PERFORMANCE PREDICTION CHALLENGE: FACT SHEET FORMAT (1 to 2 pages)

Title: Random Linear Matching Pursuit

Name, address, email: Nicolai Meinshausen, nicolai@stat.math.ethz.ch

Acronym of your best entry: ROMA

Reference:

none yet, unfortunately

Method:

- Preprocessing: none, except for NOVA (PCA)
- Feature selection feature selection is achieved automatically; no preprocessing with feature selection
- Classification
  - Generalized linear model (Binomial family); linear in the variables and all interaction terms between variables; forward selection of variables and interactions (somewhat similar to MARS), yet not the best candidate is chosen from all variables but the best in a randomly selected subset (in this regard being similar to Random Forests). An ensemble of these predictors was formed; The goals was to have a good classifier which is linear in the variables and interactions

- Model selection/hyperparameter selection
  Hyperparameter selection is not very important for this method; some tuning was done on on out-of-bag samples
- Performance prediction guess. (How did you compute the value in the . guess file). Cross-validation

Results: The reader should also know from reading the fact sheet what the strength of the method is. To that end, provide a comparison table in the following format:

Dataset	Our best entry					The challenge best entry				
	Test AUC	Test BER	BER guess	Guess error	Test score (rank)	Test AUC	Test BER	BER guess	Guess error	Test score (rank)
ADA	0.8190	0.1810	0.1590	0.0220	0.2029 (15)	0.9149	0.1723	0.1650	0.0073	0.0000
GINA	0.9442	0.0558	0.0534	0.0024	0.0578 (24)	0.9712	0.0288	0.0305	0.0017	0.0000
HIVA	0.7057	0.2943	0.2698	0.0245	0.3182 (18)	0.7671	0.2757	0.2692	0.0065	0.0000
NOVA	0.9542	0.0458	0.0506	0.0048	0.0502 (9)	0.9914	0.0445	0.0436	0.0009	0.0000
SYLVA	0.9935	0.0065	0.0053	0.0012	0.0076 (19)	0.9991	0.0061	0.0060	0.0001	0.0000
Overall	0.8833	0.1167	0.1076	0.0110	0.1274 (21)	0.8910	0.1090	0.1040	0.0079	0.0000

For the overall performance, provide the average test score (As) and in parentheses the average rank (Rk).

- easy interpretation of results as result is linear in variables and interactions; computationally attractive

Code: Implementation in R; code is to be made available later