MA8701 Advanced methods in statistical inference and learning

Part 3: Ensembles. L13: Bagging - trees - random forests

Mette Langaas

2/21/23

Table of contents

Course homepage: https://wiki.math.ntnu.no/ma8701/2023v/start

1 Ensembles - first act

1.1 Outline

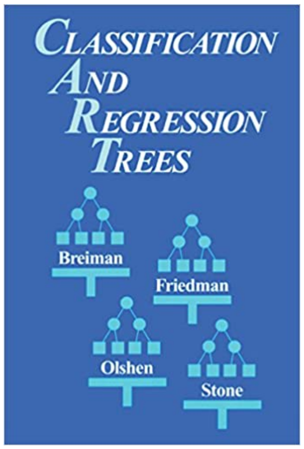
- Trees
- Many trees with bootstrap aggregation
- Many trees into a random forest
- Conclusions

1.2 Literature

- [ESL] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics, 2009) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Ebook. Chapter 8.7 (bagging), 9.2 (trees), 15 (random forest, not 15.3.3 and 15.4.3).

2 Trees

(ESL 9.2)



2.1 Leo Breiman

the inventor of CART, bagging and random forests

Quotation from Wikipedia

Leo Breiman (January 27, 1928 – July 5, 2005) was a distinguished statistician at the University of California, Berkeley. He was the recipient of numerous honors and awards, and was a member of the United States National Academy of Science.

Breiman's work helped to bridge the gap between statistics and computer science, particularly in the field of machine learning. His most important contributions were his work on classification and regression trees and ensembles of trees fit to bootstrap samples. Bootstrap aggregation was given the name bagging by Breiman. Another of Breiman's ensemble approaches is the random forest.

From Breimans obituary

BERKELEY – Leo Breiman, professor emeritus of statistics at the University of California, Berkeley.

"It is trite to say so, but Leo Breiman was indeed a Renaissance man, and we shall miss him greatly," said Peter Bickel, professor of statistics and chairman this summer of UC Berkeley's statistics department.

Breiman retired in 1993, but as a Professor in the Graduate School, he continued to get substantial National Science Foundation grants and supervised three Ph.D. students. Bickel said that some of Breiman's best work was done after retirement.

"In particular," said Bickel, "he developed one of the most successful state-of-the-art classification programs, 'Random Forest.' This method was based on a series of new ideas that he developed in papers during the last seven years, and it is extensively used in government and industry."

Breiman's best known work is considered to be "Classification and Regression Trees," a work in collaboration with three other scholars that facilitates practical applications, such as the diagnosis of diseases, from a multitude of symptoms.

2.2 Main idea

(for regression or classification)

- Derive a set of decision rules for segmenting the predictor space into a number of regions.
- We classify a new observation into one of these regions by applying the derived decision rules.
- Then we typically use the mean (regression problems) or a majority vote (classification problems) of the training observations in this region as the prediction in the region.
- Key advantage: interpretability.

We will only allow *recursive binary partitions* of the predictor space, using some stopping criterion.

2.2.1 From regions in predictor space to decision tree

(ESL Figure 9.2)
