# Foundations of ML – Additional material for lesson 8

1. Weighted version of logistic regression loss

$$-\frac{1}{\sum_{i\in I} w_i}\sum_{i\in I} w_i\left(y_i \log h_\theta(x_i)+(1-y_i)\log\left(1-h_\theta(x_i)\right)\right). \qquad h_\theta(x) := \frac{1}{1+e^{-(\langle\beta,x\rangle+\beta_0)}}.$$

2. Boosting with stumps versus trees

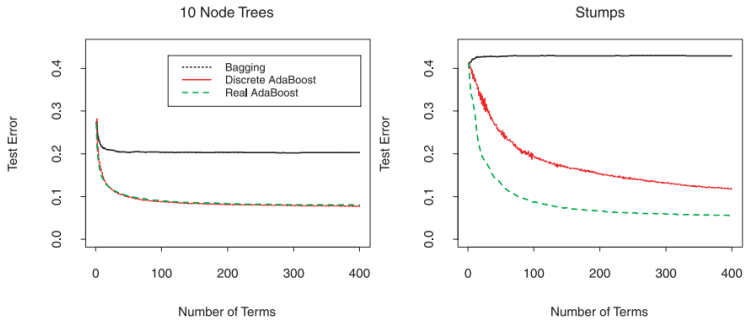
*The Annals of Statistics*
2000, Vol. 28, No. 2, 337–407

**SPECIAL INVITED PAPER**

**ADDITIVE LOGISTIC REGRESSION: A STATISTICAL VIEW OF BOOSTING**

BY JEROME FRIEDMAN,[1] TREVOR HASTIE[2,3] AND ROBERT TIBSHIRANI[2,4]

*Stanford University*




3. Wavelet-based Gradient Boosting

**Algorithm 1** Geometric Wavelets Gradient Boosting

1. Initialize $\hat{f}_0(x) = \underset{c}{argmin}\sum_{i=1}^{m} L(y_i, c)$.
2. set $\{y_i^0 = y_i\}_{i=1}^{m}$.
3. For $k = 1, 2, ..., K$
   (a) update the residuals $\left\{y_i^k = \hat{f}_{k-1}(x_i) - y_i\right\}_{i=1}^{m}$.
   (b) Choose randomly subset of $m'$ variables from the original data-set, denote by $\{x_i, y_i^k\}_{i=1}^{m'}$. based on this training set generate a tree $T(x) = \sum_j \psi_{\Omega_{k_j}}$, where $\left\{\psi_{\Omega_{k_j}}\right\}_j$ are the GW sorted by wavelet norm (see 13).
   (c) Denote the OOB subset by $OOB = \left\{(x,y) \mid (x,y) \notin \{x_i, y_i\}_{i=1}^{m'}\right\}$, then compute:

$$M_k = \underset{M}{argmin} \sum_{(x,y)\in OOB} L\left(y^k, \sum_{j=1}^{M} \psi_{\Omega_{k_j}}(x)\right).$$

   (d) Update the prediction model: $\hat{f}_k(x) = \hat{f}_{k-1}(x) + \nu\sum_{j=1}^{M_k} \psi_{\Omega_{k_j}}(x)$.
4. Output $\hat{f}(x) = \hat{f}_K(x)$.

- Initialization of approximating function with mean of all training set

$$\hat{f}_0(x) = \frac{1}{m}\sum_{i=1}^{m} \vec{y}_i \ .$$

- $l_2$ loss function for vector valued functions (regression, classification)

$$\sum_{(x,y)\in OOB} L\left(y^k, \sum_{j=1}^{M} \psi_{\Omega_{k_j}}(x)\right) = \sum_{(x,y)\in OOB} \left|\vec{y}_i^k - \sum_{j=1}^{M} \psi_{\Omega_{k_j}}(x)\right|^2 .$$

4. Wavelet based gradient boosting - experimental results

- Small datasets (comparing with XGBoost)

**Small datasets details**

| Dataset | Type | #Instances | #Features | #Unique labels |
|---|---|---|---|---|
| Airfoil-Self-Noise | R | 1503 | 5 | 1456 |
| forest-fires | R | 517 | 10 | 251 |
| Diabetes | R | 442 | 10 | 214 |
| Restaurant Revenue Prediction | R | 137 | 39 | 137 |
| Prostate | R | 97 | 8 | 85 |
| Abalone | R | 4177 | 8 | 28 |
| banknote authentication | C | 1372 | 4 | 2 |
| Blood Transfusion Service Center | C | 748 | 4 | 2 |
| breast_cancer_wisconsin | C | 569 | 30 | 2 |
| pima-indians-diabetes | C | 768 | 8 | 2 |
| Titanic | C | 891 | 6 | 2 |

**Aggregated results - Small datasets**

| Dataset | Algorithm | Avg MSE |
|---|---|---|
| Airfoil-Self-Noise | XGBoost | 2.24 |
| | WGB | 2.48 |
| forest-fires | XGBoost | 75.66 |
| | WGB | 74.55 |
| Diabetes | XGBoost | 61.44 |
| | WGB | 57.07 |
| Restaurant Revenue Prediction | XGBoost | 28.87 |
| | WGB | 28.73 |
| Prostate | XGBoost | 0.9 |
| | WGB | 0.88 |
| Abalone | XGBoost | 2.18 |
| | WGB | 2.17 |
| banknote authentication | XGBoost | 0.09 |
| | WGB | 0.09 |
| Blood Transfusion Service Center | XGBoost | 0.41 |
| | WGB | 0.4 |
| Breast cancer Wisconsin | XGBoost | 0.36 |
| | WGB | 0.36 |
| Pima indians diabetes | XGBoost | 0.42 |
| | WGB | 0.4 |
| Titanic | XGBoost | 0.37 |
| | WGB | 0.36 |

- Noisy datasets (mislabeling)

**Datasets details**

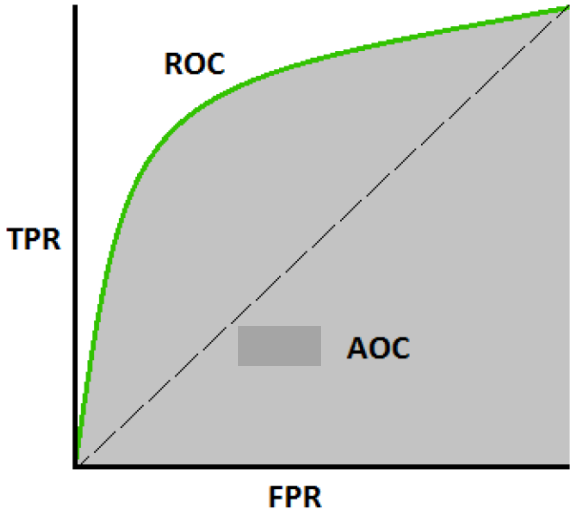
| Dataset | #Samples | #Features |
|---|---|---|
| Banana | 5299 | 2 |
| PID | 768 | 8 |
| Heart | 270 | 13 |
| TwoNorm | 7400 | 20 |

**Comparison on datasets with mislabeling (% accuracy)**

| Dataset name | Noise Level | rAdaBoost | rBoost-Fixed γ | rBoost | GBoost | MBoost | WGB |
|---|---|---|---|---|---|---|---|
| Banana | 0.1 | 86.87 ± 1.1 | 87.06 ± 0.9 | 87.04 ± 0.9 | 83.91 ± 1.6 | 78.13 ± 3.4 | **87.60 ± 1.6** |
| | 0.3 | 85.27 ± 3.0 | **85.53 ± 2.1** | 85.06 ± 2.7 | 79.38 ± 1.6 | 75.31 ± 2.5 | 85.49 ± 1.3 |
| PID | 0.1 | 74.20 ± 2.3 | 74.37 ± 1.5 | **74.80 ± 2.4** | 72.60 ± 2.0 | 75.67 ± 1.9 | 74.21 ± 6.2 |
| | 0.3 | 72.53 ± 1.9 | 70.43 ± 2.4 | 71.43 ± 2.3 | 69.40 ± 2.9 | 73.33 ± 2.3 | **75.65 ± 7.5** |
| Heart | 0.1 | 78.40 ± 3.1 | 79.70 ± 3.5 | 79.10 ± 4.4 | 76.40 ± 3.1 | 77.60 ± 3.5 | **80.74 ± 8.2** |
| | 0.3 | **78.50 ± 4.0** | 77.40 ± 6.5 | 78.10 ± 4.3 | 70.00 ± 5.5 | 75.20 ± 3.7 | 73.70 ± 11.5 |
| TwoNorm | 0.1 | 95.70 ± 0.8 | 95.58 ± 0.9 | 95.59 ± 0.7 | 90.35 ± 1.0 | 92.79 ± 0.5 | **96.40 ± 0.8** |
| | 0.3 | 93.33 ± 0.9 | 93.13 ± 1.3 | 93.40 ± 1.1 | 83.94 ± 2.0 | 91.16 ± 0.9 | **94.82 ± 0.7** |

- Area Under the Curve (AUC)

In a binary classification problem when the model outputs a probability between [0,1], we can modify the threshold for prediction of positive between [0.1] and measure the True Positive Rate (TP/P) versus the False Positive Rate (FP/N=1-TN/N). We then take the area under the curve.



- Class imbalance

**Comparison of class imbalance results (AUC)**

| Dataset name | Best Bagging-based method | Best Boosting-based method | Best Classic method | WGB |
|---|---|---|---|---|
| | UB4 | RUS 1 | SMT | WGB |
| glass1 | 0.737 | 0.763 | 0.737 | 0.816 |
| ecoli0vsl | 0.980 | 0.969 | 0.973 | 0.986 |
| Wisconsin | 0.960 | 0.964 | 0.953 | 0.985 |
| Pima | 0.760 | 0.726 | 0.725 | 0.809 |
| Iris0 | 0.990 | 0.990 | 0.990 | 1.000 |
| glass0 | 0.814 | 0.813 | 0.775 | 0.880 |
| yeast1 | 0.722 | 0.719 | 0.709 | 0.775 |
| vehicle1 | 0.787 | 0.747 | 0.730 | 0.810 |
| vehicle2 | 0.964 | 0.970 | 0.950 | 0.982 |
| vehicle3 | 0.802 | 0.765 | 0.728 | 0.805 |
| Haberman | 0.664 | 0.655 | 0.616 | 0.651 |
| glass0123vs456 | 0.904 | 0.930 | 0.923 | 0.960 |
| vehicle0 | 0.952 | 0.958 | 0.919 | 0.982 |
| ecoli1 | 0.900 | 0.883 | 0.911 | 0.951 |
| new-thyroid2 | 0.958 | 0.938 | 0.966 | 0.996 |
| new-thyroid1 | 0.964 | 0.958 | 0.963 | 0.993 |
| ecoli2 | 0.884 | 0.899 | 0.811 | 0.918 |
| Segimmt0 | 0.988 | 0.993 | 0.993 | 0.987 |
| glass6 | 0.904 | 0.918 | 0.884 | 0.935 |
| yeast3 | 0.934 | 0.925 | 0.891 | 0.957 |
| ecoli3 | 0.908 | 0.856 | 0.812 | 0.923 |
| Page-blocks0 | 0.958 | 0.948 | 0.950 | 0.990 |
| yeast2vs4 | 0.936 | 0.933 | 0.859 | 0.981 |
| yeast05679vs4 | 0.794 | 0.803 | 0.760 | 0.863 |
| vowel0 | 0.947 | 0.943 | 0.951 | 0.988 |
| glass016vs2 | 0.754 | 0.617 | 0.606 | 0.720 |
| glass2 | 0.769 | 0.780 | 0.639 | 0.690 |
| ecoli4 | 0.888 | 0.942 | 0.779 | 0.906 |
| suttle0vs4 | 1.000 | 1.000 | 1.000 | 1.000 |
| yrast1vs7 | 0.786 | 0.715 | 0.700 | 0.760 |
| glass4 | 0.846 | 0.915 | 0.887 | 0.963 |
| page-blocks13vs4 | 0.978 | 0.987 | 0.996 | 0.992 |
| abalone9vs18 | 0.719 | 0.693 | 0.628 | 0.827 |
| glass016vs5 | 0.943 | 0.989 | 0.813 | 0.946 |
| suttle2vs4 | 1.000 | 1.000 | 0.992 | 0.994 |
| yrast1458vs7 | 0.606 | 0.567 | 0.537 | 0.594 |
| glass5 | 0.949 | 0.943 | 0.881 | 0.982 |
| Yeast2vs8 | 0.783 | 0.789 | 0.834 | 0.616 |
| yeast4 | 0.855 | 0.812 | 0.712 | 0.865 |
| yeast1289vs7 | 0.734 | 0.721 | 0.683 | 0.765 |
| yeast5 | 0.952 | 0.959 | 0.934 | 0.968 |
| ecoli0137vs26 | 0.745 | 0.794 | 0.814 | 0.814 |
| yeast6 | 0.869 | 0.823 | 0.829 | 0.876 |
| Abalone19 | 0.721 | 0.631 | 0.521 | 0.594 |

5. Proof of the Jackson Theorem [7] – additional arguments

   a. For the case $r=1$. Let $\Omega'$ by a child of $\Omega$. Then,

$$\begin{aligned}
\left\|\psi_{\Omega'}\right\|_\infty &= \left\|\left(C_{\Omega'}-C_\Omega\right)1_{\Omega'}\right\|_\infty \\
&= \left|C_{\Omega'}-C_\Omega\right| \\
&= \left|\Omega'\right|^{-1/p}\left(\left|\Omega'\right|\left|C_{\Omega'}-C_\Omega\right|^p\right)^{1/p} \\
&= \left|\Omega'\right|^{-1/p}\left(\int_{\Omega'}\left|C_{\Omega'}-C_\Omega\right|^p dx\right)^{1/p} \\
&= \left|\Omega'\right|^{-1/p}\left\|\psi_{\Omega'}\right\|_p .
\end{aligned}$$

   b. The application of the $\Gamma$ function

$$\begin{aligned}
\sum_{i=1}^{I}\left|\Omega_i\right|^{-1/p}1_{\Omega_i}(x) &= \Gamma(x)^{-1/p}\sum_{i=1}^{I}\left(\frac{\Gamma(x)}{\left|\Omega_i\right|}\right)^{1/p}1_{\Omega_i}(x) \\
&\underset{(26)}{\leq} cJ\Gamma(x)^{-1/p}.
\end{aligned}$$

c. Weak $l_\tau$ norm

Strong norm $\|\beta\|_{l_\tau} = \left( \sum_k |\beta_k|^\tau \right)^{1/\tau}$,

For any $\varepsilon > 0$, $\#\{\beta_k : |\beta_k| > \varepsilon\} \varepsilon^\tau \le \sum_{k, |\beta_k| > \varepsilon} |\beta_k|^\tau \le \|\beta\|_{l_\tau}^\tau$. This implies that $\|\beta\|_{wl_\tau} \le \|\beta\|_{l_\tau}$ and $l_\tau \subset wl_\tau$.

Typical example for $wl_1$ sequence $\beta_k = 1/k$. $\beta \notin l_1$ however,

$$\#\left\{\beta_k : |\beta_k| > \frac{1}{N}\right\} N^{-1} = \frac{N-1}{N} \le 1 \Rightarrow \|\beta\|_{wl_1} = 1$$

d. Using the $wl_\tau$ norm to bound the number of wavelets in a "dyadic interval"

$$\bigcup_{\nu \le m} \Xi_\nu = \left\{ \Omega \in \mathcal{F} : \underbrace{2^{-m} \mathcal{N}_\tau(f, \mathcal{F})}_{\varepsilon} \le w_{j(\Omega)} \|\psi_\Omega\|_p \right\} \Rightarrow$$

$$\left( \#\bigcup_{\nu \le m} \Xi_\nu \right) \underbrace{2^{-m\tau} \mathcal{N}_\tau(f, \mathcal{F})^\tau}_{\varepsilon^\tau} \le \left\| \left\{ w_{j(\Omega)} \|\psi_\Omega\|_p \right\} \right\|_{l_\tau}^\tau = \mathcal{N}_\tau(f, \mathcal{F})^\tau \Rightarrow$$

$$\left( \#\bigcup_{\nu \le m} \Xi_\nu \right) \le 2^{m\tau}$$

e. The final geometric sum.

$$\sum_{\nu=m+1}^{\infty} 2^{-\nu(1-\tau/p)} = \frac{2^{-(m+1)(1-\tau/p)}}{1 - 2^{-(1-\tau/p)}} = c 2^{-m(1-\tau/p)} = c 2^{-m\tau(1/\tau - 1/p)} \le c M^{-(1/\tau - 1/p)}$$