Unsupervised Classification via Convex Absolute Value Inequalities

Olvi Mangasarian
University of Wisconsin - Madison
University of California - San Diego

January 17, 2017

Summary

- Classify completely unlabeled data

Summary

- Classify completely unlabeled data
- Utilize convex inequalities containing absolute values of the data

Summary

- Classify completely unlabeled data
- Utilize convex inequalities containing absolute values of the data
- By using absolute value inequalities (AVIs) unlabeled data can be partitioned into two classes that capture most of the labels dropped from the data

Summary

- Classify completely unlabeled data
- Utilize convex inequalities containing absolute values of the data
- By using absolute value inequalities (AVIs) unlabeled data can be partitioned into two classes that capture most of the labels dropped from the data
- Inclusion of partially labeled data leads to a semisupervised classifier

Summary

- Classify completely unlabeled data
- Utilize convex inequalities containing absolute values of the data
- By using absolute value inequalities (AVIs) unlabeled data can be partitioned into two classes that capture most of the labels dropped from the data
- Inclusion of partially labeled data leads to a semisupervised classifier
- Computational results include unsupervised and semisupervised classification of the Wisconsin Breast Cancer Wisconsin (Diagnostic) Data Set

Summary

- Classify completely unlabeled data
- Utilize convex inequalities containing absolute values of the data
- By using absolute value inequalities (AVIs) unlabeled data can be partitioned into two classes that capture most of the labels dropped from the data
- Inclusion of partially labeled data leads to a semisupervised classifier
- Computational results include unsupervised and semisupervised classification of the Wisconsin Breast Cancer Wisconsin (Diagnostic) Data Set
- Conclusion & Outlook

Introduction

- We begin with the following convex absolute value inequality (AVI):

$$|x'w - \gamma| \leq 1, \tag{1}$$

where $|\cdot|$ denotes the absolute value. Here the column vector x represents any data point in an n-dimensional space R^n, $w \in R^n$ is the normal vector to the classifying plane $x'w - \gamma = 0$, γ determines the distance from the origin of the plane, and the prime denotes the transpose of the column vector x.

Introduction

- We begin with the following convex absolute value inequality (AVI):

$$|x'w - \gamma| \leq 1, \tag{1}$$

where $|\cdot|$ denotes the absolute value. Here the column vector x represents any data point in an n-dimensional space R^n, $w \in R^n$ is the normal vector to the classifying plane $x'w - \gamma = 0$, γ determines the distance from the origin of the plane, and the prime denotes the transpose of the column vector x.

- The AVI (1) is equivalent to dividing R^n into two overlapping halfspaces by the following two linear inequalities:

$$\begin{array}{rcl} x'w & \leq & \gamma + 1, \\ x'w & \geq & \gamma - 1. \end{array} \tag{2}$$

Introduction Continued

- The key to our approach is to represent the last two inequalities of (2) by the single absolute value inequality AVI (1). Thus if $x'w - \gamma \geq 0$ then AVI (1) reduces to the first linear inequality of (2), whereas if $x'w - \gamma \leq 0$ then AVI (3) reduces to the second linear inequality of (2).

Introduction Continued

- The key to our approach is to represent the last two inequalities of (2) by the single absolute value inequality AVI (1). Thus if $x'w - \gamma \geq 0$ then AVI (1) reduces to the first linear inequality of (2), whereas if $x'w - \gamma \leq 0$ then AVI (3) reduces to the second linear inequality of (2).
- Hence if we impose the AVI (1) on an unlabeled dataset, the dataset will be divided into two categories to best fit the AVI (1) or equivalently the two linear inequalities (2).

Introduction Continued

- The key to our approach is to represent the last two inequalities of (2) by the single absolute value inequality AVI (1). Thus if $x'w - \gamma \geq 0$ then AVI (1) reduces to the first linear inequality of (2), whereas if $x'w - \gamma \leq 0$ then AVI (3) reduces to the second linear inequality of (2).
- Hence if we impose the AVI (1) on an unlabeled dataset, the dataset will be divided into two categories to best fit the AVI (1) or equivalently the two linear inequalities (2).
- Our objective will then be to minimize the overlap between the bounding planes $x'w = \gamma \pm 1$.

Unsupervised and Semisupervised Classification

- We begin with unlabeled dataset consisting of m points in the n-dimensional space R^n represented by the $m \times n$ matrix A.

Unsupervised and Semisupervised Classification

- We begin with unlabeled dataset consisting of m points in the n-dimensional space R^n represented by the $m \times n$ matrix A.
- We also have a labeled dataset consisting of k points in R^n represented by the $k \times n$ matrix H and labeled by the $k \times k$ diagonal matrix D with entries of ± 1 which denote which class of $+1$ or -1 each row of H belongs to.

Unsupervised and Semisupervised Classification

- We begin with unlabeled dataset consisting of m points in the n-dimensional space R^n represented by the $m \times n$ matrix A.
- We also have a labeled dataset consisting of k points in R^n represented by the $k \times n$ matrix H and labeled by the $k \times k$ diagonal matrix D with entries of ± 1 which denote which class of $+1$ or -1 each row of H belongs to.
- Thus we wish to find two planes $x'w - \gamma = \pm 1$ in R^n that specify the ± 1 feasible regions generated by the two inequalities of (2) and which satisfy with minimal error vector s the following inequalities:

Unsupervised and Semisupervised Classification

- We begin with unlabeled dataset consisting of m points in the n-dimensional space R^n represented by the $m \times n$ matrix A.
- We also have a labeled dataset consisting of k points in R^n represented by the $k \times n$ matrix H and labeled by the $k \times k$ diagonal matrix D with entries of ± 1 which denote which class of $+1$ or -1 each row of H belongs to.
- Thus we wish to find two planes $x'w - \gamma = \pm 1$ in R^n that specify the ± 1 feasible regions generated by the two inequalities of (2) and which satisfy with minimal error vector s the following inequalities:
-

$$\begin{array}{rcll} |Aw - e\gamma| & \leq & e, & \\ D(Hw - e\gamma) + s & \geq & e, & (3) \\ s & \geq & 0. & \end{array}$$

Minimizing Classification Error

- We now minimize the nonnegative slack variable appearing in (3) as well as:

Minimizing Classification Error

- We now minimize the nonnegative slack variable appearing in (3) as well as:
- Maximizing the 1-norm of (w, γ) in order to minimize the distance between the two overlapping feasible regions of the inequalities of (2) while μ is a positive parameter that balances the two groups of objectives of (4) as follows:

Minimizing Classification Error

- We now minimize the nonnegative slack variable appearing in (3) as well as:
- Maximizing the 1-norm of (w, γ) in order to minimize the distance between the two overlapping feasible regions of the inequalities of (2) while μ is a positive parameter that balances the two groups of objectives of (4) as follows:
-

$$\begin{aligned} \min_{(w,\gamma,s)\in R^{n+1+k}} \quad & -e'|w| - |\gamma| + \mu e's \\ \text{s.t.} \quad & |Aw - e\gamma| \leq e, \\ & D(Hw - e\gamma) + s \geq e, \\ & s \geq 0, \end{aligned} \tag{4}$$

Handling Absolute Values in Objective and Constraints

- We replace the term $|Aw - e\gamma|$ in the absolute value inequality by an upper bound r on it: $-r \leq (Aw - e\gamma) \leq r$ whose 1-norm is minimized with objective function weight μ.

Handling Absolute Values in Objective and Constraints

- We replace the term $|Aw - e\gamma|$ in the absolute value inequality by an upper bound r on it: $-r \leq (Aw - e\gamma) \leq r$ whose 1-norm is minimized with objective function weight μ.
- This results in the following linearly constrained concave minimization problem:

$$\min_{(w,\gamma,r,s)\in R^{n+1+m+k}} -e'|w| - |\gamma| + \nu e' r + \mu e' s$$

$$\text{s.t.} -r \leq \begin{array}{rcl} Aw - e\gamma & \leq & r, \\ r & \leq & e \\ -D(Hw - e\gamma) - s & \leq & -e, \\ s & \geq & 0. \end{array} \quad (5)$$

Handling Absolute Values in Objective and Constraints

- We replace the term $|Aw - e\gamma|$ in the absolute value inequality by an upper bound r on it: $-r \leq (Aw - e\gamma) \leq r$ whose 1-norm is minimized with objective function weight μ.
- This results in the following linearly constrained concave minimization problem:

$$\begin{aligned} & \min_{(w,\gamma,r,s)\in R^{n+1+m+k}} \; -e'|w| - |\gamma| + \nu e' r + \mu e' s \\ \text{s.t.} \; -r \leq & \begin{array}{rcl} Aw - e\gamma & \leq & r, \\ r & \leq & e \\ -D(Hw - e\gamma) - s & \leq & -e, \\ s & \geq & 0. \end{array} \\ & \qquad (5) \end{aligned}$$

- Will solve this problem by a finite sequence of linear programs.

Successive Linear Programming Solution of Unsupervised & Semisupervised Classification

- The objective function of our optimization problem (5) is concave and the constraints are linear.

Successive Linear Programming Solution of Unsupervised & Semisupervised Classification

- The objective function of our optimization problem (5) is concave and the constraints are linear.
- The $n+1$ columns of the $m \times (n+1)$ matrix $[A \;\; e]$ are generally linearly independent, it follows that the iterates $(w^i, \gamma^i, r^i, s^i)$ of our successive linearization algorithm below are bounded.

Successive Linear Programming Solution of Unsupervised & Semisupervised Classification

- The objective function of our optimization problem (5) is concave and the constraints are linear.
- The $n+1$ columns of the $m \times (n+1)$ matrix $[A \;\; e]$ are generally linearly independent, it follows that the iterates $(w^i, \gamma^i, r^i, s^i)$ of our successive linearization algorithm below are bounded.
- Hence, by a concave minimization theorem [OLM 1997], we have that the iterates $(w^i, \gamma^i, r^i, s^i)$ strictly decrease the objective function of (5) and terminate in a finite number of steps (typically less than five) at a point satisfying the minimum principle necessary optimality condition for our problem (5).

Successive Linear Programming Solution of Unsupervised & Semisupervised Classification

- The objective function of our optimization problem (5) is concave and the constraints are linear.
- The $n+1$ columns of the $m \times (n+1)$ matrix $[A \;\; e]$ are generally linearly independent, it follows that the iterates $(w^i, \gamma^i, r^i, s^i)$ of our successive linearization algorithm below are bounded.
- Hence, by a concave minimization theorem [OLM 1997], we have that the iterates $(w^i, \gamma^i, r^i, s^i)$ strictly decrease the objective function of (5) and terminate in a finite number of steps (typically less than five) at a point satisfying the minimum principle necessary optimality condition for our problem (5).
- We now state our successive linearization algorithm.

SLA: Successive Linearization Algorithm

(0) Choose parameter values for (μ, ν) in (5), typically $1e-4$.

SLA: Successive Linearization Algorithm

(0) Choose parameter values for (μ, ν) in (5), typically $1e-4$.

(I) Initialize the algorithm by choosing an initial nonnegative random vector in R^{n+1} for (w^0, γ^0). Set iteration number to $i = 0$.

SLA: Successive Linearization Algorithm

(0) Choose parameter values for (μ, ν) in (5), typically $1e-4$.

(I) Initialize the algorithm by choosing an initial nonnegative random vector in R^{n+1} for (w^0, γ^0). Set iteration number to $i = 0$.

(II) Solve the following linear program, which is a linearization of (5) around $(w^i, \gamma^i, r^i, s^i)$, for $(w^{i+1}, \gamma^{i+1}, r^{i+1}, s^{i+1})$:

$$\min_{(w,\gamma,r,s)} \; -sign(w^i)'w - sign(\gamma^i)\gamma + \nu e'r + \mu e's$$

$$\begin{array}{rcc} r \leq & Aw - e\gamma & \leq \quad r, \\ & r & \leq \quad e \\ & -D(Hw - e\gamma) - s & \leq \quad -e, \\ & s & \geq \quad 0, \end{array} \qquad (6)$$

SLA: Successive Linearization Algorithm

(0) Choose parameter values for (μ, ν) in (5), typically $1e-4$.

(I) Initialize the algorithm by choosing an initial nonnegative random vector in R^{n+1} for (w^0, γ^0). Set iteration number to $i = 0$.

(II) Solve the following linear program, which is a linearization of (5) around $(w^i, \gamma^i, r^i, s^i)$, for $(w^{i+1}, \gamma^{i+1}, r^{i+1}, s^{i+1})$:

$$\min_{(w,\gamma,r,s)} \; -sign(w^i)'w - sign(\gamma^i)\gamma + \nu e'r + \mu e's$$

$$\begin{array}{rcc} r \leq & Aw - e\gamma & \leq \quad r, \\ & r & \leq \quad e \\ & -D(Hw - e\gamma) - s & \leq \quad -e, \\ & s & \geq \quad 0, \end{array} \qquad (6)$$

(III) If $w^{i+1} = w^i$ stop.

SLA: Successive Linearization Algorithm

(0) Choose parameter values for (μ, ν) in (5), typically $1e-4$.

(I) Initialize the algorithm by choosing an initial nonnegative random vector in R^{n+1} for (w^0, γ^0). Set iteration number to $i = 0$.

(II) Solve the following linear program, which is a linearization of (5) around $(w^i, \gamma^i, r^i, s^i)$, for $(w^{i+1}, \gamma^{i+1}, r^{i+1}, s^{i+1})$:

$$\min_{(w,\gamma,r,s)} \; -sign(w^i)'w - sign(\gamma^i)\gamma + \nu e'r + \mu e's$$

$$\begin{array}{rcc} r \leq & Aw - e\gamma & \leq \quad r, \\ & r & \leq \quad e \\ & -D(Hw - e\gamma) - s & \leq \quad -e, \\ & s & \geq \quad 0, \end{array} \qquad (6)$$

(III) If $w^{i+1} = w^i$ stop.

(IV) Set $i = i + 1$ and go to Step (II).

Finite Termination of SLA

- **Proposition** Let $z = (w, \gamma, r, s)$, let $f(z)$ denote the concave objective function of (5) and let Z denote the feasible region of (5). Then the SLA generates a finite sequence of feasible iterates $\{z^1, z^2, \ldots, z^i\}$ of strictly decreasing function values $f(z^1) > f(z^2) >, \ldots, f(z^i)$, such that z^i satisfies the minimum principle necessary optimality condition:

$$\partial f(z^i)(z - z^i) \geq 0, \quad \forall z \in Z, \tag{7}$$

where $\partial f(z^i)$ denotes the supergradient of f at z^i.

Finite Termination of SLA

- **Proposition** Let $z = (w, \gamma, r, s)$, let $f(z)$ denote the concave objective function of (5) and let Z denote the feasible region of (5). Then the SLA generates a finite sequence of feasible iterates $\{z^1, z^2, \ldots, z^i\}$ of strictly decreasing function values $f(z^1) > f(z^2) >, \ldots, f(z^i)$, such that z^i satisfies the minimum principle necessary optimality condition:

$$\partial f(z^i)(z - z^i) \geq 0, \quad \forall z \in Z, \tag{7}$$

  where $\partial f(z^i)$ denotes the supergradient of f at z^i.

- We note that all the above results can be extended to nonlinear kernel classification by replacing the linear absolute value inequality (1) by one representing a nonlinear surface $|K(x', A')u - \gamma| \leq 1$ that is still linear in the unknowns (u, γ), but nonlinear in the data variable $x \in R^n$, where K is any nonlinear kernel.

Computational Results 1

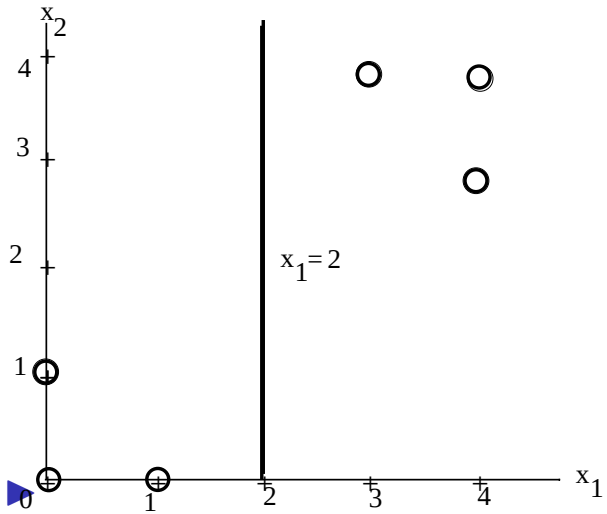
- We begin with a simple 2-dimensional unlabeled example consisting of six points in R^2 three of which lie in the southwest corner of a square and three of which lie on the northeast corner of the square: $[0, 0; 0, 1; 1, 0; 4, 4; 3, 4; 4, 3]$.

Computational Results 1

- We begin with a simple 2-dimensional unlabeled example consisting of six points in R^2 three of which lie in the southwest corner of a square and three of which lie on the northeast corner of the square: $[0, 0; 0, 1; 1, 0; 4, 4; 3, 4; 4, 3]$.
- Starting with a random (w, γ), the SLA algorithm terminates after solving three linear programs with the separating line $x_1 = 2$, which is quite appropriate for the given six unlabeled data points.

Computational Results 1

- We begin with a simple 2-dimensional unlabeled example consisting of six points in R^2 three of which lie in the southwest corner of a square and three of which lie on the northeast corner of the square: $[0, 0; 0, 1; 1, 0; 4, 4; 3, 4; 4, 3]$.
- Starting with a random (w, γ), the SLA algorithm terminates after solving three linear programs with the separating line $x_1 = 2$, which is quite appropriate for the given six unlabeled data points.



Computational Results 2

- Our second example is one of the most popular datasets available from the University of California Irvine Machine Learning Repository:

Computational Results 2

- Our second example is one of the most popular datasets available from the University of California Irvine Machine Learning Repository:
- The Wisconsin Diagnostic Breast Cancer Dataset WDBC. For our purposes we have extracted from WDBC an $m \times n$ matrix A with $m = 569$ and $n = 30$. Here m is the total number of patients in WDBC and n is the total number of features obtained from the fine needle aspirate of each patient.

Computational Results 2

- Our second example is one of the most popular datasets available from the University of California Irvine Machine Learning Repository:
- The Wisconsin Diagnostic Breast Cancer Dataset WDBC. For our purposes we have extracted from WDBC an $m \times n$ matrix A with $m = 569$ and $n = 30$. Here m is the total number of patients in WDBC and n is the total number of features obtained from the fine needle aspirate of each patient.
- We have rearranged the rows of A so that the first 357 rows are the data of benign aspirates, while the last 212 rows A are those of malignant aspirates.

Computational Results 2

- Our second example is one of the most popular datasets available from the University of California Irvine Machine Learning Repository:
- The Wisconsin Diagnostic Breast Cancer Dataset WDBC. For our purposes we have extracted from WDBC an $m \times n$ matrix A with $m = 569$ and $n = 30$. Here m is the total number of patients in WDBC and n is the total number of features obtained from the fine needle aspirate of each patient.
- We have rearranged the rows of A so that the first 357 rows are the data of benign aspirates, while the last 212 rows A are those of malignant aspirates.
- We first ran SLA on the whole matrix A as a completely unsupervised problem and obtained a correctness of 65.55%.

Computational Results 2 Continued

- Next we ran SLA twice with two different ten labeled cases, five benign and five malignant cases and obtained correctness values of 75.40% and 81.72% respectively. These results are summarized in the Table below.

Computational Results 2 Continued

- Next we ran SLA twice with two different ten labeled cases, five benign and five malignant cases and obtained correctness values of 75.40% and 81.72% respectively. These results are summarized in the Table below.

Unlabeled Data Used	Labeled Data Rows Used	Linear Programs Solved	Correctness
569	None	2	65.55%
559	Benign 201-205 Malignant 361-365	3	75.40%
559	Benign 1-5 Malignant 565-569	4	81.72%

Conclusion and Outlook

- We have proposed the use of absolute value inequalities for classifying unlabeled data.

Conclusion and Outlook

- We have proposed the use of absolute value inequalities for classifying unlabeled data.
- We have also combined the approach with standard methods for classifying labeled data.

Conclusion and Outlook

- We have proposed the use of absolute value inequalities for classifying unlabeled data.
- We have also combined the approach with standard methods for classifying labeled data.
- It will be interesting to utilize other absolute value inequality formulations to handle unlabeled data, as well as utilizing nonlinear kernels in addition to the linear kernels employed here.

Conclusion and Outlook

- We have proposed the use of absolute value inequalities for classifying unlabeled data.
- We have also combined the approach with standard methods for classifying labeled data.
- It will be interesting to utilize other absolute value inequality formulations to handle unlabeled data, as well as utilizing nonlinear kernels in addition to the linear kernels employed here.
- Hopefully our absolute value approach will lead to effective tools for handling large unlabeled data.

References

- O.L. Mangasarian: "Unsupervised Classification via Convex Absolute Value Inequalities", ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/14-01.pdf, **Optimization** 2015, Vol. 64, No. 1, 81-86.

References

- O.L. Mangasarian: "Unsupervised Classification via Convex Absolute Value Inequalities", ftp://ftp.cs.wisc.edu/pub/dmi/tech-reports/14-01.pdf, **Optimization** 2015, Vol. 64, No. 1, 81-86.
- O. L. Mangasarian: "Solution of General Linear Complementarity Problems via Nondifferentiable Concave Minimization". Acta Mathematica Vietnamica 1997, 22(1), 199-205. ftp://ftp.cs.wisc.edu/math-prog/tech-reports/96-10.ps