Nonsmooth optimization models in Data Mining

Adil Bagirov

SITE, University of Ballarat, Victoria, Australia

August 28, 2013, Turku University, Finland

Outline

- Introduction
- Cluster analysis via nonsmooth optimization
- Data classification via nonsmooth optimization
- SVM and nonsmooth optimization
- Clusterwise regression via nonsmooth optimization
- Nonsmooth optimization methods in machine learning
- Incremental algorithms
- Conclusions

Introduction

Data analysis is a process of gathering, modeling, and transforming data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making.

Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to learn based on data, such as from databases.

Data mining is the process of extracting hidden patterns from data.

Introduction

- Mathematical programming provides powerful tools for solving many data mining problems;
- Nonsmooth optimization algorithms are increasingly playing vital role in modeling and solving data mining problems;
- In this talk we concentrate on nonsmooth optimization approaches to data classification, clustering and clusterwise regression problems.

Cluster analysis via nonsmooth optimization

In cluster analysis we assume that we have been given a finite set of points A in the n-dimensional space $\mathbb{R}^n$, that is

$$A = \{a^1, \ldots, a^m\}, \text{ where } a^i \in \mathbb{R}^n, \ i = 1, \ldots, m.$$

We consider the hard unconstrained partition clustering problem, that is the distribution of the points of the set A into a given number k of disjoint subsets $A^j, \ j = 1, \ldots, k$ with respect to predefined criteria such that:

1) $A^j \neq \emptyset, \ j = 1, \ldots, k$;

2) $A^j \bigcap A^l = \emptyset, \ j, l = 1, \ldots, k, \ j \neq l$;

3) $A = \bigcup\limits_{j=1}^{k} A^j$;

4) no constraints on the clusters $A^j, \ j = 1, \ldots, k$.

Cluster analysis via nonsmooth optimization

$$\text{minimize} \quad \psi(x,w) = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k} w_{ij}\|x^j - a^i\|^2 \tag{1}$$

subject to

$$x = (x^1, \ldots, x^k) \in \mathbb{R}^{n\times k}, \tag{2}$$

$$\sum_{j=1}^{k} w_{ij} = 1, \;\; i = 1, \ldots, m, \tag{3}$$

and

$$w_{ij} = 0 \text{ or } 1, \;\; i = 1, \ldots, m, \;\; j = 1, \ldots, k \tag{4}$$

where w_{ij} is the association weight of pattern a^i with cluster j, given by

Cluster analysis via nonsmooth optimization

$$w_{ij} = \begin{cases} 1 & \text{if pattern } a^i \text{ is allocated to cluster } j \\ 0 & \text{otherwise} \end{cases}$$

and

$$x^j = \frac{\sum_{i=1}^{m} w_{ij} a^i}{\sum_{i=1}^{m} w_{ij}}, \quad j = 1, \ldots, k.$$

Cluster analysis via nonsmooth optimization

$$\text{minimize} \quad f_k(x^1, \ldots, x^k) \quad \text{subject to } x^i \in \mathbb{R}^n, \ i = 1, \ldots, k,$$

where

$$f_k(x^1, \ldots, x^k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} \|x^j - a^i\|^2.$$

(Bock, 1974, Bagirov, Rubinov, Sukhorukova and Yearwood, TOP, 2003, Bagirov and Yearwood, EJOR, 2006)

Cluster analysis via nonsmooth optimization

- nonsmooth, nonconvex optimization problem; the number of local minimizers increases exponentially as the number of clusters increase;

- The number of variables increases substantially as the number of clusters increases and the problem becomes large scale ($n \times k$).

k-plane clustering

Collection of hyperplanes: $\{x^i, y_i\}, \quad x^i \in \mathbb{R}^n, \ y_i \in \mathbb{R}, \ i = 1, \ldots, k.$

$$\text{minimize} \quad f_k(X, Y),$$

subject to $X = (x^1, \ldots, x^k) \in \mathbb{R}^{k \times n}$, $Y = (y_1, \ldots, y_k) \in \mathbb{R}^k$,

$$\|x^i\|^2 = 1, \ \ i = 1, \ldots, k.$$

Here

$$f_k(X, Y) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} \left(\langle x^j, a^i \rangle - y_j \right)^2.$$

Data classification via nonsmooth optimization

Let A and B be given disjoint sets containing m and p n-dimensional vectors, respectively:

$$A = \{a^1, \ldots, a^m\}, a^i \in \mathbb{R}^n, j = 1, \ldots, m,$$
$$B = \{b^1, \ldots, b^p\}, b^j \in \mathbb{R}^n, j = 1, \ldots, p.$$

Linear separability

The sets A and B are linearly separable if there exists a hyperplane $\{x, y\}$, with $x \in \mathbb{R}^n,\ y \in \mathbb{R}^1$ such that

1) for any $j = 1, \ldots, m$

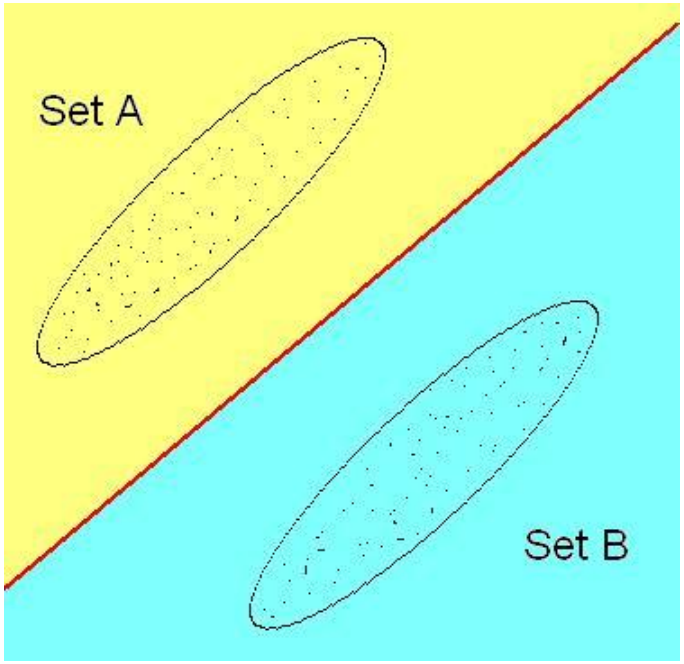
$$\langle x, a^j \rangle - y < 0,$$

2) for any $k = 1, \ldots, p$

$$\langle x, b^k \rangle - y > 0.$$

Here $\langle \cdot, \cdot \rangle$ is an inner product in $\mathbb{R}^n$.

Data classification via nonsmooth optimization



Data classification via nonsmooth optimization

The sets A and B are linearly separable if and only if $\text{co}\, A \bigcap \text{co}\, B = \emptyset$. Here co denotes the convex hull of a set.

$$\text{minimize} \quad f(x,y) \quad \text{subject to} \quad (x,y) \in \mathbb{R}^n \times \mathbb{R}^1 \tag{5}$$

where

$$f(x,y) = \frac{1}{m}\sum_{i=1}^{m} \max\left(0, \langle x, a^i\rangle - y + 1\right) + \frac{1}{p}\sum_{j=1}^{p} \max\left(0, -\langle x, b^j\rangle + y + 1\right)$$

is an error function.

Data classification via nonsmooth optimization

The problem (5) is equivalent to the following linear program:

$$\text{minimize} \quad \frac{1}{m}\sum_{i=1}^{m} t_i + \frac{1}{p}\sum_{j=1}^{p} z_j$$

subject to

$$t_i \geq \langle x, a^i \rangle - y + 1, \quad i = 1, \ldots, m,$$

$$z_j \geq -\langle x, b^j \rangle + y + 1, \quad j = 1, \ldots, p,$$

$$t \geq 0, \quad z \geq 0,$$

where t_i is nonnegative and represents the error for the point $a^i \in A$ and z_j is nonnegative and represents the error for the point $b^j \in B$.

The sets A and B are linearly separable if and only if $f^* = f(x^*, y_*) = 0$. The trivial solution $x = 0$ cannot occur.

Data classification via nonsmooth optimization

Polyhedral separability (Astorino and Gaudioso, 2002)
The sets A and B are h-polyhedrally separable if there exists a set of h hyperplanes $\{x^i, y_i\}$, with $x^i \in \mathbb{R}^n,\ y_i \in \mathbb{R}^1,\ i = 1, \ldots, h$ such that

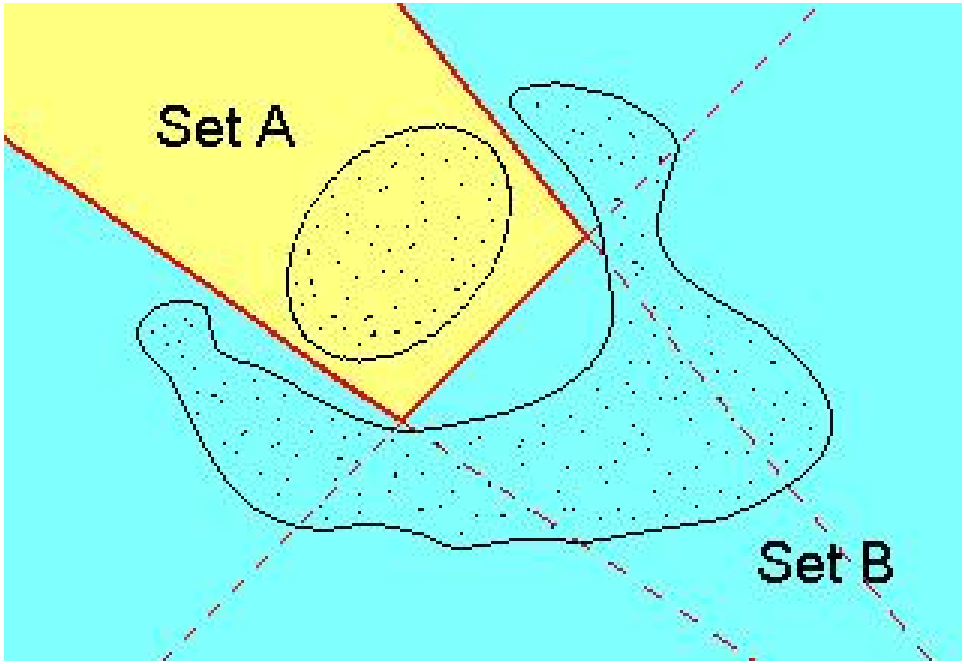
1) for any $j = 1, \ldots, m$ and $i = 1, \ldots, h$

$$\langle x^i, a^j \rangle - y_i < 0,$$

2) for any $k = 1, \ldots, p$ there exists at least one $i \in \{1, \ldots, h\}$ such that

$$\langle x^i, b^k \rangle - y_i > 0.$$

Data classification via nonsmooth optimization



Data classification via nonsmooth optimization

The sets A and B are h-polyhedrally separable, for some $h \leq p$ if and only if

$$\text{co}\, A \bigcap B = \emptyset.$$

$$\text{minimize} \quad f(x,y) \quad \text{subject to} \quad (x,y) \in \mathbb{R}^{hn} \times \mathbb{R}^h \tag{6}$$

where

$$f(x,y) = \frac{1}{m} \sum_{j=1}^{m} \max \left[0, \max_{1 \leq i \leq h} \left\{ \langle x^i, a^j \rangle - y_i + 1 \right\} \right] +$$

$$\frac{1}{p} \sum_{k=1}^{p} \max \left[0, \min_{1 \leq i \leq h} \left\{ -\langle x^i, b^k \rangle + y_i + 1 \right\} \right]$$

is an error function.

Data classification via nonsmooth optimization

- This function is a nonconvex piecewise linear function.
- $x^i = 0,\ i = 1, \ldots, h$ cannot be the optimal solution.
- Let $(\bar{x}, \bar{y}) = (\{\bar{x}^i, \bar{y}_i\}, i = 1, \ldots, h)$ be a global solution to the problem (6). The sets A and B are h-polyhedrally separable if and only if $f(\bar{x}, \bar{y}) = 0$.
- If there exists a nonempty set $\bar{I} \subset \{1, \ldots, h\}$ such that $x^i = 0,\ i \in \bar{I}$, then the sets A and B are $(h - |\bar{I}|)$-polyhedrally separable.

Data classification via nonsmooth optimization

Max-min separability (Bagirov, 2005)
Consider a collection of hyperplanes

$$H = \left\{ \{x^{ij}, y_{ij}\},\ j \in J_i,\ i \in I \right\},$$

where

$$x^{ij} \in \mathbb{R}^n,\ y_{ij} \in \mathbb{R}^1,\ j \in J_i,\ i \in I$$

and

$$I = \{1, \ldots, l\},\ l > 0,\ J_i \neq \emptyset\ \forall i \in I.$$

Data classification via nonsmooth optimization

This collection of hyperplanes defines the following max-min function on $\mathbb{R}^n$:

$$\varphi(z) = \max_{i \in I} \min_{j \in J_i} \left\{ \langle x^{ij}, z \rangle - y_{ij} \right\}, \ z \in \mathbb{R}^n.$$

Here $\langle \cdot, \cdot \rangle$ is an inner product in $\mathbb{R}^n$.

Data classification via nonsmooth optimization

The sets A and B are max-min separable if there exist a finite number of hyperplanes $\{x^{ij}, y_{ij}\}$ with $x^{ij} \in \mathbb{R}^n,\ y_{ij} \in \mathbb{R}^1,\ j \in J_i,\ i \in I$ such that

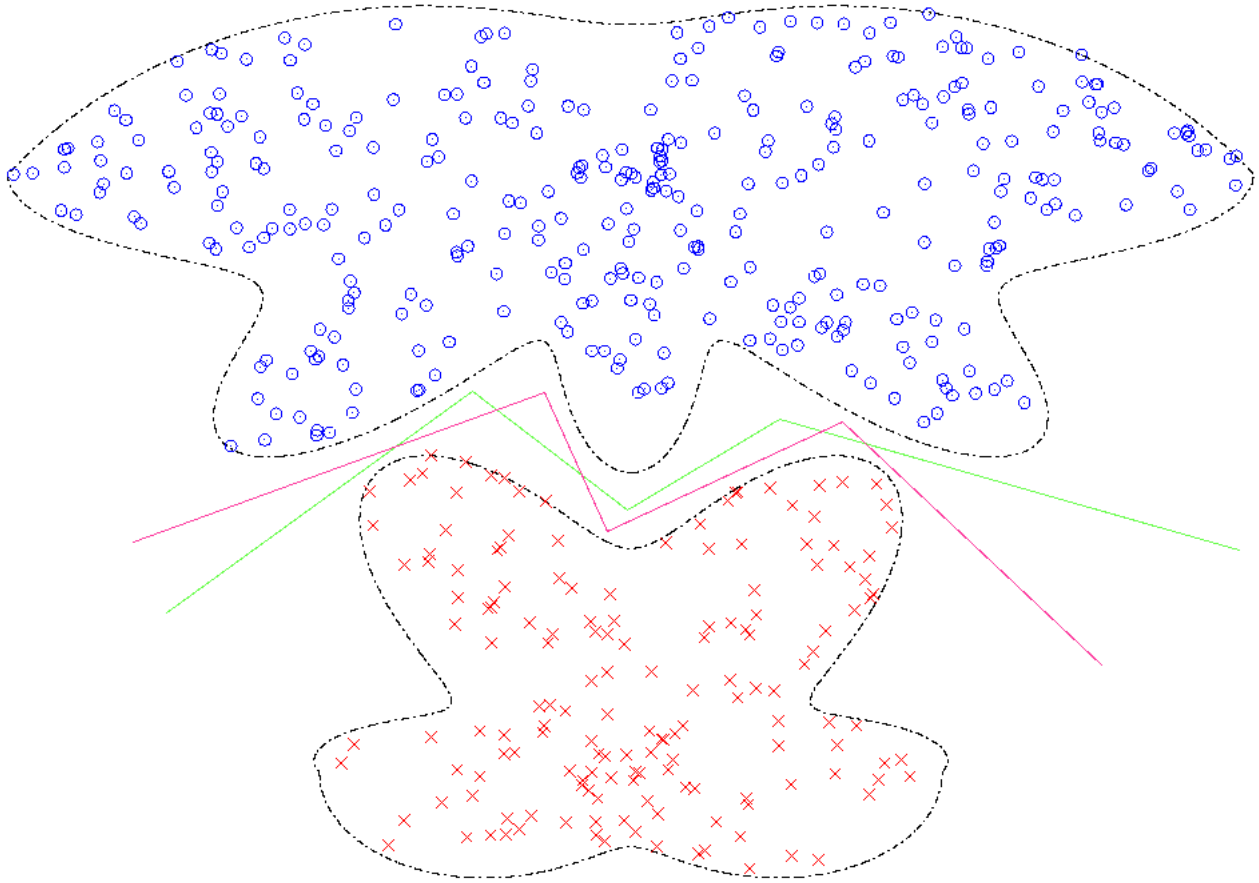
1. for all $i \in I$ and $a \in A$

$$\min_{j \in J_i} \left\{ \langle x^{ij}, a \rangle - y_{ij} \right\} < 0;$$

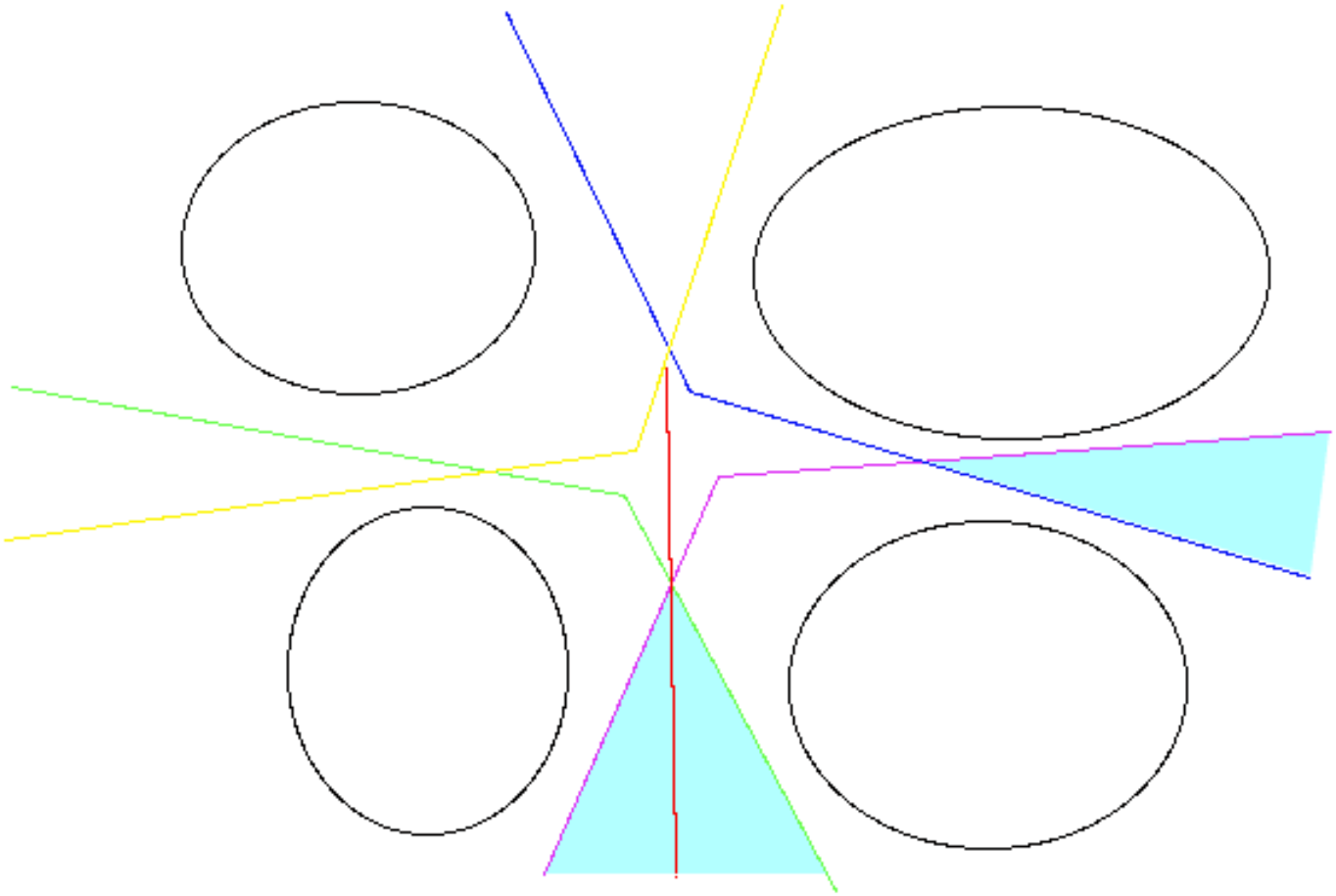
2. for any $b \in B$ there exists at least one $i \in I$ such that

$$\min_{j \in J_i} \left\{ \langle x^{ij}, b \rangle - y_{ij} \right\} > 0.$$

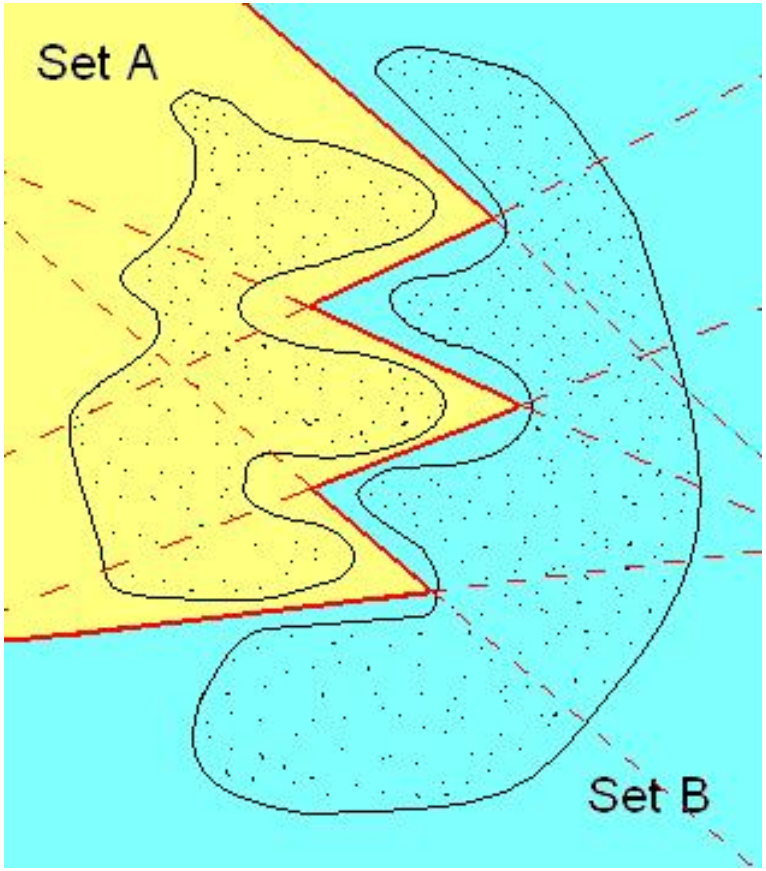
Data classification via nonsmooth optimization



Data classification via nonsmooth optimization



Data classification via nonsmooth optimization



Data classification via nonsmooth optimization

- The sets A and B are max-min separable then $\varphi(a) < 0$ for any $a \in A$ and $\varphi(b) > 0$ for any $b \in B$. Thus the sets A and B can be separated by a function represented as a max-min of linear functions. Therefore this kind of separability is called a max-min separability.

- The notions of max-min and piecewise linear separability are equivalent.

- The sets A and B are max-min separable if and only if they are disjoint: $A \bigcap B = \emptyset$, that is any two disjoint finite point sets are max-min separable.

Data classification via nonsmooth optimization

An averaged error function:

$$f(X,Y) = f_1(X,Y) + f_2(X,Y)$$

$$f_1(X,Y) = (1/m)\sum_{k=1}^{m} \max\left[0, \max_{i\in I}\min_{j\in J_i}\left\{\langle x^{ij}, a^k\rangle - y_{ij} + 1\right\}\right],$$

$$f_2(X,Y) = (1/p)\sum_{t=1}^{p} \max\left[0, \min_{i\in I}\max_{j\in J_i}\left\{-\langle x^{ij}, b^t\rangle + y_{ij} + 1\right\}\right],$$

$$X = (x^{11}, \ldots, x^{lq_l}) \in \mathbb{R}^{nL}, \;\; Y = (y_{11}, \ldots, y_{lq_l}) \in \mathbb{R}^{L},$$

$$L = \sum_{i\in I} q_i, \;\; q_i = |J_i|, \;\; i \in I = \{1, \ldots, l\}$$

Data classification via nonsmooth optimization

- $f(X,Y) \geq 0$ for all $X \in \mathbb{R}^{nL}$ and $Y \in \mathbb{R}^{L}$;
- The sets A and B are max-min separable if and only if there exists a set of hyperplanes $\{x^{ij}, y_{ij}\}, j \in J_i,\ i \in I = \{1,\ldots,l\}$ such that $f(X,Y) = 0$.
- $X = 0 \in \mathbb{R}^{nL}$ cannot be an optimal solution.

minimize $f(X,Y)$ subject to $(X,Y) \in \mathbb{R}^{(n+1)L}$

(**Bagirov, Optimization Methods and Software, 2005**)

Data classification via nonsmooth optimization

- nonsmooth, nonconvex optimization problem; the number of local minimizers increases exponentially as the number of hyperplanes increases;

- The number of variables increases substantially as the number of hyperplanes increases and the problem becomes large scale $((n+1) \times L)$.

Data classification via nonsmooth optimization

Ellipsoidal separability. (Gaudioso, 2005)
In particular, the set A is ellipsoidally separable from B if and only if there exists an ellipsoid

$$E(Q) = \left\{ x \in \mathbb{R}^n : \ (x - x^0)^T Q (x - x^0) \leq 1 \right\}$$

centered in $x^0 \in \mathbb{R}^n$, with $Q \in R^{n \times n}$ symmetric and positive definite, such that

$$(a^i - x^0)^T Q (a^i - x^0) \leq 1, \ \ i = 1, \ldots, m$$

$$(b^j - x^0)^T Q (b^j - x^0) > 1, \ \ j = 1, \ldots, p.$$

Such an ellipsoid might exist only if the intersection of the convex hull of A with the set B is empty.

Data classification via nonsmooth optimization

If the center x^0 is fixed, the ellipsoidal separation of A and B can be obtained by minimizing the following error function:

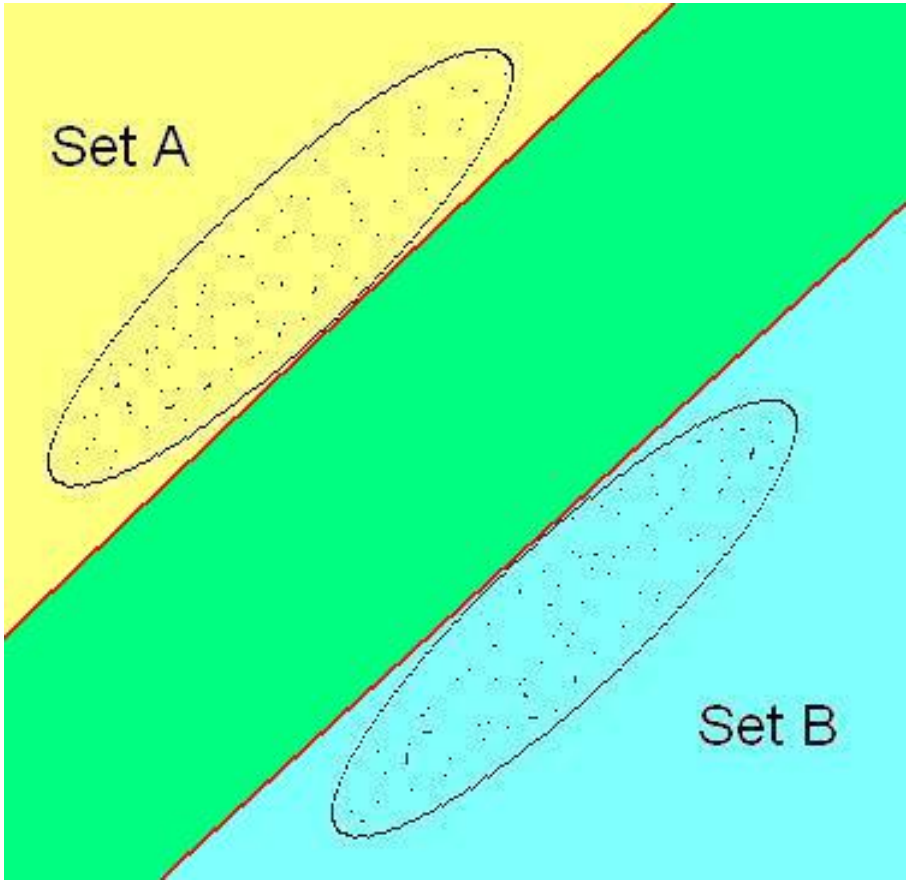
$$f(Q) = \sum_{i=1}^{m} \max\left\{0, (a^i - x^0)^T Q (a^i - x^0) - 1\right\} + \sum_{j=1}^{p} \max\left\{0, 1 - (b^j - x^0)^T Q (b^j - x^0)\right\}.$$

Support vector machines and nonsmooth optimization

The output of an SVM model is a hyperplane staying in the middle between two parallel hyperplanes, each of them supporting, respectively, one set. The distance between these two parallel hyperplanes is called the margin and it is a measure of the ability to correctly classify a new sample point (generalization capability). In particular, the optimal separation hyperplane is constructed by minimizing the following error function:

$$f(x,y) = \frac{1}{2}\|x\|^2 + C\sum_{i=1}^{m} \max\left(0, \langle x, a^i\rangle - y + 1\right) + C\sum_{j=1}^{p} \max\left(0, -\langle x, b^j\rangle + y + 1\right).$$

Data classification via nonsmooth optimization



Support vector machines and nonsmooth optimization

Although the objective function above is not differentiable, its minimization is equivalent to the following quadratic program:

$$\text{minimize} \quad \frac{1}{2}\|x\|^2 + C\left(\sum_{i=1}^{m} t_i + \sum_{j=1}^{p} z_j\right)$$

subject to

$$t_i \geq \langle x, a^i \rangle - y + 1, \;\; i = 1, \ldots, m,$$

$$z_j \geq -\langle x, b^j \rangle + y + 1, \;\; j = 1, \ldots, p,$$

$$t_i \geq 0, \;\; i = 1, \ldots, m, \;\; z_j \geq 0, \;\; j = 1, \ldots, p.$$

Support vector machines and nonsmooth optimization

Transductive Support Vector Machines (Astorino and Fuduli, 2007)

Given the two sets of labelled points A and B (training set) a set (test set) of q unlabelled points in $\mathbb{R}^n$

$$\{c^1, \ldots, c^q\}$$

the problem of finding a hyperplane far away from both the labelled and unlabelled points can be formulated as follows:

$$\text{minimize} \quad \frac{1}{2}\|x\|^2 + C_1 \left(\sum_{i=1}^{m} t_i + \sum_{j=1}^{p} z_j \right) + C_2 \sum_{k=1}^{q} r(\langle x, c^k \rangle - y)$$

where r is the margin penalty function.

$$r(t) = \begin{cases} 1 - |t| & \text{for } -1 \leq t \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Support vector machines and nonsmooth optimization

$$\text{minimize} \quad h(x,y), \quad \text{subject} \quad \text{to} \quad x \in \mathbb{R}^n, \ y \in R$$

where

$$h(x,y) = \frac{1}{2}\|x\|^2 + C_1 \left(\sum_{i=1}^{m} \max\{0, \langle x, a^i \rangle - y + 1\} + \sum_{j=1}^{p} \max\{0, -\langle x, b^j \rangle + y + 1\} \right)$$

$$+ C_2 \sum_{k=1}^{q} \max\{0, 1 - |\langle x, c^k \rangle - y|\}.$$

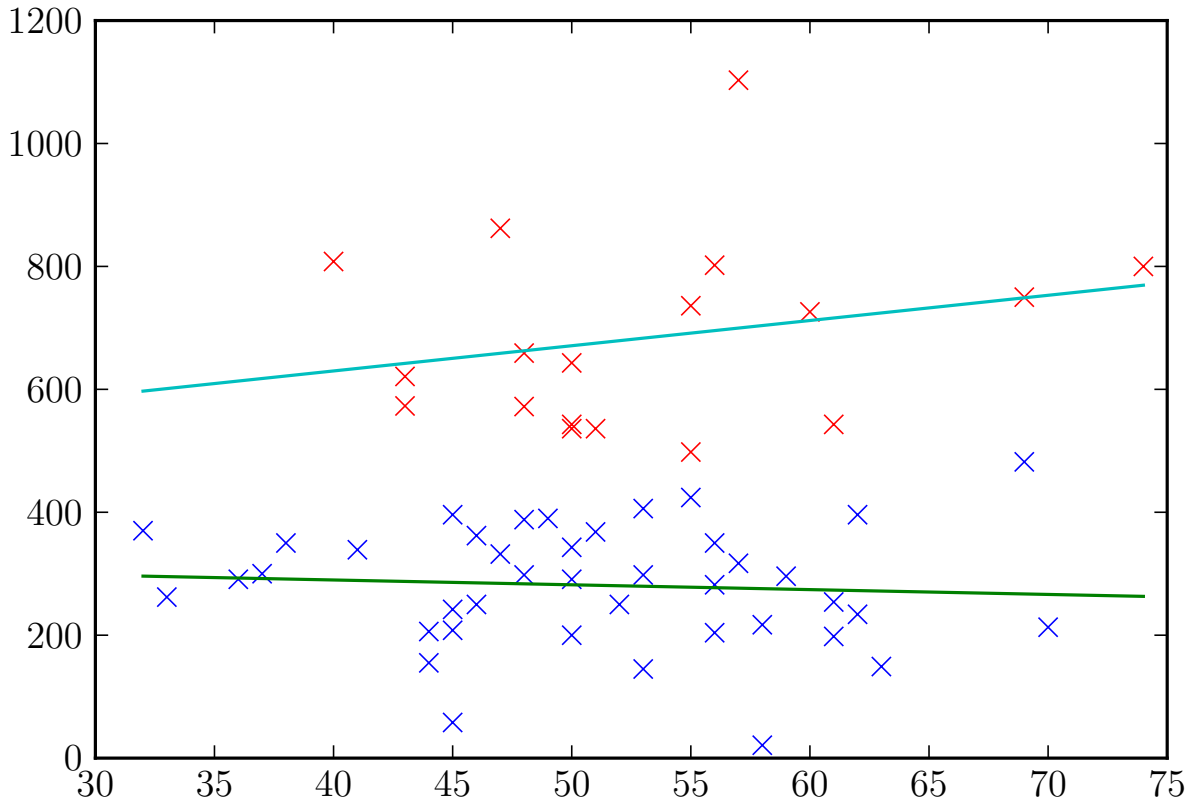
Clusterwise regression via nonsmooth optimization

Unsupervised classification, or clustering consists in finding subsets of similar points in a data set, in order to find patterns in the data. Regression analysis consists in fitting a function (often linear) to the data to discover how one or more variables vary as a function of another.

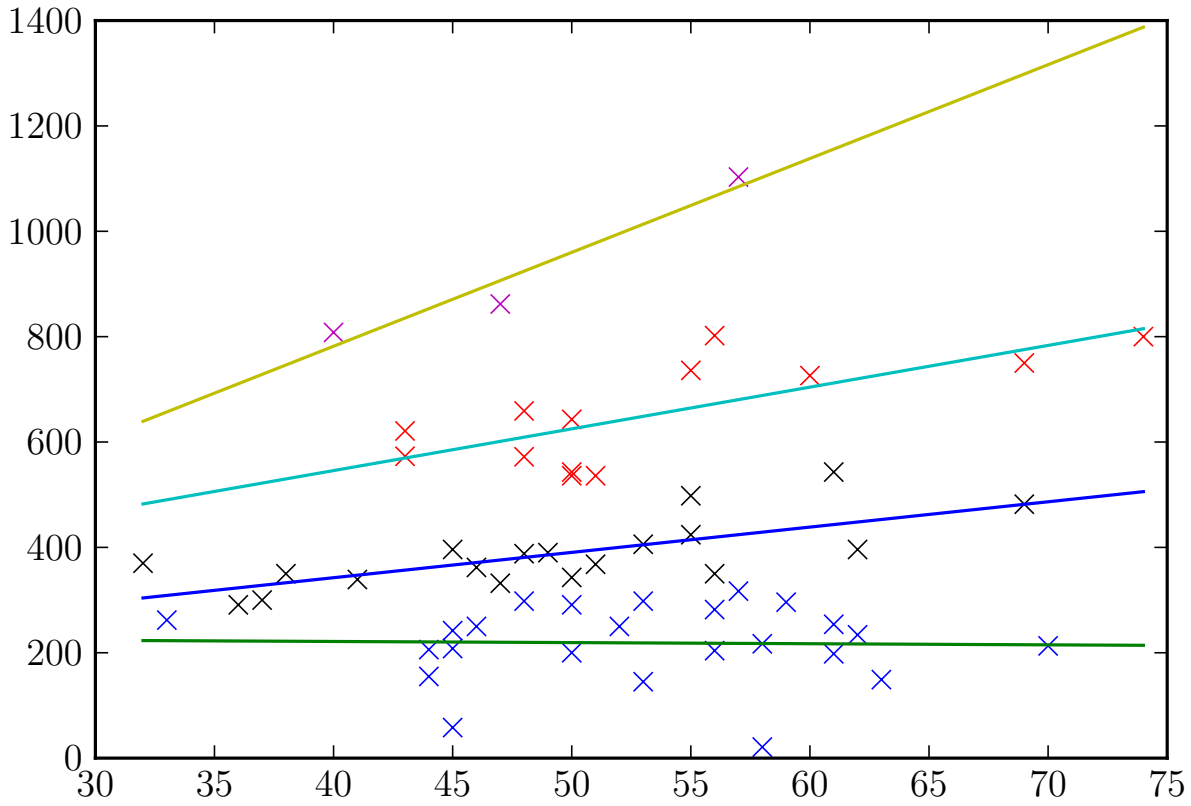
Clusterwise regression via nonsmooth optimization

The aim of clusterwise regression is to combine both of these techniques, to discover trends within data, when more than one trend is likely to exist. Clusterwise regression has applications for instance in market segmentation, where it allows one to gather information on customer behaviors for several unknown groups of customers. It is also applied to investigate the stock-exchange data and the so-called benefit segmentation. The presence of nonlinear relationships, heterogeneous subjects, or time series in these applications necessitate the use of two or more regression functions to best summarize the underlying structure of the data.

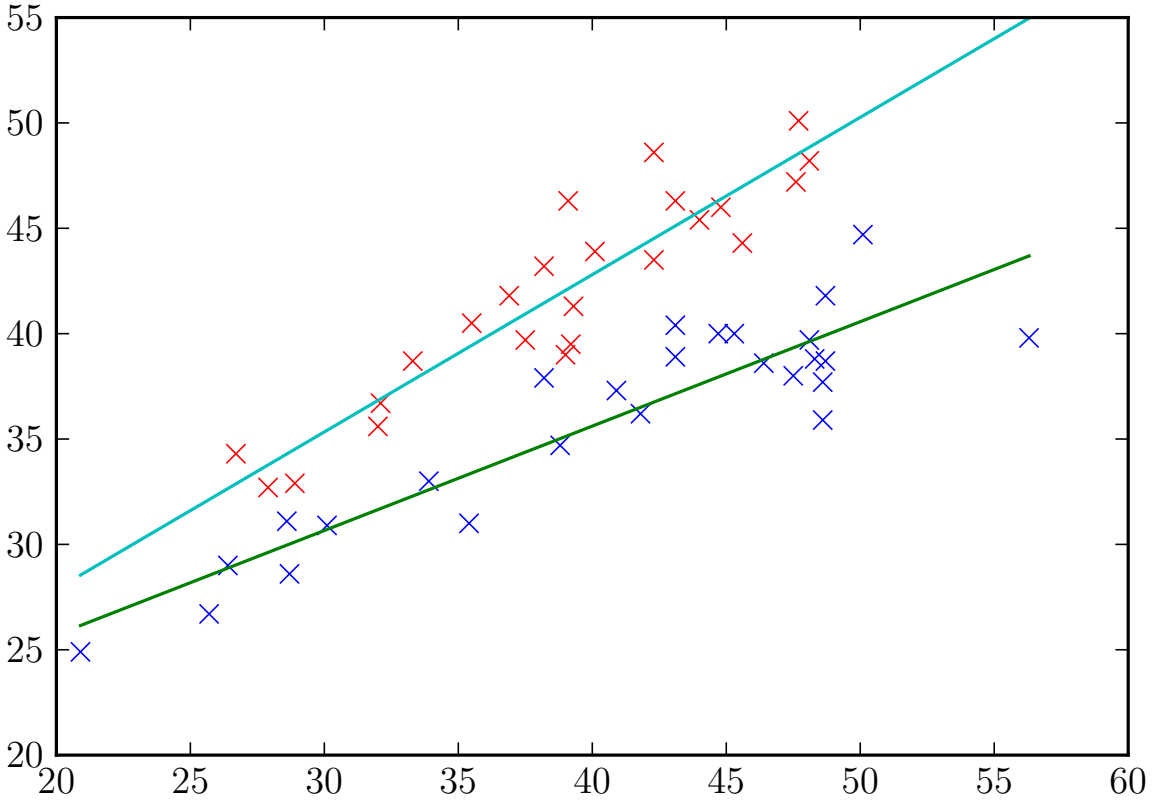
Clusterwise regression via nonsmooth optimization



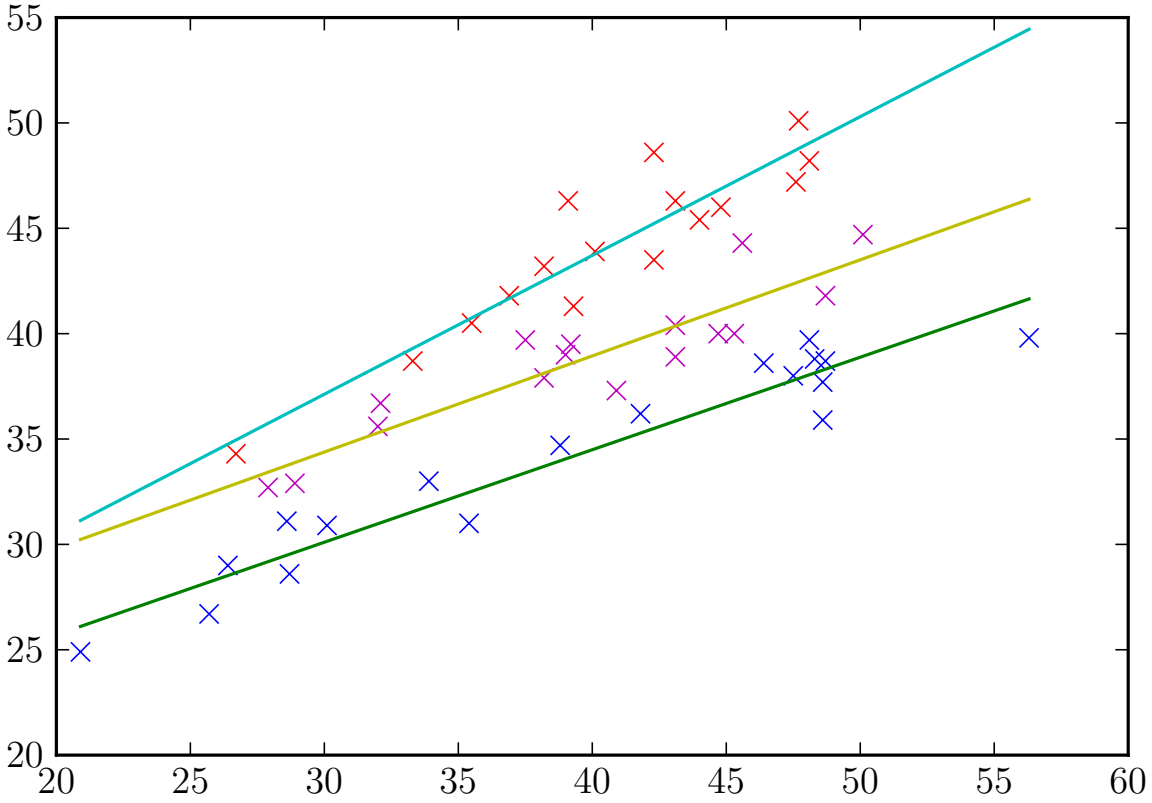
Clusterwise regression via nonsmooth optimization



Clusterwise regression via nonsmooth optimization



Clusterwise regression via nonsmooth optimization



Clusterwise regression via nonsmooth optimization

Given a data set $A = \{(a^i, b_i) \in \mathbb{R}^n \times \mathbb{R} : \ i = 1, \ldots, m\}$, the aim of the clusterwise linear regression is to find simultaneously an optimal partition of data in k clusters and regression coefficients $\{x^j, y_j\}, \ j = 1, \ldots, k$ within clusters in order to minimize the overall fit. Let $A^j, \ j = 1, \ldots, k$ be clusters such that

$$A^j \neq \emptyset, \ \ A^j \bigcap A^t = \emptyset, \ \ j, t = 1, \ldots, k, \ \ t \neq j \quad \text{and} \quad A = \bigcup_{j=1}^{k} A^j.$$

Clusterwise regression via nonsmooth optimization

Let $\{x^j, y_j\}$ be linear regression coefficients computed using only data points from the cluster A^j, $j = 1, \ldots, k$. Then for the given data point (a^i, b_i) and coefficients $\{x^j, y_j\}$ the regression error $h(x^j, y_j, a^i, b_i)$ is:

$$h(x^j, y_j, a^i, b_i) = \left| \langle x^j, a^i \rangle + y_j - b_i \right|^p .$$

Here $p > 0$. We associate a data point with the cluster whose regression error at this point is smallest. Then the overall fit function is:

$$f_k(x, y) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} h(x^j, y_j, a^i, b_i),$$

where $x = (x^1, \ldots, x^k) \in \mathbb{R}^{k \times n}$ and $y \in \mathbb{R}^k$.

Clusterwise regression via nonsmooth optimization

Thus the k-clusterwise linear regression problem is formulated as follows:

$$\text{minimize} \quad f_k(x,y) \quad \text{subject to} \quad x \in \mathbb{R}^{k\times n},\ y \in \mathbb{R}^k.$$

In general, the objective function f_k in this problem is nonsmooth nonconvex. One can consider any positive values of p to define regression errors. If $p = 1$ then the function f_k is piecewise linear and if $p = 2$ then it is piecewise quadratic. Moreover, if $k = 1$ then in both cases the objective function is convex and if $k > 1$ it becomes nonconvex.

(**Bagirov, Ugon and Mirzayeva, EJOR, 2013)**

Methods of nonsmooth optimization

Subdifferential:

$$\partial f(x) = \text{co}\left\{v \in \mathbb{R}^n : \exists\{x^k\} \subset D(f), x = \lim_{k\to\infty} x^k \text{ and } v = \lim_{k\to\infty} \nabla f(x^k)\right\}.$$

- Bundle methods
- Discrete gradient - derivative free method
- Smoothing techniques
- Stochastic algorithms.

Methods of nonsmooth optimization

Subdifferential:

Bundle methods use convex piecewise linear underestimators to find descent directions (Lemarechal, Kiwiel, Gaudioso, Mifflin, Wolfe, Zowe, Mäkelä, Karmitsa and others).

Discrete gradient uses approximations to subgradients to find descent directions (Bagirov)

Smoothing techniques replaces component maximum and minimum functions by their smooth approximations (Polak, Xavier, Bagirov).

Methods of nonsmooth optimization

Let f be a scalar function defined on an open set $D_0 \subseteq \mathbb{R}^n$ containing a closed set $D \subseteq \mathbb{R}^n$.

The function f is called partially separable if there exists a family of $n \times n$ diagonal matrices $U_i,\ i = 1, \ldots, M$ such that the function f can be represented as follows:

$$f(x) = \sum_{i=1}^{M} f_i(U_i x).$$

The matrices U_i are binary, that is they contain only 0 and 1.

The function f is said to be piecewise partially separable if there exists a finite family of closed sets $D_1, \ldots, D_m$ such that

$$\bigcup_{i=1}^{m} D_i = D$$

and the function f is partially separable on each set $D_i,\ i = 1, \ldots, m$.

Incremental approach

Incremental learning algorithms are becoming increasingly popular in supervised and unsupervised data classification. This type of approach breaks up the data set into observations that can be classified using simple separators, and observations that require more elaborate ones. This allows one to simplify the learning task by eliminating the points that can be more easily classified.

Incremental approach

Furthermore, at each iteration, information gathered during prior iterations can be exploited. In the case of piecewise linear classifiers, this approach allows us to compute as few hyperplanes as needed to separate the sets, without any prior information. Additionally, this approach allows us to reach a near global solution of error functions by using solutions obtained at a given iteration as a starting point for the next iteration. Thus it reduces computational effort and avoids possible overfitting.

Incremental approach

- Objective functions, considered above, are piecewise partially separable. The use of this property may significantly reduce the complexity of the problems;

- One can develop incremental algorithms to reduce complexity and to find at least "near" global solution;

Incremental algorithms for clustering

Modified global k-means algorithm (Bagirov, 2008)

Assume that $k > 1$ and the cluster centers $x^1, \ldots, x^{k-1}$ for $(k-1)$-partition problem are known.

$$d^i_{k-1} = \min\left\{\|x^1 - a^i\|^2, \ldots, \|x^{k-1} - a^i\|^2\right\}.$$

$$\bar{f}_k(y) = \frac{1}{m}\sum_{i=1}^{m} \min\left\{d^i_{k-1}, \|y - a^i\|^2\right\}.$$

The function $\bar{f}_k$ is called an *auxiliary cluster function*.

$$\bar{f}_k(y) = f_k(x^1, \ldots, x^{k-1}, y).$$

Incremental algorithms for clustering

Consider the sets

$$\overline{D} = \left\{ y \in \mathbb{R}^n : \|y - a^i\|^2 \geq d^i_{k-1} \right\}, \quad D_0 = \mathbb{R}^n \setminus \overline{D}.$$

For any $y \in D_0$ consider the following sets:

$$S_1(y) = \left\{ a^i \in A : \|y - a^i\|^2 = d^i_{k-1} \right\},$$

$$S_2(y) = \left\{ a^i \in A : \|y - a^i\|^2 < d^i_{k-1} \right\},$$

$$S_3(y) = \left\{ a^i \in A : \|y - a^i\|^2 > d^i_{k-1} \right\}.$$

The set $S_2(y) \neq \emptyset$ for any $y \in D_0$.

Incremental algorithms for clustering

Algorithm 1 An algorithm for finding a starting point.

Step 1. For each $a^i \in D_0 \bigcap A$ compute the set $S_2(a^i)$, its center c^i and $\bar{f}_{k,a^i} = \bar{f}_k(c^i)$.

Step 2. Compute

$$\bar{f}_{k,min} = \min_{a^i \in D_0 \bigcap A} \bar{f}_{k,a^i},$$

$$a^j = \text{argmin}_{a^i \in D_0 \bigcap A} \bar{f}_{k,a^i},$$

the corresponding center c^j and the set $S_2(c^j)$.

Step 3. Recompute the set $S_2(c^j)$ and its center until no more data points escape or return to this cluster.

Incremental algorithms for clustering

Algorithm 2 An incremental algorithm for clustering problems.

Step 1. (Initialization). Select $\varepsilon > 0$. Compute the center $x^1 \in \mathbb{R}^n$ of the set A. Let f^1 be the corresponding value of the clustering function. Set $k = 1$.

Step 2. (Computation of the next cluster center). Set $k = k + 1$. Let $x^1, \ldots, x^{k-1}$ be the cluster centers for $(k-1)$-partition problem. Apply Algorithm 1 to find a starting point $\bar{y} \in \mathbb{R}^n$ for the k-th cluster center.

Step 3. (Refinement of all cluster centers). Select $(x^1, \ldots, x^{k-1}, \bar{y})$ as a new starting point, apply k-means algorithm to solve k-partition problem. Let $y^1, \ldots, y^k$ be a solution to this problem and f^k be the corresponding value of the clustering function.

Step 4. (Stopping criterion). If

$$\frac{f^{k-1} - f^k}{f^1} < \varepsilon$$

then stop, otherwise set $x^i = y^i,\ i = 1, \ldots, k$ and go to Step 2.

Incremental algorithms for clustering

Data sets	Number of instances	Number of attributes
Shuttle control	58000	9

We implemented the algorithm in Lahey Fortran 95 on PC Intel CPU 1.83 GHz and RAM 1 GB.

Incremental algorithm for clustering

k	KM		GKM		MGKM	
	E	CPU	E	CPU	E	CPU
2			0.00	212.19	0.00	0.36
10			0.02	1980.22	0.00	103.39
20			0.00	4286.66	0.00	546.95
30			0.00	6591.28	0.00	1310.20
40			1.66	8644.75	0.00	1651.88
50			0.93	10436.01	0.00	1998.30
60			0.00	12281.36	1.21	2397.91
80			0.00	15930.31	0.45	2957.89
100			0.10	19538.72	0.00	3536.83

Incremental algorithms for data classification

A piecewise linear classifier requires very low memory usage while providing real-time classification. This makes it suitable for many applications where many classifiers cannot be used. These applications include small reconnaissance robots, autonomous mobile robots, intelligent cameras, imbedded and real-time systems, portable devices, industrial vision systems, automated visual surveillance systems, monitoring systems.

Incremental algorithms for data classification

(Bagirov, Ugon and Webb, PAAA and IS, 2011)

Data set A with $q \geq 2$ classes $A_1, \ldots, A_q$ is given.

$$\alpha_j^i = \min_{a \in A_i} a_j, \quad \beta_j^i = \max_{a \in A_i} a_j$$

and define vectors $\alpha^i = (\alpha_1^i, \ldots, \alpha_n^i)$, $\beta^i = (\beta_1^i, \ldots, \beta_n^i)$, $i = 1, \ldots, q$. We define the following hyperboxes:

$$H(A_i) = [\alpha^i, \beta^i] \doteq \{x \in \mathbb{R}^n : \ \alpha_j^i \leq x_j \leq \beta_j^i, \ j = 1, \ldots, n\}, \ i = 1, \ldots, q.$$

Incremental algorithms for data classification

The intersection between two hyperboxes is also a hyperbox, which is obtained using the following:

$$H(A_i, A_j) = [c^{ij}, d^{ij}]$$

where

$$c_k^{ij} = \max\{\alpha_k^i, \alpha_k^j\}, \quad d_k^{ij} = \min\{\beta_k^i, \beta_k^j\}, \ \ k = 1, \ldots, n.$$

If $d_k^{ij} < c_k^{ij}$ for at least one $k \in \{1, \ldots, n\}$ then the hyperbox $H(A_i, A_j)$ is empty. The indeterminate region between i-th class and the rest of the data set is defined as:

$$D = \bigcup_{i=1,\ldots,q} \bigcup_{j=1,\ldots,q} H(A_i, A_j).$$

The set of obviously classified points in class A_i is:

$$O_i = A_i \setminus D.$$

Incremental algorithms for data classification

Let $\varepsilon_1 > 0, \varepsilon_2 > 0, \varepsilon_3 > 0$ be given tolerances and $\delta > 0$ be a sufficiently small number.

Algorithm 3 *An incremental algorithm*

Step 1. Initialization*) Set* $D^1 = D, \overline{A_u}^1 = \overline{A_u}, \underline{A_u}^1 = \underline{A_u}$*. Select any starting point* $(X^1, Y_1) = (x^1, y_1), x^1 \in \mathbb{R}^n, y_1 \in \mathbb{R}^1$ *. Set* $I_1 = 1, J_1^1 = 1, f_1 = f(x^1, y_1), r_1 = |I| = 1, s_1 = |J_1^1| = 1$*, the number of hyperplanes* $l = 1$ *and iteration counter* $k = 1$*.*

Step 2. *(*Computation of piecewise linear function*) Solve max-min separability problem over the set* D^k *starting from the point* $(X^k, Y_k) \in \mathbb{R}^{(n+1)l}$*. Let* (X^{k*}, Y_{k*}) *be solution to this problem,* f_k^* *corresponding objective function value, and* $f_{1,k}^*$ *and* $f_{2,k}^*$ *values of*

functions f_1 and f_2 respectively. Let E_k be the error rate at iteration k over the set D, that is

$$E_k = \frac{|\{a \in \overline{A_u} : \varphi^k(a) > 0\} \cup \{b \in \underline{A_u} : \varphi^k(b) < 0\}|}{|D|},$$

$$\varphi^k(a) = \max_{i \in I_k} \min_{j \in J_i^k} \left(\langle x^{ij*}, a \rangle - y_{ij*} \right).$$

Step 3. *(*The first stopping criterion*) If $f_k^* \leq 2\varepsilon_1$ then stop. (X^{k*}, Y_{k*}) is the final solution.*

Step 4. *(*The second stopping criterion*) If $k \geq 2$ and $f_{k-1}^* - f_k^* \leq \varepsilon_2$ then stop. $(X^{k-1,*}, Y_{k-1,*})$ is the final solution.*

Step 5. *(*The third stopping criterion*) If $E_k < \varepsilon_3$ then stop. (X^{k*}, Y_{k*}) is the final solution.*

Step 6. *(*Refinement of indeterminate regions*) Compute*

$$f_{k,max} = \max_{a \in \overline{A_u}^k} \max_{i \in I_k} \min_{j \in J_i^k} \left(\langle x^{ij*}, a \rangle - y_{ij*} \right),$$

$$f_{k,min} = \min_{a \in \underline{A_u}^k} \max_{i \in I_k} \min_{j \in J_i^k} \left(\langle x^{ij*}, a \rangle - y_{ij*} \right).$$

and the following sets:

$$C_1 = \left\{ a \in \overline{A_u}^k : \max_{i \in I_k} \min_{j \in J_i^k} \langle x^{ij*}, a \rangle - y_{ij*} \leq (1+\delta) f_{k,min} \right\},$$

$$C_2 = \left\{ a \in \underline{A_u}^k : \max_{i \in I_k} \min_{j \in J_i^k} \langle x^{ij*}, a \rangle - y_{ij*} \geq (1+\delta) f_{k,max} \right\}$$

Refine the indeterminate region as follows:

$$D^{k+1} = D^k \setminus \{C_1 \bigcup C_2\}, \quad \overline{A_u}^k = A_u \cap D^k, \quad \underline{A_u}^k = D^k \setminus A_u.$$

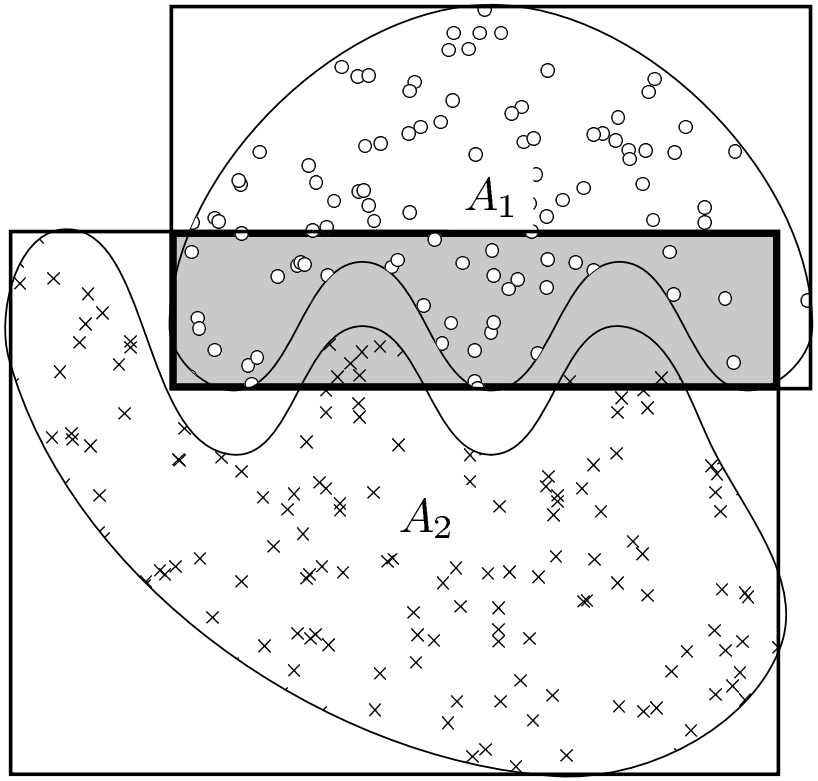
Step 7. *(*Adding new hyperplanes*)*

- *If* $f^*_{1,k} > \varepsilon_1$ *then set* $s_{k+1} = s_k + 1$, $J_i^{k+1} = J_i^k \cup \{s_{k+1}\}$ *for all* $i \in I_k$. *Set* $x^{ij} = x^{i,j-1,*}, y_{ij} = y_{i,j-1,*}$, $i \in I_k$, $j = s_{k+1}$.

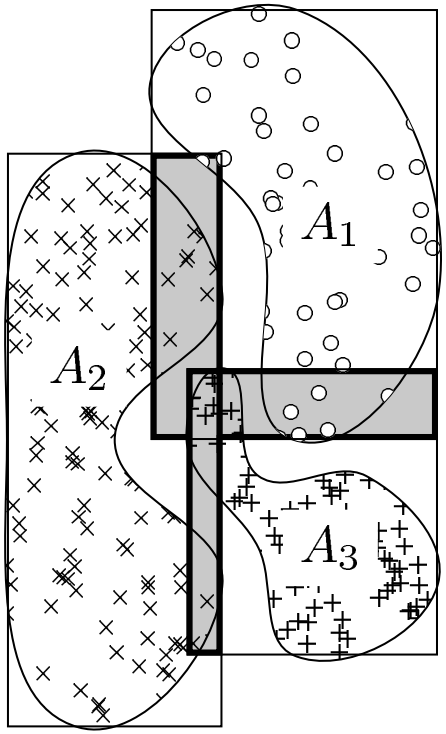
- *If* $f^*_{2,k} > \varepsilon_1$ *then set* $r_{k+1} = r_k + 1$, $I_{k+1} = I_k \cup \{r_{k+1}\}$, $J^{k+1}_{r_{k+1}} = J_{r_k}$. *Set* $x^{ij} = x^{i-1,j,*}, y_{ij} = y_{i-1,j,*}$, $i = r_{k+1}$, $j \in J^k_{r_k}$.

Step 8. *(*New starting point*) Set* $X^{k+1} = (X^{k*}, x_{ij},\ i \in I_{k+1},\ j \in J^i_{k+1})$, $Y_{k+1} = (Y_{k*}, y_{ij},\ i \in I_{k+1},\ j \in J^i_{k+1})$, $l = \sum_{i \in I_{k+1}} |J_i^{k+1}|$, $k = k+1$ *and go to Step 2.*

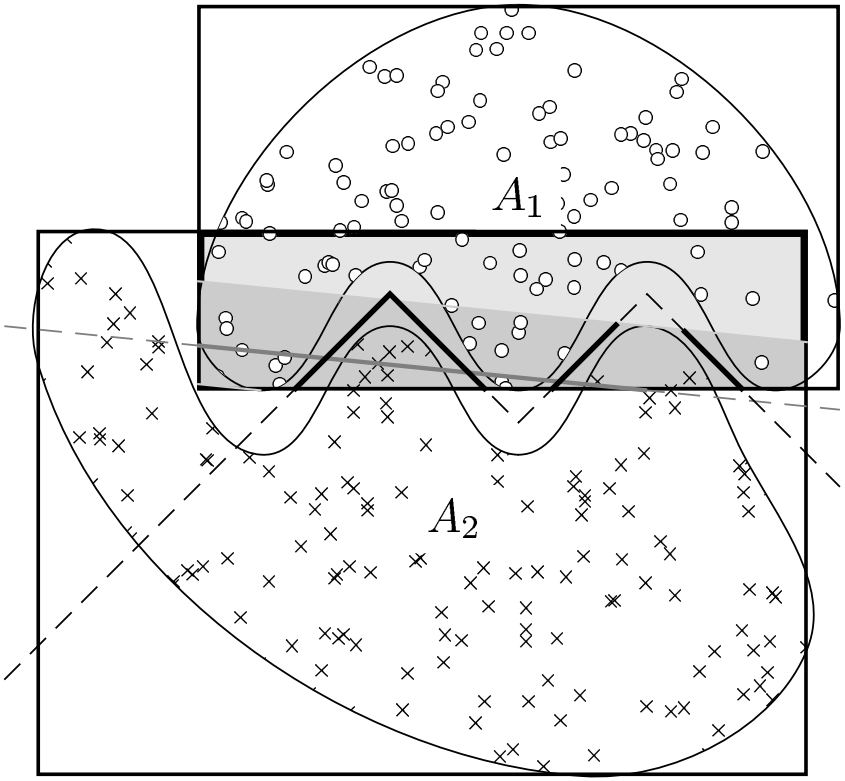
Incremental algorithm for data classification



Incremental algorithm for data classification



Incremental algorithm for data classification



Incremental algorithm for data classification

Data sets	(train,test)	No. of attributes	No. of classes
Shuttle control	(43500, 14500)	10	7

Incremental algorithm for data classification

Dataset	Shuttle	
Algorithm	Test set	CPU
NB(Kernel)	98.32	0.74
Logistic	96.83	4018.47
MLPerceptron	99.75	1017.70
SMO (NPOL)	96.81	9008.80
SMO (PUK)	99.50	3190.44
Polyhedral	99.29	1247.02
Incremental	99.90	1059.72
Hyperboxes	99.94	76.84

Conclusions

- nonsmooth optimization approaches to some data mining problems allow one to significantly reduce their complexity and develop efficient algorithm

- Data mining can benefit from many methodologies of the nonsmooth optimization, where problems of the min-max, min-max-min and min-sup type are tackled.

THANK YOU!