Optimization in Data Mining

Olvi L. Mangasarian

with

G. M. Fung, J. W. Shavlik, Y.-J. Lee, E.W. Wild

& Collaborators at ExonHit – Paris

University of Wisconsin – Madison

&

University of California- San Diego
Occam’s Razor
A Widely Held “Axiom” in Machine Learning & Data Mining

“Entities are not to be multiplied beyond necessity"
William of Ockham (English Philosopher & Theologian)
1287 Surrey - 1347 Munich

“Everything should be made as simple as possible, but not simpler”
Albert Einstein
1879 Munich - 1955 Princeton

“Simplest is Best”
What is Data Mining?

- Data mining is the process of analyzing data in order to extract useful knowledge such as:
  - Clustering of unlabeled data
    - Unsupervised learning
  - Classifying labeled data
    - Supervised learning
  - Feature selection
    - Suppression of irrelevant or redundant features

- Optimization plays a fundamental role in data mining via:
  - Support vector machines or kernel methods
    - State-of-the-art tool for data mining and machine learning
What is a Support Vector Machine?

- An optimally defined surface
- Linear or nonlinear in the input space
- Linear in a higher dimensional feature space
- Feature space defined by a linear or nonlinear kernel

\[ K(A, X) \rightarrow Y, \]
\[ A \in \mathbb{R}^{m \times n}, \ X \in \mathbb{R}^{n \times k}, \ \text{and} \ Y \in \mathbb{R}^{m \times k} \]
Principal Topics

- Data clustering as a concave minimization problem
  - K-median clustering and feature reduction
  - Identify class of patients that benefit from chemotherapy
- Linear and nonlinear support vector machines (SVMs)
  - Feature and kernel function reduction
- Enhanced knowledge-based classification
  - LP with implication constraints
- Generalized Newton method for nonlinear classification
  - Finite termination with or without stepsize
- Drug discovery based on gene macroarray expression
  - Identify class of patients likely to respond to new drug
- Multisurface proximal classification
  - Nonparallel classifiers via generalized eigenvalue problem
Clustering in Data Mining

General Objective

- **Given**: A dataset of $m$ points in $n$-dimensional real space
- **Problem**: Extract *hidden distinct properties* by clustering the dataset into $k$ clusters
Concave Minimization Formulation

1-Norm Clustering: k-Median Algorithm

- **Given:** Set $A$ of $m$ points in $\mathbb{R}^n$ represented by the matrix $A \in \mathbb{R}^{m \times n}$, and a number $k$ of desired clusters.

- **Find:** Cluster centers $C_1, \ldots, C_k \in \mathbb{R}^n$ that minimize the sum of 1-norm distances of each point: $A_1, A_2, \ldots, A_m$, to its closest cluster center.

- **Objective Function:** Sum of $m$ minima of $k$ linear functions, hence it is piecewise-linear concave.

- **Difficulty:** Minimizing a general piecewise-linear concave function over a polyhedral set is NP-hard.
Clustering via Finite Concave Minimization

- Minimize the sum of 1-norm distances between each data point \( A_i \) and the closest cluster center \( C_\ell \):

\[
\min_{C_\ell \in \mathbb{R}^n, \, D_{i\ell} \in \mathbb{R}^n} \sum_{i=1}^{m} \min_{\ell=1 \ldots, k} \{ e' D_{i\ell} \}
\]

s.t. \(- D_{i\ell} \leq A'_i - C_\ell \leq D_{i\ell},\)
\(i = 1, \ldots, m, \ell = 1, \ldots, k,\)

where \( e \) is a column vector of ones.
K-Median Clustering Algorithm
Finite Termination at Local Solution
Based on a Bilinear Reformulation

Step 0 (Initialization): Pick $k$ initial cluster centers

Step 1 (Cluster Assignment): Assign points to the cluster with the nearest cluster center in 1-norm

Step 2 (Center Update) Recompute location of center for each cluster as the cluster median (closest point to all cluster points in 1-norm)

Step 3 (Stopping Criterion) Stop if the cluster centers are unchanged, else go to Step 1

Algorithm terminates in a finite number of steps, at a local solution
Breast Cancer Patient Survival Curves With & Without Chemotherapy

Kaplan-Meier survival curves: NoChemo & Chemo

113 NoChemo Patients
140 Chemo Patients

p-value
NoChemo vs. Chemo: 0.0018
Survival Curves for 3 Groups: Good, Intermediate & Poor Groups
(Generated Using k-Median Clustering)
Survival Curves for Intermediate Group: Split by Chemo & NoChemo

Intermediate Kaplan–Meier survival curves: NoChemo & Chemo

- 44 NoChemo Patients
- 67 Chemo Patients

$p$-value

1. Chemo vs. 1. NoChemo: 0.0306
Feature Selection in k-Median Clustering

- Find a reduced number of input space features such that clustering in the reduced space closely replicates the clustering in the full dimensional space
Basic Idea

- Based on nondifferentiable optimization theory, make a simple but fundamental modification in the second step of the k-median algorithm.
- In each cluster, find a point closest in the 1-norm to all points in that cluster and to the zero median of ALL data points.
- Based on increasing weight given to the zero data median, more features are deleted from the problem.
- Proposed approach can lead to a feature reduction as high as 69%, with clustering comparable to within 4% to that with the original set of features.
3-Class Wine Dataset
178 Points in 13-dimensional Space
Support Vector Machines

- Linear & nonlinear classifiers using kernel functions
Support Vector Machines
Maximize the Margin between Bounding Planes

\[ x'w = \gamma + 1 \]

\[ x'w = \gamma - 1 \]

\[ \frac{2}{||w||} \]

A–

A+

\( x \)
Support Vector Machine
Algebra of 2-Category Linearly Separable Case

- Given $m$ points in $n$ dimensional space
- Represented by an $m$-by-$n$ matrix $A$
- Membership of each $A_i$ in class $+1$ or $-1$ specified by:
  - An $m$-by-$m$ diagonal matrix $D$ with $+1$ & $-1$ entries
- Separate by two bounding planes, $x'w = \gamma \pm 1$:
  $$A_iw \geq \gamma + 1, \quad \text{for} \quad D_{ii} = +1,$$
  $$A_iw \leq \gamma - 1, \quad \text{for} \quad D_{ii} = -1.$$
- More succinctly:
  $$D(Aw - e\gamma) \geq e,$$
where $e$ is a vector of ones.
Feature-Selecting 1-Norm Linear SVM

- 1-norm SVM:
  \[
  \begin{align*}
  \min_{y \geq 0, w, \gamma} & \quad \nu e'y + \|w\|_1 \\
  \text{s. t.} & \quad D(Aw - e\gamma) + y \geq e,
  \end{align*}
  \]
  where \(D_{ii} = \pm 1\) are elements of the diagonal matrix \(D\) denoting the class of each point \(A_i\) of the dataset matrix \(A\).

- Very effective in feature suppression
  - For example, 5 out of 30 cytological features are selected by the 1-norm SVM for breast cancer diagnosis with over 97% correctness.
  - In contrast, 2-norm and \(\infty\)-norm SVMs suppress no features.
1- Norm Nonlinear SVM

- **Linear SVM:** (Linear separating surface: \( x'w = \gamma \))

\[
\min_{y \geq 0, w, \gamma} \nu e'y + \| w \|_1 \\
\text{s.t.} \quad D(Aw - e\gamma) + y \geq e
\]

Change of variable \( w = A'Du \) and maximizing the margin in the “dual space”, gives:

\[
\min_{y \geq 0, u, \gamma} \nu e'y + \| u \|_1 \\
\text{s.t.} \quad D(AA'Du - e\gamma) + y \geq e
\]

- **Replace** \( AA' \) **by a nonlinear kernel** \( K(A, A') \):

\[
\min_{y \geq 0, u, \gamma} \nu e'y + \| u \|_1 \\
\text{s.t.} \quad D(K(A, A')Du - e\gamma) + y \geq e
\]
2- Norm Nonlinear SVM

\[
\begin{align*}
\min_{y \geq 0, \, u, \, \gamma} & \quad \frac{\nu}{2} \| y \|_2^2 + \frac{1}{2} \| u, \, \gamma \|_2^2 \\
\text{s.t.} & \quad D(K(A, A')Du - e\gamma) + y \geq e
\end{align*}
\]

Equivalently:

\[
\begin{align*}
\min_{u, \, \gamma} & \quad \frac{\nu}{2} \| (e - D(KA, A')Du - e\gamma) \|_+^2 + \frac{1}{2} \| u, \, \gamma \|_2^2
\end{align*}
\]
The Nonlinear Classifier

- The nonlinear classifier:
  \[
  K(x', A') Du = \gamma
  \]
  \[
  K(A, A') : \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times m} \rightarrow \mathbb{R}^{m\times m}
  \]

- \( K \) is a nonlinear kernel, e.g.:
  - Gaussian (Radial Basis) Kernel:
    \[
    K(A, A')_{ij} = \varepsilon^{-\mu\|A_i - A_j\|_2^2}, \quad i, j = 1, \ldots, m
    \]
  - The \( ij \)-entry of \( K(A, A') \) represents “similarity” between the data points \( A_i \) and \( A_j \) (Nearest Neighbor)
  - Can generate highly nonlinear classifiers
Data Reduction in Data Mining

- RSVM: Reduced Support Vector Machines
Difficulties with Nonlinear SVM for Large Problems

- The nonlinear kernel $K(A, A') \in \mathbb{R}^{m \times m}$ is fully dense
  - Long CPU time to compute $m \times m$ elements of nonlinear kernel $K(A, A')$
  - Runs out of memory while storing $m \times m$ elements of $K(A, A')$
- Computational complexity depends on $m$
  - Complexity of nonlinear SSVM $\sim O((m + 1)^3)$
- Separating surface depends on almost entire dataset
  - Need to store the entire dataset after solving the problem
Overcoming Computational & Storage Difficulties
Use a “Thin” Rectangular Kernel

- Choose a small random sample $\overline{A} \in \mathbb{R}^{m \times n}$ of $A$
  - The small random sample $\overline{A}$ is a representative sample of the entire dataset
  - Typically $\overline{A}$ is 1% to 10% of the rows of $A$
- Replace $K(A, A')$ by $K(A, \overline{A}') \in \mathbb{R}^{m \times \overline{m}}$ with corresponding $\overline{D} \subset D$ in nonlinear SSVM
  - Only need to compute and store $m \times \overline{m}$ numbers for the rectangular kernel
- Computational complexity reduces to $O((\overline{m} + 1)^3)$

- The nonlinear separator only depends on $\overline{A}$

  🔥 Using $K(\overline{A}, \overline{A}')$ gives lousy results!
Reduced Support Vector Machine Algorithm

Nonlinear Separating Surface: $K(x', \bar{A}') \bar{D} \bar{u} = \gamma$

(i) Choose a random subset matrix $\bar{A} \in \mathbb{R}^{m \times n}$ of entire data matrix $A \in \mathbb{R}^{m \times n}$

(ii) Solve the following problem by a generalized Newton method with corresponding $\bar{D} \subset D$:

$$\min_{(\bar{u}, \gamma) \in \mathbb{R}^{m+1}} \frac{\nu}{2} \|(e - D(K(A, \bar{A}') \bar{D} \bar{u} - e\gamma))_+\|_2^2 + \frac{1}{2} \|\bar{u}, \gamma\|_2^2$$

(iii) The separating surface is defined by the optimal solution $(\bar{u}, \gamma)$ in step (ii):

$$K(x', \bar{A}') \bar{D} \bar{u} = \gamma$$
A Nonlinear Kernel Application
Checkerboard Training Set: 1000 Points in $\mathbb{R}^2$
Separate 486 Asterisks from 514 Dots
Conventional SVM Result on Checkerboard Using 50 Randomly Selected Points Out of 1000

\[ K(\overline{A}, \overline{A}') \in \mathbb{R}^{50 \times 50} \]
RSVM Result on Checkerboard

Using SAME 50 Random Points Out of 1000

$K(A, \overline{A}') \in \mathbb{R}^{1000 \times 50}$
Knowledge-Based Classification

- Use prior knowledge to improve classifier correctness
Conventional Data-Based SVM

\[ x'w = \gamma + 1 \]

\[ x'w = \gamma \]

\[ x'w = \gamma - 1 \]
Knowledge-Based SVM via Polyhedral Knowledge Sets
Suppose that the knowledge set: \( \{ x \mid Bx \leq b \} \) belongs to the class A+. Hence it must lie in the halfspace:
\[
\{ x \mid x'w \geq \gamma + 1 \}
\]
We therefore have the implication:
\[
Bx \leq b \implies x'w \geq \gamma + 1
\]
This implication is equivalent to a set of constraints that can be imposed on the classification problem.
Knowledge Set Equivalence Theorem

\[ Bx \leq b \implies x'w \geq \gamma + 1, \]

or, for a fixed \((w, \gamma)\):

\[ Bx \leq b, \quad x'w < \gamma + 1, \text{ has no solution } x \]

\[ \iff \{ x \mid Bx \leq b \} \neq \emptyset \]

\[ \exists u : B'u + w = 0, \quad b'u + \gamma + 1 \leq 0, \quad u \geq 0 \]
Knowledge-Based SVM Classification

Adding one set of constraints for each knowledge set to the 1-norm SVM LP, we have:

\[
\begin{align*}
\min_{w, \gamma, y, u^i, v^j} & \quad \nu e'y + \|w\|_1 \\
\text{s.t.} & \quad D(Aw - e\gamma) + y \geq e \\
& \quad y \geq 0 \\
& \quad B^i'u^i + w = 0 \\
& \quad b^i'u^i + \gamma + 1 \leq 0 \\
& \quad u^i \geq 0, \quad i = 1, \ldots, k \\
& \quad C^j'v^j - w = 0 \\
& \quad c^j'v^j - \gamma + 1 \leq 0 \\
& \quad v^j \geq 0, \quad j = 1, \ldots, \ell
\end{align*}
\]
Promoter: Short DNA sequence that precedes a gene sequence.

A promoter consists of 57 consecutive DNA nucleotides belonging to \{A,G,C,T\}.

Important to distinguish between promoters and nonpromoters.

This distinction identifies starting locations of genes in long uncharacterized DNA sequences.
The Promoter Recognition Dataset
Numerical Representation

- Input space mapped from 57-dimensional nominal space to a real valued $57 \times 4 = 228$ dimensional space.

57 nominal values

57 x 4 = 228 binary values
Promoter Recognition Dataset Prior Knowledge Rules as Implication Constraints

- Prior knowledge consist of the following 64 rules:

\[
\left[ \begin{array}{c}
R_1 \\
\text{or} \\
R_2 \\
\text{or} \\
R_3 \\
\text{or} \\
R_4
\end{array} \right] \land \left[ \begin{array}{c}
R_5 \\
\text{or} \\
R_6 \\
\text{or} \\
R_7 \\
\text{or} \\
R_8
\end{array} \right] \land \left[ \begin{array}{c}
R_9 \\
\text{or} \\
R_{10} \\
\text{or} \\
R_{11} \\
\text{or} \\
R_{12}
\end{array} \right] \implies \text{PROMOTER}
\]
Promoter Recognition Dataset
Sample Rules

\[ R_4 : \ (p_{-36} = T) \land (p_{-35} = T) \land (p_{-34} = G) \]
\[ \land (p_{-33} = A) \land (p_{-32} = C), \]

\[ R_8 : \ (p_{-12} = T) \land (p_{-11} = A) \land (p_{-07} = T), \]

\[ R_{10} : \ (p_{-45} = A) \land (p_{-44} = A) \land (p_{-41} = A). \]

A sample rule is:

\[ R_4 \land R_8 \land R_{10} \implies PROMOTER \]
The Promoter Recognition Dataset
Comparative Algorithms

- **KBANN** Knowledge-based artificial neural network [Shavlik et al]
- **BP**: Standard back propagation for neural networks [Rumelhart et al]
- **NN**: Nearest neighbor with k=3 [Cost et al]
- **ID3**: Quinlan’s decision tree builder [Quinlan]
- **SVM1**: Standard 1-norm SVM [Bradley et al]
The Promoter Recognition Dataset
Comparative Test Results
with Linear KSVM

Total leave-one-out error
KSVM & other classification algorithms
106-point Promoter Dataset: 53 Promoters, 53 Nonpromoters

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Errors (out of 106)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KBANN</td>
<td>4</td>
</tr>
<tr>
<td>KSVM</td>
<td>5</td>
</tr>
<tr>
<td>BP</td>
<td>8</td>
</tr>
<tr>
<td>SVM₁</td>
<td>9</td>
</tr>
<tr>
<td>O’Neill</td>
<td>12</td>
</tr>
<tr>
<td>NN</td>
<td>13</td>
</tr>
<tr>
<td>ID3</td>
<td>19</td>
</tr>
</tbody>
</table>
Finite Newton Classifier

- Newton for SVM as an unconstrained optimization problem
Fast Newton Algorithm for SVM Classification

Standard quadratic programming (QP) formulation of SVM:

\[
\min_{w, \gamma, y} \quad \frac{\nu}{2} \| y \|_2^2 + \frac{1}{2} \| w, \gamma \|_2^2 \\
\text{s.t.} \quad D(Aw - e\gamma) + y \geq e \\
y \geq 0,
\]

At solution of QP:

\[ y = (e - D(Aw - e\gamma))_+, \]

where \((\cdot)_+ = \max \{\cdot, 0\}\). Hence QP is equivalent to the nonsmooth SVM:

\[
\min_{w, \gamma} \quad \frac{\nu}{2} \|(e - D(Aw - e\gamma))_+\|_2^2 + \frac{1}{2} \| w, \gamma \|_2^2
\]

Once, but not twice differentiable. However Generalized Hessian exists!
Generalized Newton Algorithm

\[ f(z) = \frac{\nu}{2} \| (Cz - h)_+ \|^2 + \frac{1}{2} \| z \|^2 \]

\[ z^{i+1} = z^i - \partial^2 f(z^i)^{-1} \nabla f(z^i) \]

\[ \nabla f(z) = \nu C'(Cz - h)_+ + z \]

\[ \partial^2 f(z) = \nu C' diag(Cz - h)_+ C + I \]

where \( (Cz - h)_+ = 0 \) if \((Cz - h) \leq 0\), else \((Cz - h)_+ = 1\).

- Newton algorithm terminates in a finite number of steps
  - With an Armijo stepsize (unnecessary computationally)
- Termination at global minimum
- Error rate decreases linearly
- Can generate complex nonlinear classifiers
  - By using nonlinear kernels: \( K(x,y) \)
Nonlinear Spiral Dataset

94 Red Dots & 94 White Dots
SVM Application to Drug Discovery

- Drug discovery based on gene expression
Breast Cancer Drug Discovery Based on Gene Expression
Joint with ExonHit - Paris (Curie Dataset)

- 35 patients treated by a drug cocktail
- 9 partial responders; 26 nonresponders
- 25 gene expressions out of 692 selected by ExonHit
- 1-Norm SVM and greedy combinatorial approach selected 5 genes out of 25
- Most patients had 3 distinct replicate measurements
- Distinguishing aspects of this classification approach:
  - Separate convex hulls of replicates
  - Test on mean of replicates
Separation of Convex Hulls of Replicates

10 Synthetic Nonresponders: 26 Replicates (Points)
5 Synthetic Partial Responders: 14 Replicates (Points)
Linear Classifier in 3-Gene Space
35 Patients with 93 Replicates
26 Nonresponders & 9 Partial Responders

In 5-gene space, leave-one-out correctness was 33 out of 35, or 94.2%
Generalized Eigenvalue Classification

- Multisurface proximal classification via generalized eigenvalues
Multisurface Proximal Classification

- Two distinguishing features:
  - Replace halfspaces containing datasets A and B by planes proximal to A and B
  - Allow nonparallel proximal planes

- First proximal plane: $x^' w^1-\gamma^1=0$
  - As close as possible to dataset A
  - As far as possible from dataset B

- Second proximal plane: $x^' w^2-\gamma^2=0$
  - As close as possible to dataset B
  - As far as possible from dataset A
Classical Exclusive “Or” (XOR) Example

\[ A = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}; \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \]
Multisurface Proximal Classifier
As a Generalized Eigenvalue Problem

\[ \min_{(w, \gamma) \neq 0} \frac{\| Aw - e\gamma \|^2}{\| [w] \|^2} \]

Simplifying and adding regularization terms gives:

\[ \min_{(w, \gamma) \neq 0} \frac{\| Bw - e\gamma \|^2}{\| [w] \|^2} + \delta \| [w] \|^2 \]

Define:

\[
G := [A - e]'[A - e] + \delta I, \quad H := [B - e]'[B - e] + \delta I, \\
z := [w \gamma],
\]
Generalized Eigenvalue Problem

The optimization problem reduces to minimizing the Rayleigh quotient:

$$\min_{z \neq 0} r(z) := \frac{z'Gz}{z'Hz}.$$ 

- The Rayleigh quotient ranges over the interval $[\lambda_1, \lambda_{n+1}]$ for $\|z\|_2 = 1$.
- $\lambda_1$ and $\lambda_{n+1}$ are the minimum and maximum eigenvalues of the generalized eigenvalue problem:

$$Gz = \lambda Hz, \quad z \neq 0.$$ 

The eigenvectors $z^1$ corresponding to the smallest eigenvalue $\lambda_1$ and $z^{n+1}$ corresponding to the largest eigenvalue $\lambda_{n+1}$ determine the two nonparallel proximal planes. $\text{eig}(G, H)$
A Simple Example

Linear Classifier

Generalized Eigenvalue Classifier

80% Correctness

100% Correctness

Also applied successfully to real world test problems
Conclusion

- Variety of optimization-based approaches to data mining
  - Feature selection in both clustering & classification
  - Enhanced knowledge-based classification
  - Finite Newton method for nonlinear classification
  - Drug discovery based on gene macroarrays
  - Proximal classification via generalized eigenvalues

- Optimization is a powerful and effective tool for data mining, especially for implementing Occam’s Razor
  - “Simplest is best”