

Applied inductive learning - Lecture 9

Louis Wehenkel

Department of Electrical Engineering and Computer Science
University of Liège

Montefiore - Liège - November 28, 2005

Find slides: <http://montefiore.ulg.ac.be/~lwh/AIA/>

Ensemble-based methods

The intuitive idea

Boosting

Stacking

Further reading

The intuitive ideas behind ensemble-based methods

- ▶ Perturb and combine
 - ▶ Bagging: perturbation of learning samples by bootstrap resampling
 - ▶ Randomized trees: perturbation of learning algorithm by introducing random choices
 - ▶ Various combinations...
- ▶ Boosting
 - ▶ Iteratively correct errors of previous learners
- ▶ Stacking
 - ▶ Learn to combine models produced by “orthogonal learners”

Boosting

- ▶ Boosting for regression: residual fitting
- ▶ Boosting for classification: ADABOOST
- ▶ Stacking: some comments
- ▶ Further reading

Boosting for regression.

Gradient boost (Friedman 1998)

- ▶ Let A be a *weak* regression algorithm, and $l_s = \{(x_i, y_i)\}_{i=1}^N$ a sample.
- ▶ Set $l_s^1 = \{x_i, y_i^1\}_{i=1}^N$ with $y_i^1 = y_i$ for all i ;
- ▶ For $t = 1, \dots, T$ do
 - ▶ $\hat{y}^t = A(l_s^t)$;
 - ▶ $l_s^{t+1} = \{x_i, y_i^{t+1}\}_{i=1}^N$ with $y_i^{t+1} = y_i^t - \hat{y}^t(x_i)$ for all i .
- ▶ Return $\hat{y}^T = \sum_{t=1}^T \hat{y}^t$.

NB. e.g. with regression trees of limited complexity (e.g. stumps).

NB. \exists Further refinements...

Boosting for classification.

ADABOOST.M1 (Freund and Schapire 1996)

- ▶ Let A be a *weak* classification algorithm able to handle weighted samples,

Let $ls = \{(x_i, y_i)\}_{i=1}^N$ be a sample with $y_i \in \{1, \dots, k\}$.

- ▶ Set $w_i^1 = 1/N$ for all i ;
- ▶ For $t = 1, \dots, T$ do
 - ▶ Construct a model: $\hat{y}^t = A(ls, w^t)$
 - ▶ Calculate error: $err_t = \sum_{i: \hat{y}^t(x_i) \neq y_i} w_i^t$
If $err_t > 1/2$ set $T = t - 1$ and abort loop;
 - ▶ Set $\beta_t = err_t / (1 - err_t)$
 - ▶ Update the weights by

$$w_i^{t+1} = \frac{w_i^t}{Z_t} \times \begin{cases} \beta_t & \text{if } \hat{y}^t(x_i) = y_i \\ 1 & \text{otherwise} \end{cases}$$

where Z_t is a normalizing constant (such that $\sum_i w_i^{t+1} = 1$).

- ▶ Return $\hat{y}^T = \arg \max_k \sum_{t: \hat{y}^t = k} \log \frac{1}{\beta_t}$.

Further refinements, variants...

Stacking

(for classification and regression)

- ▶ Let $l_s = \{(x_i, y_i)\}_{i=1}^N$ be a sample.
- ▶ Let $A^t, t = 0, \dots, T$ be $t + 1$ learning algorithms
- ▶ For $t = 1, \dots, T$ do
 - ▶ Construct a model: $\hat{y}^t = A^t(l_s)$
 - ▶ Calculate predictions: $y_i^t = \hat{y}^t(x_i)$
- ▶ Set $l_s^0 = \{(x_i^0, y_i)\}$ with $x_i^0 = (y_i^t)_{t=1}^T$.
- ▶ Return $\hat{y} = A^0(l_s^0)$

Further reading

NB. Read in the order suggested (pdf versions of papers will be provided on the AIA web page).

- ▶ P. Geurts, D. Ernst, L. Wehenkel
Extremely randomized trees
To appear in Machine Learning, 2005
- ▶ Y. Freund, R. E. Schapire
Experiments with a new boosting algorithm
Proc. of 13th ICML (Int. Conf. on Mach. Learnng), 1996
- ▶ J. Friedman, T. Hastie, R. Tibshirani
Additive logistic regression: a statistical view of boosting
The Annals of Statistics, Vol. 28, No. 2, pp. 337-407, 2000.