Faculté des sciences médicales et paramédicales
Aix-Marseille Université

SESSTIM-Sciences Economiques et Sociales de la Santé &
Traitement de l'Information Médicale
Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université

Master of Public Health, specialization: AI4PH

TU MET-MALE: Methods of machine learning

Case study: Prostate cancer

The treatment of prostate cancer varies depending on the condition of the lymph nodes surrounding the prostate. In order to avoid a major surgical operation that consists of opening the abdominal cavity, doctors can make a preliminary assessment of the state of the lymph nodes according to some explanatory variables. In this application you will predict the binary outcome Y whether the cancer has reached the lymphatic network or not.

- `Y`: Y = 0 if the cancer has not reached the lymphatic network; Y = 1 if the cancer has reached the lymphatic network
- `age`: age of the patient at the time of diagnosis
- `acid` : acid phosphatase level in serum
- `rayx`: result of a ray analysis (0= negative, 1=positive)
- `size`: size of the tumor (0=small, 1=large)
- `grade`: pathological state of the tumor determined by biopsy (0=medium, 1=severe)

Questions

1. Load the data from `healthdata` package.
   (link: https://github.com/ielbadisy/healthdata).
2. Perform a descriptive analysis of the `prostatecancer` dataset.
3. First consider your outcome Y as a factor variable and build a classification tree with all the predictors.
4. Because we trained our tree on all the data, we cannot properly evaluate its predictive power. Now divide the dataset into train (70%) and test (30%). Fix the seed to ensure reproducibility.
5. Re-train your tree using only the training set and compute the test error.
6. Briefly describe the cross-validation procedure.
7. Use cross-validation to find the most optimal tree in terms of complexity through cost complexity pruning.
8. Plot the CV misclassifications as a function of size. What is the optimal number of nodes?
9. Now using the optimal node tree number, and prune the tree. Plot and interpret this smaller tree.
10. How well does this pruned tree perform on the test set? Compute the misclassification error.
11. What is the utility of the pruning using the CV process? Could it reduce the misclassification error?
12. Fit a Random Forest for the data with all the predictors. You can keep the same train/test partition done in question 4.
13. What does the OBB mean?
14. Evaluate the accuracy of your RF model on the test set. Interpret your results.
15. What are the most important variables in your RF? Interpret you results.

© Imad EL BADISY, SESSTIM, Faculté des Sciences Médicales et Paramédicales, Aix-Marseille Université