Bibliographical Notes on Tree-structured classifiers.

Vittorio Castelli

The main source for the lecture on trees is Chapter 8, sections 1 to 4 inclusive, of the textbook [5].

CART is described in the lovely book by Breiman, Friedman, Olshen and Stone [1], which contains an extensive discussion of all the main points we touched during the lecture.

A machine-learning point of view is contained in Chapter 3 of Mitchell's book [8], which is a good reference for ID3 and C4.5.

Devroye, Györfi, and Lugosi devote Chapters 20 and 21 of their book [4] to the topic. As usual, they devote a substantial effort to consistency proofs for classifiers, and provide us with interesting examples. Unlike the previous references, they also consider trees constructed using only the observations and not their labels.

The material described in class came from two papers [3, 6].

The problem of matching impurity functions to loss functions was addressed by P.A. Chou in 1991 [3]. From this work we derived the analysis of the selection of optimal labels for nodes and leaves, and the theory of how to optimally choose splitting points. Chou's paper also describes an algorithm that performs optimal (greedy) splitting in linear time (in the number of samples, in the number of classes and in the number of dimensions).

The description of pruning methods provided during the lecture comes essentially from a paper by Esposito, et al.,[6]. Other pruning methods are also described in this work and in [7].

Two interesting interesting conclusions were reached in [6]: the use of a pruning set is often a bad idea, and some common methods exhibit a marked tendency to overprune/underprune, which one would not easily infer from their description.

References

- L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth & Brooks/Cole, 1984.
- [2] Carla E. Brodley. Automatic selection of split criterion during tree growing based on node location. In *International Conference on Machine Learning*, pages 73–80, 1995.
- [3] P.A. Chou. Optimal partitioning for classification and regression trees. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(4):340–354, April 1991.
- [4] L. Devroye, L. Györfi, and G. Lugosi. A probabilistic theory of pattern recognition. Springer, New York, 1996.
- [5] R.O. Duda, P.E. Hart, and D.G. Sytork. *Pattern Classification*. John Wiley & Sons, second edition, 2001.
- [6] F. Esposito, D. Malerba, and G. Semeraro. A comparative analysis of methods for pruning decision trees. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(5):476–491, May 1997.
- [7] D. Malerba, F. Esposito, and G. Semeraro. A further comparison of simplification methods for decision-tree induction, 1996.
- [8] T.M. Mitchell. Machine Learning. McGraw-Hill, Boston, MA, 1997.