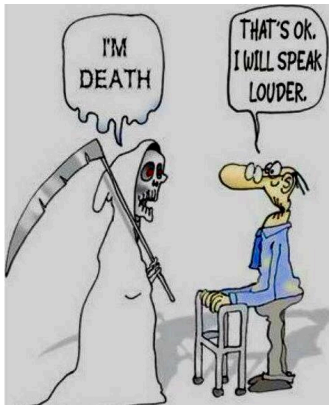


Figure 7: *Decision Tree produced by the machine learning and classification algorithm C5.0.*

C5.0

1. The predictor that provides the highest normalized information gain is the one selected for the decision (see also Woehrling et al., 2009, who provide classification a regional French varieties, using a different decision tree method).
2. Typically, each split is also an interpretation of the variation or impurity in the data.
3. The algorithm will stop when a criterion is met, such as when there are not enough data left to split.
4. Finally, C5.0 provides both tree and rule models (for an application of C4.5, which is an earlier iteration of C5.0, on accent classification, see Vieru et al. (2011) and for the classification of stressed and unstressed fricatives using C5.0, see Themistocleous et al. (2016)).

Classification of Fricatives



Fricatives Database

Next Class

- Markov Chains
- HMMs