Large Margin LAD Models and LAD-based Regression

Tibérius O. Bonates

Princeton Consultants, Inc.

DIMACS - RUTCOR Workshop on Boolean and
Pseudo-Boolean Functions in Memory of Peter L. Hammer

January, 2009

Summary

- **Introduction**
- Large Margin LAD Classifiers
- LAD-Based Regression
- Conclusions and Future Work

Conjunctions and Patterns

Consider a dataset $\Omega = \Omega^+ \cup \Omega^- \subset \{0,1\}^n$, with $\Omega^+ \cap \Omega^- = \emptyset$.

A conjunction is a clause involving literals from $\{x_1, \ldots, x_n, \overline{x_1}, \ldots, \overline{x_n}\}$. A conjunction defines a subcube of $\{0,1\}^n$ in which a subset of the components is fixed to 0 or 1.

A positive pattern is a homogeneous conjunction, i.e. a subcube having:

(i) a nonempty intersection with Ω^+,

(ii) an empty intersection with Ω^-.

A negative pattern is defined similarly.

The concept of a pattern is frequently relaxed to allow the inclusion of a <u>small number of points</u> of the other set.

LAD Models and Discriminants

A pattern P is said to "cover" a point ω if ω is in the subcube defined by P.

LAD Model: collection $M = M^+ \cup M^-$ of positive and negative patterns so that every point in Ω is covered by at least one pattern of M.

Let $M^+ = \{P_1, \ldots, P_r\}$ and $M^- = \{N_1, \ldots, N_s\}$.

LAD Discriminant Function: for a point $\omega \in \{0,1\}^n$ the discriminant function associated with model M is given by

$$\Delta(\omega) = \sum_{j=1}^{r} \alpha_j P_j(\omega) - \sum_{j=1}^{s} \beta_j N_j(\omega),$$

where α, β are positive real vectors and $P_j(\omega) = 1$ if ω is covered by P_j, and $P_j(\omega) = 0$ otherwise.

LAD Classification

The weights α and β are chosen so that

$$\Delta(\omega) \geq 0, \text{ for every } \omega \in \Omega^+$$
$$\Delta(\omega) \leq 0, \text{ for every } \omega \in \Omega^-.$$

Given a model M and an associated discriminant function Δ, the LAD classification of a new point $\omega \in \{0,1\}^n$ is as follows:

If $\Delta(\omega) > 0$, then ω is classified as positive

If $\Delta(\omega) < 0$, then ω is classified as negative

If $\Delta(\omega) = 0$, then ω is <u>not classified</u>

Summary

- Introduction
- **Large Margin LAD Classifiers**
- LAD-Based Regression
- Conclusions and Future Work

LAD Models and Margin of Separation

Dataset:

Class	1	2	3	4	5	6	7
$+$	1	0	0	0	0	1	1
$+$	0	1	1	1	0	0	1
$+$	1	1	0	1	0	1	0
$-$	0	0	0	1	0	1	0
$-$	0	1	1	0	1	1	1
$-$	1	0	1	1	1	0	0

LAD Models and Margin of Separation

Dataset:

Class	1	2	3	4	5	6	7	
$+$	**1**	0	**0**	0	0	1	1	$\Delta = +0.5$
$+$	0	**1**	1	**1**	0	0	1	$\Delta = +0.5$
$+$	**1**	**1**	**0**	**1**	0	1	0	$\Delta = +1.0$
$-$	0	0	0	1	0	1	0	
$-$	0	1	1	0	1	1	1	
$-$	1	0	1	1	1	0	0	

Positive patterns: $x_1\overline{x_3}$ and x_2x_4 (both with a 0.5 weight).

LAD Models and Margin of Separation

Dataset:

Class	1	2	3	4	5	6	7	
$+$	**1**	0	**0**	0	0	1	1	$\Delta = +0.5$
$+$	0	**1**	1	**1**	0	0	1	$\Delta = +0.5$
$+$	**1**	**1**	**0**	**1**	0	1	0	$\Delta = +1.0$
$-$	**0**	0	0	1	0	**1**	0	$\Delta = -0.5$
$-$	**0**	1	**1**	0	**1**	**1**	1	$\Delta = -1.0$
$-$	1	0	**1**	1	**1**	0	0	$\Delta = -0.5$

Positive patterns: $x_1\overline{x_3}$ and x_2x_4 (both with a 0.5 weight).
Negative patterns: x_3x_5 and $\overline{x_1}x_6$ (both with a –0.5 weight).

Margin of separation: $+0.5 + |-0.5| = 1.0$.

LAD Models and Margin of Separation

Dataset:

Class	1	2	3	4	5	6	7	
+	**1**	0	**0**	0	**0**	1	**1**	$\Delta = +0.66$
+	0	**1**	1	**1**	**0**	0	**1**	$\Delta = +0.66$
+	**1**	**1**	**0**	**1**	0	1	0	$\Delta = +0.66$
−	**0**	0	0	1	0	**1**	0	$\Delta = -0.5$
−	**0**	1	**1**	0	**1**	**1**	1	$\Delta = -1.0$
−	1	0	**1**	1	**1**	0	0	$\Delta = -0.5$

Positive patterns: $x_1\overline{x_3}$, x_2x_4, and $\mathbf{x_5x_7}$ (all with a 0.33 weight).
Negative patterns: x_3x_5 and $\overline{x_1}x_6$ (both with a –0.5 weight).

Margin of separation: $+0.66 + |-0.5| = 1.16$.

Large Margin LAD Models

Problem: Construct a LAD model and an associated discriminant function maximizing the margin of separation between points in the training set.

Recall that

$$\Delta(\omega) = \sum_j \alpha_j P_j(\omega) - \sum_k \beta_k N_k(\omega),$$

and that we want

$$\begin{aligned} \Delta(\omega) &\geq r, && \text{for } \omega \in \Omega^+, \\ \Delta(\omega) &\leq -s, && \text{for } \omega \in \Omega^-, \end{aligned}$$

for some $r, s > 0$. The margin of separation is given by $r + s$.

We want to find a <u>set of patterns</u>, values for $\underline{r}$, $\underline{s}$, and the <u>weights α and β</u> so that $r + s$ is maximized.

Large Margin LAD Models: Formulation

We can obtain an optimal discriminant function by solving the following linear program (MP):

$$\text{maximize} \quad r + s \; - \; C \sum_{\omega \in \Omega} \epsilon_\omega$$

$$\text{subject to:}$$

$$\Delta(\omega) + \epsilon_\omega \geq r, \;\; \forall\, \omega \in \Omega^+ \tag{1}$$

$$\Delta(\omega) - \epsilon_\omega \leq -s, \;\; \forall\, \omega \in \Omega^- \tag{2}$$

$$\sum_{P_i \in \mathbb{P}} \alpha_i = \sum_{N_j \in \mathbb{N}} \beta_j = 1, \tag{3}$$

with $\alpha, \beta \geq \mathbf{0}$, $\quad r \geq 0$, $\quad s \geq 0$, $\quad \epsilon_\omega \geq 0$, for every $\omega \in \Omega$, and C being a penalty factor for the violating margin of separation.

Typically MP cannot be solved directly. We apply column generation to iteratively construct an optimal discriminant function, starting from a simple set of patterns $\mathcal{P} \cup \mathcal{N}$.

Large Margin LAD Models: Subproblem

Let λ and μ be the dual variables associated to (1) and (2).

To find a conjunction with positive reduced cost we solve:

$$\text{maximize} \quad \sum_{\omega \in \Omega^+} (-\lambda_\omega) y_\omega + \sum_{\gamma \in \Omega^-} \mu_\gamma y_\omega$$

subject to:

$$\sum_{i:\omega_i=0} x_i + \sum_{j:\omega_j=1} x_j^c + n y_\omega \leq n, \forall\, \omega \in \Omega$$

$$\sum_{i:\omega_i=0} x_i + \sum_{j:\omega_j=1} x_j^c + \; y_\omega \geq \; 1, \forall\, \omega \in \Omega$$

$$x, x^c \in \{0,1\}^n$$

$$y \in \{0,1\}^{|\Omega|}.$$

Large Margin LAD Models: Subproblem

Also known as...

$$(S1) \quad \text{maximize} \quad \sum_{\omega \in \Omega} \left(\prod_{i:\omega_i=0} \overline{p_i} \prod_{j:\omega_j=1} \overline{p_j^c} \right) \beta_\omega$$

$$\text{subject to:} \quad p_j, p_j^c \in \{0,1\},\ j = 1, \ldots, n,$$

We solve $(S1)$ approximately with a simple branch-and-bound procedure, branching on terms.

Large Margin LAD Models

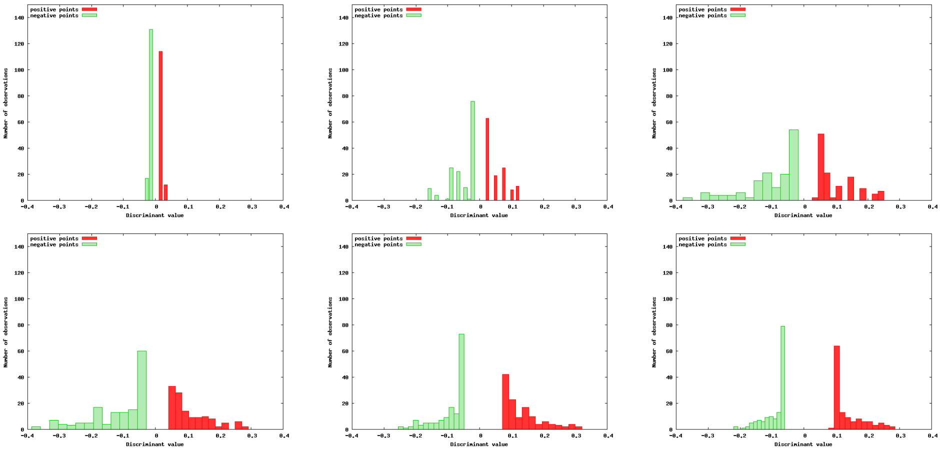
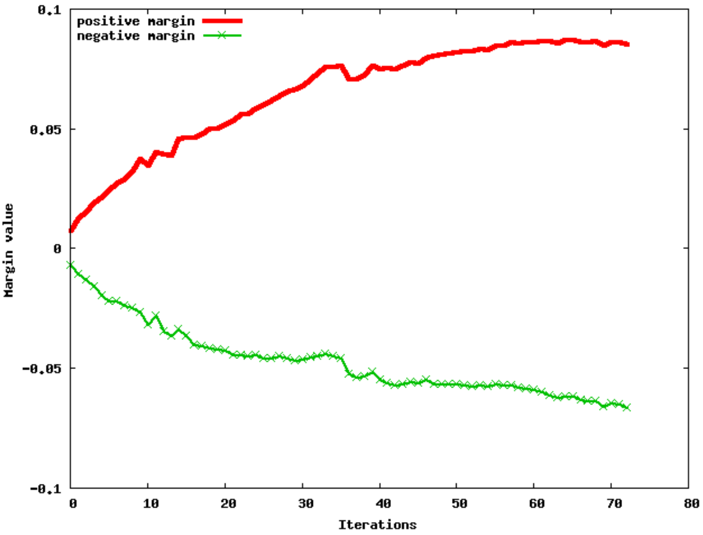


Table: Histograms of discriminant values of positive and negative points.

(— positive points; — negative points)

Large Margin LAD Models



Large Margin LAD Classifiers: Results

Dataset	SMO	J48	Rand.For.	Mult.Perc.	LM-LAD
breast-w	0.965 ± 0.011	0.939 ± 0.012	0.967 ± 0.009	0.956 ± 0.012	0.942 ± 0.024
credit-a	0.864 ± 0.025	0.856 ± 0.031	0.882 ± 0.027	0.831 ± 0.032	0.815 ± 0.044
hepatitis	0.772 ± 0.084	0.652 ± 0.086	0.722 ± 0.101	0.727 ± 0.065	0.738 ± 0.091
krkp	0.996 ± 0.003	0.994 ± 0.003	0.992 ± 0.003	0.993 ± 0.002	0.962 ± 0.031
boston	0.889 ± 0.028	0.837 ± 0.045	0.875 ± 0.024	0.893 ± 0.031	0.840 ± 0.045
bupa	0.701 ± 0.045	0.630 ± 0.041	0.731 ± 0.046	0.643 ± 0.020	0.678 ± 0.034
heart	0.837 ± 0.039	0.799 ± 0.052	0.834 ± 0.051	0.815 ± 0.025	0.814 ± 0.033
pima	0.727 ± 0.029	0.722 ± 0.026	0.736 ± 0.030	0.726 ± 0.023	0.682 ± 0.023
sick	0.824 ± 0.027	0.926 ± 0.020	0.832 ± 0.023	0.852 ± 0.049	0.815 ± 0.041
voting	0.961 ± 0.018	0.960 ± 0.015	0.961 ± 0.016	0.944 ± 0.025	0.945 ± 0.025

Table: Classification accuracy of Weka algorithms and LM-LAD.

	SMO	J48	Rand.For.	Mult.Perc.	CAP-LAD
J48	4-1-5				
Rand.For.	1-1-8	4-1-5			
Mult.Perc.	0-4-6	2-2-6	0-2-8		
S.Log.	1-1-8	1-2-7	0-2-8	1-2-7	
CAP-LAD	0-1-9	2-1-7	0-0-10	2-1-7	
LM-LAD	0-0-10	0-1-9	0-1-9	0-0-10	0-1-9

Table: Matrix of wins, losses and ties (95% confidence interval).

Summary

- Introduction
- Large Margin LAD Classifiers
- **LAD-Based Regression**
- Conclusions and Future Work

LAD-Based Regression

Consider a dataset $\Omega \subset \{0,1\}^n$ and the values of an unknown *target* function $r : \Omega \to \mathbb{R}$, and let $y^i = r(\omega^i)$, $\omega^i \in \Omega$. We want to find a function $f : \Omega \to \mathbb{R}$ that approximates r "well enough".

Measures of interest:

- Least Absolute Residual (LAR) measure: $\sum_{i=1}^{|\Omega|} |f(\omega^i) - y^i|$.
- Correlation coefficient between the values of f and r over Ω.

Approach:

Construct a regression function in the space of conjunctions (i.e. use conjunctions as independent variables).

LAD-Based Regression: Formulation

Let $\mathcal{C}^0 = \{C_1, \ldots, C_n\}$ be the set of conjunctions consisting of the n positive literals, i.e. $C_j = x_j$ $(j = 1, \ldots, n)$. The LAR-best linear approximation of r using $\mathcal{C}^0$ can be found by solving

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i=1}^{m} e_i \\
\text{subject to:} \quad & e_i + \beta_0 + \sum_{C_j \in \mathcal{C}^0} \beta_j \, \mathcal{C}_j(\omega^i) \geq y^i, \; i = 1, \ldots, m \qquad (4) \\
& e_i - \beta_0 - \sum_{C_j \in \mathcal{C}^0} \beta_j \, \mathcal{C}_j(\omega^i) \geq -y^i, \; i = 1, \ldots, m \qquad (5) \\
& \beta_j \geq 0, \quad j = 0, \ldots, n \\
& e_i \geq 0, \quad i = 1, \ldots, m.
\end{aligned}
$$

Let λ^* and μ^* be the optimal vectors of dual variables associated to constraints (4) and (5), respectively.

LAD-Based Regression: Subproblem

Find a new conjunction whose inclusion in $\mathcal{C}^0$ is likely to improve the current LAR-approximation by solving the following problem (S2):

$$\begin{array}{ll} \text{maximize} & \displaystyle\sum_{\omega\in\Omega}\left(\prod_{i:\omega_i=0}\overline{p_i}\prod_{j:\omega_j=1}\overline{p_j^c}\right)(\lambda_\omega^*-\mu_\omega^*) \\ \text{subject to:} & p_j, p_j^c \in \{0,1\},\ j=1,\ldots,n, \end{array}$$

with p_j being a binary decision variable corresponding to the inclusion of literal x_j in the resulting conjunction. Similarly, p_j^c corresponds to the inclusion of $\overline{x_j}$.

Problem (S2) is an instance of (S1) and is solved with the B&B algorithm previously mentioned.

LAD-Based Regression: Results

	Mean Absolute Error					
Algorithms	AB	BH	MPG	RAK	SV	Borda
LR	1.59	3.33	2.73	0.16	0.87	9
MP	1.62	2.94	2.90	0.13	0.44	14
SVR	1.54	3.17	2.63	0.16	0.70	12
PBR	1.82	3.11	2.37	0.15	0.29	15

Table: Mean absolute error of regression algorithms applied to 5 datasets.

	Correlation					
Algorithms	AB	BH	MPG	RAK	SV	Borda
LR	0.73	0.86	0.89	0.64	0.67	9
MP	0.75	0.91	0.92	0.82	0.90	19
SVR	0.73	0.84	0.89	0.64	0.63	9
PBR	0.51	0.86	0.89	0.68	0.94	13

Table: Correlation of regression algorithms applied to 5 datasets.

Summary

- Introduction
- Large Margin LAD Classifiers
- LAD-Based Regression
- **Conclusions and Future Work**

Conclusions and Future Work

- ✓ Large margin LAD models: parameter-free, accurate models
- ✓ Extension of LAD methodology to regression problems
- ☕ Quadratic objective functions
- ☕ More applications and different loss functions for regression

Reading Material

- Discrete Applied Mathematics: http://dx.doi.org/10.1016/j.dam.2007.06.004
- Annals of Operations Research: vol. 148, pp. 203–225, 2006
- Annals of Mathematics and Artificial Intelligence: vol. 49, pp. 265-312, 2007
- RUTCOR Research Reports: 9-2006, 3-2007, 21-2007, 22-2007
- Coming soon...

⇒ Tiberius.Bonates@rutgers.edu