STATISTICAL LEARNING SYSTEMS

LECTURE 7: ENSEMBLE METHODS FOR CLASSIFICATION AND BACK TO REGRESSION

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Ensemble methods - algorithms and their regularization

From the earlier course on ML we know well what the algorithms of bagging and boosting are. We know equally well what the random forests of Breiman, perhaps the most fascinating idea of building an ensemble, are. (During the lecture we shall briefly recall these algorithms; exact form of the algorithm will be given only for real AdaBoost.)

The idea behind the real AdaBoost algorithm is for each individual classifier (in a two-class problem) to estimate posterior probabilities

$$\hat{p}(y = -1|\mathbf{x})$$

and

$$\hat{p}(y = 1|\mathbf{x}) = 1 - \hat{p}(y = -1|\mathbf{x}).$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Ensemble methods - algorithms and their regularization

The real AdaBoost algorithm has the form:

1. Set weights $w_i = \frac{1}{n}, i = 1, \ldots, n$.
2. For $c = 1, \ldots, C$:
   (a) using weights w_i, $i = 1, \ldots, n$, construct

$$\hat{p}_c(\mathbf{x}) \equiv \hat{p}_w(y = 1|\mathbf{x}) \in [0, 1];$$

   (b) substitute

$$f_c(\mathbf{x}) \longleftarrow \frac{1}{2} \ln \frac{\hat{p}_c(\mathbf{x})}{1 - \hat{p}_c(\mathbf{x})};$$

   (c) substitute

$$w_i \longleftarrow w_i \exp[-y_i f_c(x_i)], \quad i = 1, \ldots, n,$$

   and renormalize (so as $\sum_{i=1}^{n} w_i = 1$).
3. Give

$$\operatorname{sgn}[\sum_{c=1}^{C} f_c(\mathbf{x})].$$

Ensemble methods - algorithms and their regularization

A way to generalize AdaBoost to $g > 2$:

1. Let the training sample be of the form (and size ng):

$$((\mathbf{x}_i, 1), y_{i1}), ((\mathbf{x}_i, 2), y_{i2}), \ldots, ((\mathbf{x}_i, g), y_{ig}),$$

   $i = 1, \ldots, n$, where y_{ik} is a label of observation $(\mathbf{x}_i, k)$, $y_{ik} \in$ $\in \{-1, 1\}$ and $y_{ik} = 1$ if $(\mathbf{x}_i, k)$ is in class k, while $y_{ik} = -1$ if $(\mathbf{x}_i, k)$ does not belong to k.

2. Apply the real AdaBoost to the training sample to get

$$F : \mathcal{X} \times \{1, 2, \ldots, g\} \to R, \quad F(\mathbf{x}, k) = \sum_{c=1}^{C} f_c(\mathbf{x}, k).$$

3. Give

$$\underset{k}{\operatorname{argmax}}\, F(\mathbf{x}, k).$$

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Ensemble methods - algorithms and their regularization

We must have noticed by now that regularization is prevalent when it comes to statistical prediction (whether regression or classification).

Recently, we noticed that the SVMs include regularization, earlier we discussed regularized regression methods.

We should have noticed that pruning CART or MARS or MART is in fact their regularization too.

Last but not least, pruning the number of individual classifiers used in an ensemble is also a way to regularize the ensemble classifier.

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund

Ensemble methods - algorithms and their regularization

Regularization can often be seen from another angle, at least when algorithm's outcome has an additive form, e.g., when it is built recursively,

$$\hat{f}_m(\mathbf{x}) = \hat{f}_{m-1}(\mathbf{x}) + \alpha_m f_m(\mathbf{x}),$$

$m = 1, \ldots, M$; just as MARS, MART and, for that matter AdaBoost (whether discrete or real) are. Here, to get the algorithm in its regularized version, it suffices to include a regularization coefficient γ, $0 < \gamma \leqslant 1$:

$$\hat{f}_m(\mathbf{x}) = \hat{f}_{m-1}(\mathbf{x}) + \gamma\alpha_m f_m(\mathbf{x}).$$

Most interestingly, boosting needs regularization if it is to have the property of Bayes-risk consistency; see *Ann. Statist.* **32** (2004), 1–134 (in particular, papers by Lugosi and Vayatis and Jiang), Bühlmann and Hothorn (with discussion) in *Statistical Science* **22** (2007), 477–522; see also Rosset, Zhu and Hastie, Boosting as a regularized path to a maximum margin classifier, *Journal of Machine Learning Research* **5**, 941–973, 2004).

Back to regression

SVM for regression: Suppose that we want to estimate a linear regression function,

$$f(\mathbf{x}) = \mathbf{w} \cdot \mathbf{x} + b,$$

given a training sample $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$ and using the following criterion:

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} [|y_i - (\mathbf{x}_i \cdot \mathbf{w} + b)| - \varepsilon]_+ + \lambda \|\mathbf{w}\|^2,$$

where ε and λ are given positive constants.
The similarity to the problem of finding the best SVM for classification,

$$\min_{\mathbf{w},b} \sum_{i=1}^{n} [1 - y_i(\mathbf{x}_i \cdot \mathbf{w} + b)]_+ + \lambda \|\mathbf{w}\|^2,$$

is not only striking, but makes the two problems solved in practically the same way with only inner products in their respective solutions.
Accordingly, using a kernel trick, we readily obtain an SVM for regression functions of practically any nonlinear form.

Back to regression - kernel ridge and PLS regression models

For kernel ridge regression please see a separate handout.

Kernel PLS regression will be discussed following the lines of the paper by K.P. Bennet and M.J Embrechts,

http://homepages.rpi.edu/∼bennek/papers/KB-ME-PLS.pdf

PhD STUDIES
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI
UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY
The project is co-financed by the European Union within the framework of European Social Fund