

Title: Bagging and regression trees in individual claims reserving
Author: Jan Janoušek

THE SUMMARY OF THE THESIS

The first chapter gives a summary exposition, including certain details, of the methodology known as “classification and regression trees” (CART), starting from certain general aspects of classification tasks and evaluation of their quality (including apparent error rate called “resubstitution estimate“ in the thesis, sample splitting, and k -fold cross-validation). The classification trees and methods of their construction (via impurities, like the Gini index or cross-entropy) are depicted in considerable detail (including the question when to stop splitting). The subsequent pruning, to attain the optimum cost complexity, is also given a comprehensive description. Somewhat abridged, the whole description is then repeated adapted to the regression (numeric response) framework. The final part (very briefly) introduces bootstrap aggregating aka bagging.

The second chapter then introduces and again describes in considerable detail the task of insurance claims reserving, and finally applies the methodology exposed in the first chapter to it. This application is not just straightforward, but some pertinent details are elaborated there; the result is then demonstrated on the existing data from actuarial practice.

THE OVERALL EVALUATION OF THE THESIS

Overall, this is a pretty good thesis, and it is hard to find any substantial objections here. My major objections would concern the overall “big picture“. Given the final objective of the thesis, the application in insurance claims reserving, the treatment of classification part is somewhat too much extensive – after all, the actuarial application and the specification (“assignment of the thesis“) uses merely regression trees, not classification ones. The extended treatment of the latter would not be a drawback per se, but it takes space that could be otherwise devoted to what is important in the present context: regression trees, and above all, bagging – the treatment of the latter is rather brief and lacks some fine aspects regarding tuning, especially the wonderful idea of using “out-of-bag“ samples to tune the whole procedure.

Nonetheless, as I understand the departmental criteria for the evaluation of master theses, the onus there is not on what is potentially missing, but on the quality of what actually is contained in the thesis. From this viewpoint, my objections expressed above may be rather irrelevant, and I cannot but assert that the assignment of the thesis, as formulated by its advisor, was fulfilled – despite certain shift of focus.

The specific contribution of the author lies in the application of the tree methodology in the insurance claims reserving context – but not only there: the author also indulges in formal mathematical language of exposition, which, while not much appreciated in much of the statistical literature, may be viewed in the present environment as a sort of an asset – as it, after all, exhibits an individuality of the author, and thus attests to his independent treatment of the topics he treats.

Finally, the style, mathematical culture and typesetting, and references to the literature in the thesis are all flawless – here I could not find any shortcoming.

CONCLUSION

Once again, this is pretty good thesis, and despite certain aspects I point out above, I recommend it to be accepted as a master thesis – with rather a higher grade, which I will communicate to the committee chair orally; it should also depend on the candidate’s ability to answer the questions below.

OTÁZKY

1. I believe there is no application of classification in Chapter 2 of your thesis, and it is only regression trees that are mentioned in its assignment. However: would you know about any other application of classification trees, in actuarial science, or more generally, in finance?
2. Are you aware of how the “out-of-bag“ samples can be used to tune bagging (rather than sample splitting or cross-validation) and would you be able to briefly explain the idea?
3. In the conclusion of the thesis (page 60), you mention that all the methods you used are “overestimating the reserves“. Would you have any opinion why is it so?