An Introduction to Boosting
Brown Bag Seminar

Marcel Dettling

Institut für Datenanalyse und Prozessdesign

Zürcher Hochschule für Angewandte Wissenschaften

marcel.dettling@zhaw.ch

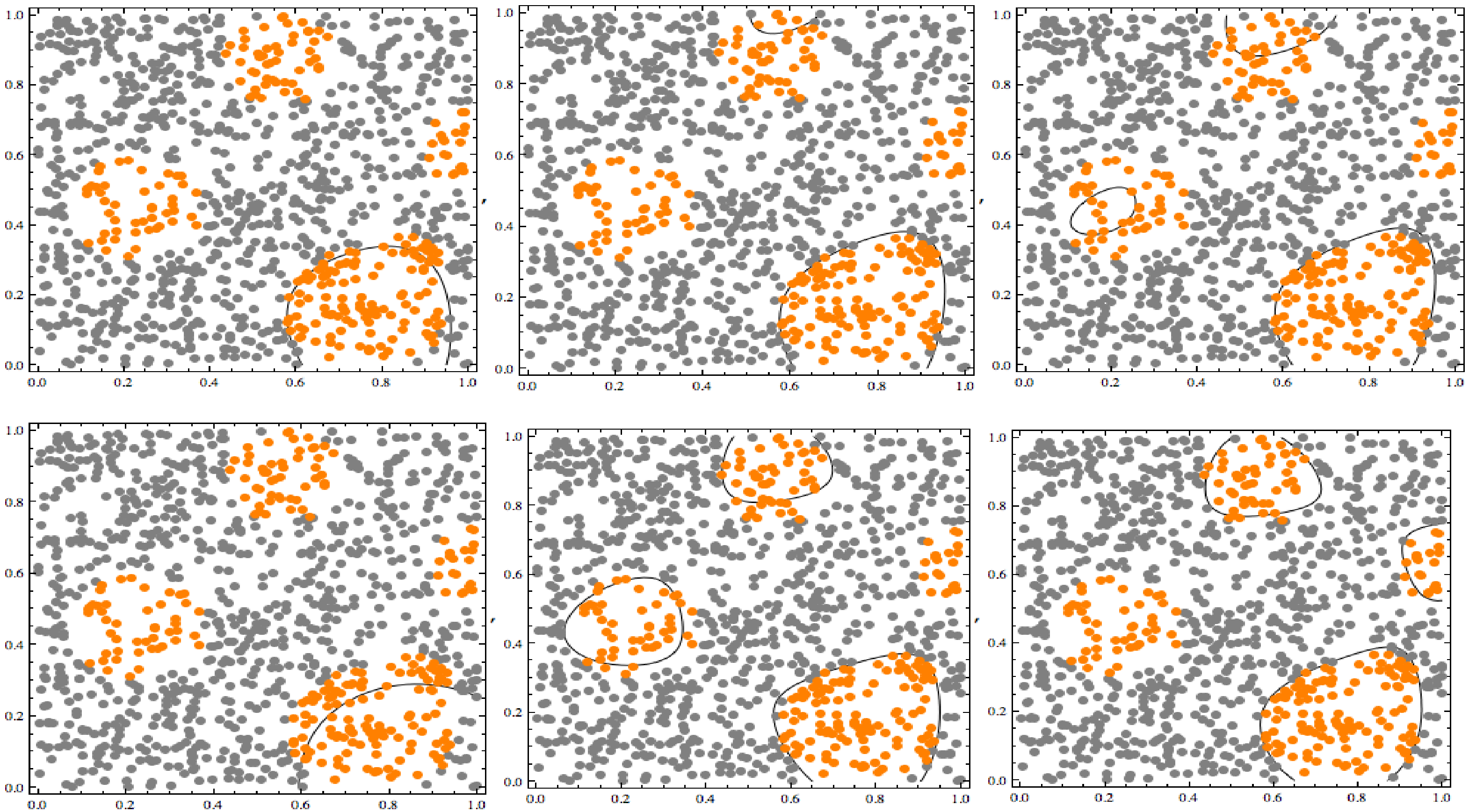
http://home.zhaw.ch/~dtli

Winterthur, 15. Januar 2014

An Introduction to Boosting
Brown Bag Seminar

Boosting…

A machine learning approach for supervised problems!



An Introduction to Boosting
Brown Bag Seminar

Setup

We are given data: $(x_1, y_1), \ldots, (x_n, y_n)$ i.i.d., with predictor variables $x_i \in \mathbb{R}^p$ and response $y_i \in \{0,1\}$.

Example: Prediction of response to a marketing campaign, based on all the information that is available about this customer.

Remarks:
- p is typically (very) large
- n can be large, but must not
- boosting is well suited if $p \gg n$

There are extensions of boosting to multiple class prediction where $y_i \in \{0,1,\ldots,J-1\}$, to regression with $y_i \in \mathbb{R}$, etc.

Goal

In our situation, we require a classifier function:

$$F(x) = \hat{y} \in \{-1,1\}, \text{ resp. } F(x) = \hat{y}^* \in \{0,1\}$$

or even better, an estimate of the conditional probability function:

$$F(x) = \hat{P}[y = 1 \mid X = x] \in [0,1]$$

In our example, this corresponds to the probability for positive response to the marketing campaign, given all the <u>relevant</u> properties of the customer. This potentially includes the identification and selection of the relevant features.

→ **Boosting can provide this with** $F_M(x) = \sum_{m=1}^{M} \alpha_m f_m(x)$

Historical View

Due to *Freund & Schapire, 1990-1997*, with AdaBoost: *Boosting is an ensemble method based on iterative data reweighting.*

Base Procedure (e.g. a classification tree):

$$(x_1, y_1), \ldots, (x_n, y_n) \longrightarrow \hat{f}(\cdot)$$

Boosting Steps

reweighted data $w_1; (x_1, y_1), \ldots, (x_n, y_n) \longrightarrow \hat{f}_1(\cdot)$

reweighted data … …

reweighted data $w_M; (x_1, y_1), \ldots, (x_n, y_n) \longrightarrow \hat{f}_M(\cdot)$

Aggregated Classifier

based on a weighted ensemble: $F_M(\cdot) = \sum_{m=1}^{M} \alpha_m f_m(\cdot)$

An Introduction to Boosting
Brown Bag Seminar

Iterative Reweighting Approach

Basic Idea:

- start with identical weights $w_i = 1/n$, fit a learner $f_1(\cdot)$ and evaluate its insample prediction performance for y_i
- depending on whether or how heavily an observation i was misclassified, increase its weight w_i. Hence the learner is forced to focus on the difficult-to-classify instances.
- the contribution of the learner $f_1(\cdot)$ to the final classifier $F(\cdot)$ is gauged by the averaging weight α_1. It is large if $f_1(\cdot)$ performed well, and small otherwise.

→ Repeat this process M times to obtain the solution $F_M(\cdot)$

An Introduction to Boosting
Brown Bag Seminar

AdaBoost

Algorithm:

1) Set $y_i \in \{-1,+1\}$ and start with identical weights $w_i = 1/n$
2) Repeat for $m = 1, 2, \ldots, M$:
   a) Fit the classifier $f_m(x) \in \{-1,+1\}$ using weights w_i
   b) Compute the weighted error $err_m = \sum_i w_i \cdot I[y_i \neq f_m(x_i)]$
   c) Compute the aggregation weight $\alpha_m = \log((1-err_m)/err_m)$
   d) Set $w_i \leftarrow w_i \cdot \exp(\alpha_m \cdot I[y_i \neq f_m(x_i)])$; normalize to $\sum_i w_i = 1$
3) Output $F_M(x) = sign \sum_{m=1}^{M} \alpha_m f_m(x)$

Statistical View of Boosting

Due to *Breiman (1999)* and *Friedman/Hastie/Tibshirani (2000)*:

All boosting algorithms fit a stagewise additive model of the form

$$F_M(x) = \sum_{m=1}^{M} \alpha_m f_m(x)$$

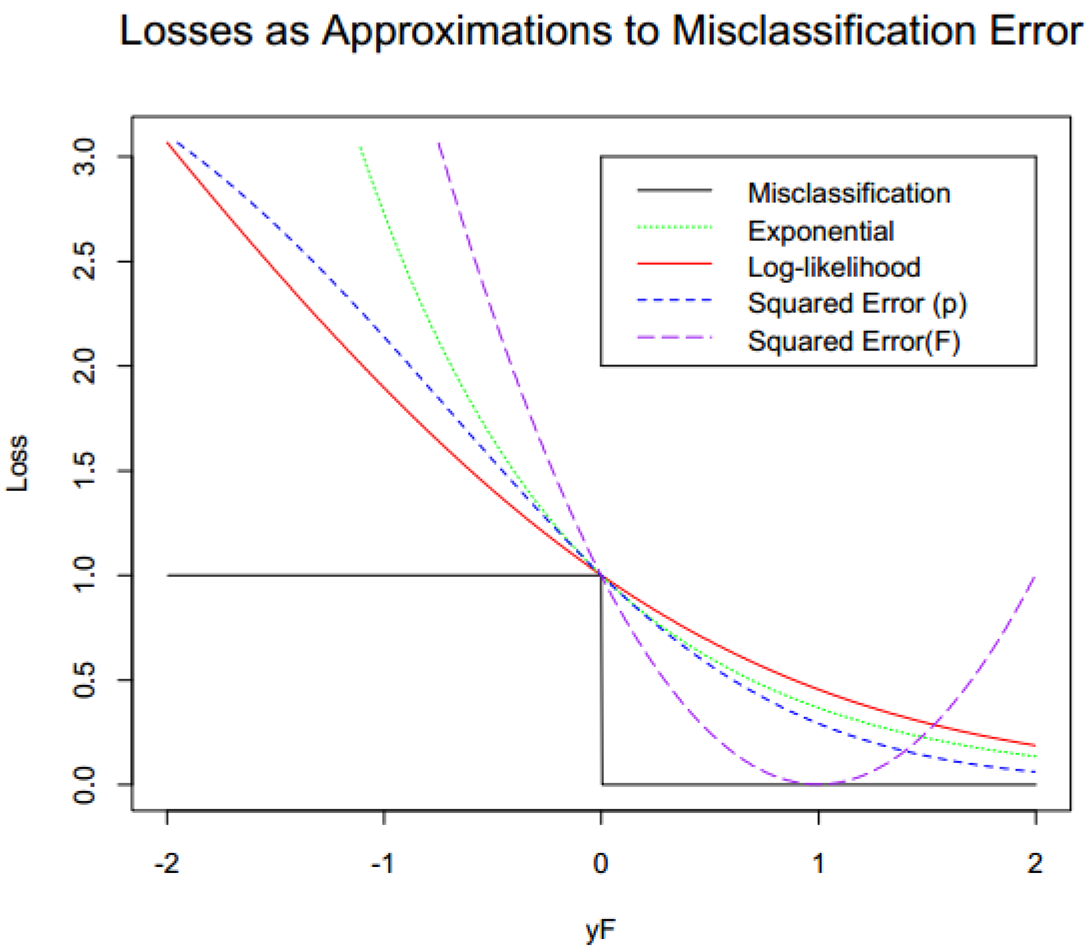
by steepest gradient descent minimization of a loss function, i.e.

$$F(x) = \arg\min_{f(\cdot)} E[L(y, f(x))]$$

The loss function is typically assumed to be differentiable and convex with respect to the second argument. This notion opens the door for novel, powerful boosting algorithms that are better and more flexible than AdaBoost.

An Introduction to Boosting
Brown Bag Seminar

Loss Functions



Misclassification

$$L(y,F) = I[y \neq sign(F)]$$

Exponential / AdaBoost

$$L(y,F) = \exp(-yF)]$$

LogLik / LogitBoost

$$L(y,F) = \log(1+\exp(-2yF))$$

Quadratic / L2-Boost

$$L(y,F) = (y-F)^2$$

Logit or Binomial Boosting

LogitBoost fits and additive logistic regression model by numerical optimization of the Bernoulli log-likelihood.

1) Set $y_i^* \in \{0,1\}$, $w_i = 1/n$, $F(x) = 0$ and $p(x) = 1/2$

2) Repeat for $m = 1, 2, \ldots, M$:

   a) Compute the working response and weights:

$$z_i = \frac{y_i^* - p(x_i)}{p(x_i)(1 - p(x_i))}; \;\; w_i = p(x_i)(1 - p(x_i))$$

   b) Fit $f_m(x)$ by weighted LS-regression of z_i on x_i with w_i

   c) Update $F \leftarrow F + f_m$ and $p \leftarrow \exp(F)/(\exp(F) + \exp(-F))$

3) Output $\hat{y}_i^* = sign(F(x_i))$ and/or $p(x_i)$