Short Course
Robust Optimization and Machine Learning

Lecture 6:
Robust Optimization in Machine Learning

Laurent El Ghaoui

EECS and IEOR Departments
UC Berkeley

Spring seminar TRANSP-OR, Zinal, Jan. 16-19, 2012

Outline

Outline

Supervised learning problems

Many supervised learning problems (*e.g.*, classification, regression) can be written as

$$\min_{w} \mathcal{L}(X^T w)$$

where $\mathcal{L}$ is convex, and X contains the data.

Penalty approach

Often, optimal value and solutions of optimization problems are sensitive to data.

A common approach to deal with sensitivity is via penalization, *e.g.*:

$$\min_x \; \mathcal{L}(X^T w) + \|Wx\|_2^2 \qquad (W = \text{weighting matrix}).$$

- How do we choose the penalty?
- Can we choose it in a way that reflects knowledge about problem structure, or how uncertainty affects data?
- Does it lead to better solutions from machine learning viewpoint?

Support Vector Machine

Support Vector Machine (SVM) classification problem:

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+$$

- $Z := [z_1, \ldots, z_m] \in \mathbf{R}^{n \times m}$ contains the *data points* .
- $y \in \{-1, 1\}^m$ contain the *labels* .
- $x := (w, b)$ contains the *classifier parameters* , allowing to classify a new point z via the rule

$$y = \mathbf{sgn}(z^T w + b).$$

Robustness to data uncertainty

Assume the data matrix is only partially known, and address the robust optimization problem:

$$\min_{w,b} \max_{U \in \mathcal{U}} \sum_{i=1}^{m} (1 - y_i((z_i + u_i)^T w + b))_+,$$

where $U = [u_1, \ldots, u_m]$ and $\mathcal{U} \subseteq \mathbf{R}^{n \times m}$ is a set that describes additive uncertainty in the data matrix.

Measurement-wise, spherical uncertainty

Assume

$$\mathcal{U} = \{U = [u_1, \ldots, u_m] \in \mathbf{R}^{n \times m} \ : \ \|u_i\|_2 \le \rho\},$$

where $\rho > 0$ is given.

Robust SVM reduces to

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b) + \rho \|w\|_2)_+.$$

Link with classical SVM

Classical SVM contains l_2-norm regularization term:

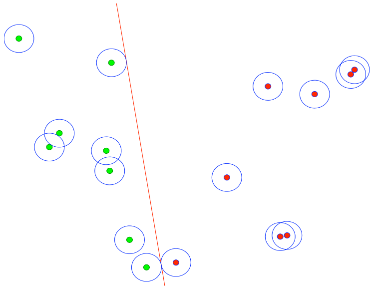
$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+ + \lambda \|w\|_2^2.$$

where $\lambda > 0$ is a penalty parameter.

With spherical uncertainty, robust SVM is similar to classical SVM.

When data is separable, the two models are equivalent ...

Separable data



Maximally robust classifier for separable data, with spherical uncertainties around each data point. In this case, the robust counterpart reduces to the classical maximum-margin classifier problem.

Interval uncertainty

Assume

$$\mathcal{U} = \{U \in \mathbf{R}^{n \times m} \; : \; \forall (i,j), \;\; |U_{ij}| \le \rho\},$$

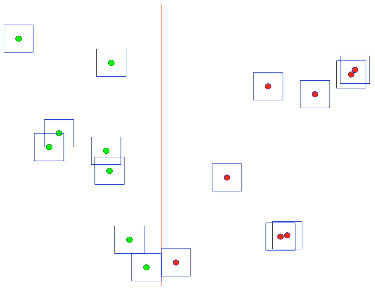
where $\rho > 0$ is given.

Robust SVM reduces to

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b) + \rho \|w\|_1)_+.$$

The l_1-norm term encourages sparsity, and may not regularize the solution.

Separable data



Maximally robust classifier for separable data, with box uncertainties around each data point. This uncertainty model encourages sparsity of the solution.

Other uncertainty models

We may generalize the approach to other uncertainty models, retaining tractability:

- "Measurement-wise" uncertainty models: perturbations affect each data point independent of each other.
- Other models couple the way uncertainties affect each measurement; for example we may control the number of errors across all the measurements.
- Norm-bound models allow for uncertainty of data matrix that is bounded in matrix norm.
- A whole theory is presented in [1].

Thresholding and robustness

Consider standard l_1-penalized SVM:

$$\phi_\lambda(X) := \min_{w,b} \sum_{i=1}^{m} (1 - y_i(w^T x_i + b))_+ + \lambda \|w\|_1$$

Constrained counterpart:

$$\psi_c(X) := \min_{w,b} \frac{1}{m} \sum_{i=1}^{m} (1 - y_i(x_i^T w + b))_+ \ : \ \|w\|_1 \le c$$

- *Basic goal:* solve these problems in the large-scale case.
- *Approach:* use robustness to sparsify the data matrix in a controlled way.

Thresholding data

We threshold the data using an absolute level t:

$$(x_i(t))_j := \begin{cases} 0 & \text{if } |x_{i,j}| \leq t \\ 1 & \text{otherwise} \end{cases}$$

This will make the data sparser, resulting in memory and time savings.

Handling thresholding errors

Handle thresholding errors via robust counterpart:

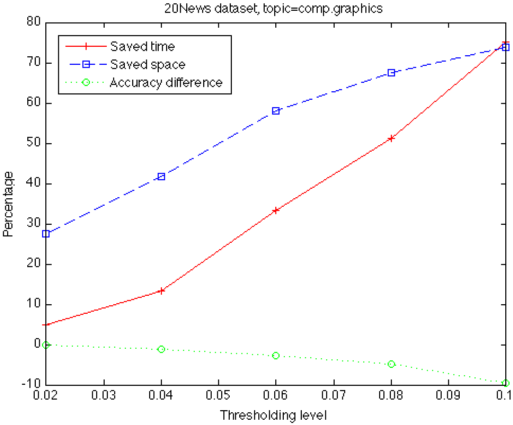
$$(w(t), b(t)) := \arg\min_{w,b} \max_{\|Z-X\|_\infty \le t} \sum_{i=1}^{m} (1 - y_i(w^T z_i + b))_+ + \lambda \|w\|_1 .$$

Above problem is tractable.

The solution $w(t)$ at threshold level t satisfies

$$0 \le \frac{1}{m} \sum_{i=1}^{m} (1 - y_i(x_i^T w(t) + b(t)))_+ + \lambda \|w(t)\|_1 - \phi_\lambda(X) \le \frac{2t}{\lambda}.$$

Results
20 news groups data set



Dataset size: $20,000 \times 60,000$. Thresholding of data matrix of TF-IDF scores.

Results

UCI NYTimes Dataset

1	stock	11	bond
2	nasdaq	12	forecast
3	portfolio	13	thomson financial
4	brokerage	14	index
5	exchanges	15	royal bank
6	shareholder	16	fund
7	fund	17	marketing
8	investor	18	companies
9	alan greenspan	19	bank
10	fed	20	merrill

Top 20 keywords for topic '**stock**'. Dataset size: $100,000 \times 102,660$, $\approx$ 30,000,000 non-zeros. Thresholded dataset (by TF-IDF scores) with level 0.05 $\approx$ 850,000 non-zeroes (2.8 %). Total run time: 4317s.

Robust SVM with Boolean data

- *Data:* boolean $Z \in \{0,1\}^{n\times m}$ (eg, co-occurence matrix)
- *Nominal problem:* SVM

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+,$$

- *Uncertainty model:* assume each data value can be flipped, total budget of flips is constrained:

$$\mathcal{U} = \left\{ U = [u_1, \ldots, u_m] \in \mathbf{R}^{l\times m} \;:\; u_i \in \{-1,0,1\}^l, \;\; \|u\|_1 \leq k \right\}.$$

Robust counterpart

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b) + \phi(w))_+,$$

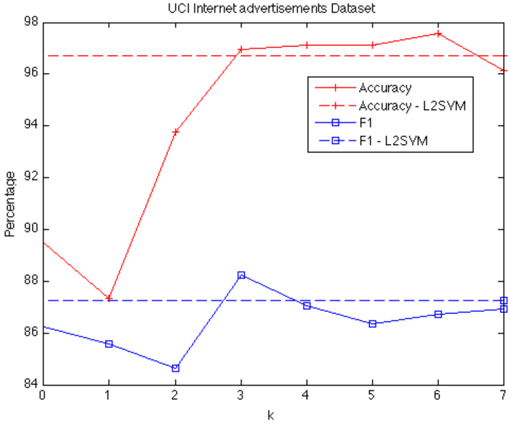
where

$$\phi(w) := \min_{s} \; k\|w - s\|_\infty + \|s\|_1$$

- Penalty is a combination of l_1, l_∞ norms.
- Problem is tractable (doubles number of variables over nominal).
- Still needs regularization.

Results

UCI Internet advertisement data set



Dataset size: 3279×1555. $k = 0$ corresponds to nominal SVM problem. Best performance at $k = 3$.

Refined model

We can impose $u_i \in \{0, 1 - 2x_i\}$. This leads to a new penalty:

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(x_i^T w + b) + \phi_i(w))_+,$$

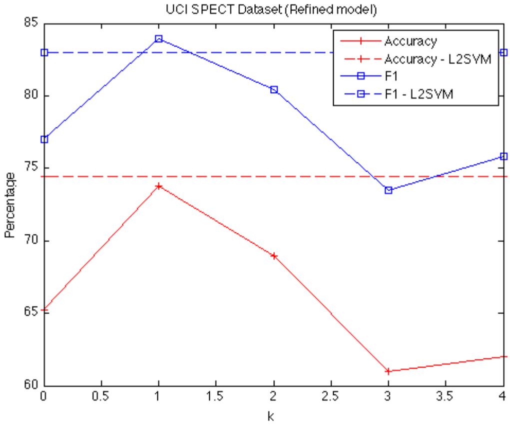
with

$$\phi_i(w) := \min_{\mu \geq 0} \; k\mu + \sum_{j=1}^{n} (y_i w_j (2x_{ij} - 1) - \mu)_+$$

Problem can still be solved via LP.

Results

UCI Heart data set



Dataset size: 267×22. $k = 0$ corresponds to nominal SVM problem. Best performance at $k = 1$.

Outline

Nominal problem

$$\min_{\theta \in \Theta} \mathcal{L}(Z^T\theta),$$

where

- $Z := [z_1, \ldots, z_m] \in \mathbf{R}^{n \times m}$ is the data matrix
- $\mathcal{L} : \mathbf{R}^m \to \mathbf{R}$ is a convex loss function
- Θ imposes "structure" (eg, sign) constraints on parameter vector θ

Loss function: assumptions

We assume that

$$\mathcal{L}(r) = \pi(\mathbf{abs}(P(r))),$$

where $\mathbf{abs}(\cdot)$ acts componentwise, $\pi : \mathbf{R}^m_+ \to \mathbf{R}$ is a convex, monotone function on the non-negative orthant, and

$$P(r) = \begin{cases} r & (\text{"symmetric case"}) \\ r_+ & (\text{"asymmetric case"}) \end{cases}$$

with r_+ the vector with components $\max(r_i, 0)$, $i = 1, \ldots, m$.

Loss function: examples

- l_p-norm regression
- hinge loss
- Huber, Berhu loss

Robust counterpart

$$\min_{\theta \in \Theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta).$$

where $\mathcal{Z} \subseteq \mathbf{R}^{n \times m}$ is a set of the form

$$\mathcal{Z} = \{ Z + \Delta \ : \ \Delta \in \rho\mathcal{D}, \},$$

with $\rho \geq 0$ a measure of the size of the uncertainty, and $\mathcal{D} \subseteq \mathbf{R}^{l \times m}$ is given.

Generic analysis

For a given vector θ, we have

$$\max_{\mathbf{Z}\in\mathcal{Z}} \mathcal{L}(\mathbf{Z}^T\theta) = \max_u u^T Z^T \theta - \mathcal{L}^*(u) + \rho\phi_{\mathcal{D}}(uv^T),$$

where $\mathcal{L}^*$ is the conjugate of $\mathcal{L}$, and

$$\phi_{\mathcal{D}}(X) := \max_{\Delta\in\mathcal{D}} \langle X, \Delta\rangle$$

is the support function of $\mathcal{D}$.

Assumptions on uncertainty set $\mathcal{D}$

Separability condition: there exist two semi-norms ϕ, ψ such that

$$\phi_{\mathcal{D}}(uv^T) := \max_{\Delta \in \mathcal{D}} u^T \Delta v = \phi(u)\psi(v).$$

- Does not completely characterize (the support function of) the set $\mathcal{D}$
- Given ϕ, ψ, we can construct a set $\mathcal{D}_{\text{out}}$ that obeys condition
- The robust counterpart only depends on ϕ, ψ.

WLOG, we can replace $\mathcal{D}$ by its convex hull.

Examples

- Largest singular value model: $\mathcal{D} = \{\Delta \;:\; \|\Delta\| \leq \rho\}$, with ϕ, ψ Euclidean norms.
- Any norm-bound model involving an induced norm (ϕ, ψ are then the norms dual to the norms involved).
- Measurement-wise uncertainty models, where each column of the perturbation matrix is bounded in norm, independently of the others, correspond to the case with $\psi(v) = \|v\|_1$.

Other examples

Bounded-error model: there are (at most K) errors affecting data

$$\mathcal{D} = \left\{ \Delta = [\lambda_1 \delta_1, \ldots, \lambda_m \delta_m] \in \mathbf{R}^{l \times m} \ : \ \begin{array}{l} \|\delta_i\| \leq 1, \quad i = 1, \ldots, m, \\ \displaystyle\sum_{i=1}^{m} \lambda_i \leq K, \quad \lambda \in \{0,1\}^m \end{array} \right\}.$$

for which $\phi(\cdot) = \| \cdot \|_*$, $\psi(v) =$ sum of the K largest magnitudes of the components of v.

Examples (follow'd)

- The set

$$\mathcal{D} = \left\{ \Delta = [\lambda_1\delta_1, \ldots, \lambda_m\delta_m] \in \mathbf{R}^{l\times m} \;:\; \delta_i \in \{-1,0,1\}^l, \;\; \|\delta\|_1 \le k \right\}$$

models measurement-wise uncertainty affecting Boolean data (we can impose $\delta_i \in \{x_i - 1, x_i\}$ to be more realistic)
In this case, we have $\psi(\cdot) = \|\cdot\|_1$ and

$$\phi(u) = \|u\|_{1,k} := \min_w \; k\|u - w\|_\infty + \|w\|_1 .$$

Main result

For a given vector θ, we have

$$\min_{\theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta) = \min_{\theta, \kappa} \mathcal{L}_{\rm wc}(Z^T \theta, \kappa) \ : \ \kappa \geq \phi(U^T \theta)$$

where

$$\mathcal{L}(r, \kappa) := \max_{v} \ v^T r - \mathcal{L}^*(v) + \kappa \psi(v)$$

is the worst-case loss function of the robust problem.

Worst-case loss function

The tractability of the robust counterpart is directly linked to our ability to compute optimal solutions v^* for

$$\mathcal{L}(r, \kappa) = \max_v \; v^T r - \mathcal{L}^*(v) + \kappa\psi(v)$$

Dual representation (assume $\psi(\cdot) = \|\cdot\|$ is a norm):

$$\mathcal{L}(r, \kappa) = \max_\xi \; \mathcal{L}(r + \kappa\xi) \;:\; \|\xi\|_* \leq 1$$

When ψ is the Euclidean norm, robust regularization of $\mathcal{L}$ (Lewis, 2001).

Special cases

- When $\psi(\cdot) = \| \cdot \|_p$, $p = 1, \infty$, problem reduces to simple, tractable convex problem (assuming nominal problem is).
- For $p = 2$, problem can be reduced to such a simple form, for the hinge, l_q-norm and Huber loss functions.

Lasso

In particular, the least-squares problem with lasso penalty

$$\min_{\theta} \|X^T\theta - y\|_2 + \rho\|\theta\|_1$$

is the robust counterpart to a least-squares problem with uncertainty on X, with additive perturbation bounded in the norm

$$\|\Delta\|_{1,2} := \max_{1 \le i \le l} \sqrt{\sum_{j=1}^{n} \Delta_{ij}^2}.$$

Globalized robust counterpart

The robust counterpart is based on the worst-case value of the loss function assuming a bound on the data uncertainty ($Z \in \mathcal{Z}$):

$$\min_{\theta \in \Theta} \max_{\mathbf{Z} \in \mathcal{Z}} \mathcal{L}(\mathbf{Z}^T \theta).$$

The approach does not control the degradation of the loss outside the set $\mathcal{Z}$.

Globalized robust counterpart: formulation

In globalized robust counterpart, we fix a “rate” of degradation of the loss, which controls the amount of degradation of the loss as the data matrix Z goes “away from” the set $\mathcal{Z}$.
We seek to minimize τ, such that

$$\forall \, \Delta \;:\; \mathcal{L}((Z+\Delta)^T\theta) \leq \tau + \alpha\|\Delta\|,$$

where $\alpha > 0$ controls the rate of degradation, and $\|\cdot\|$ is a matrix norm.

Globalized robust counterpart

Examples

- For the SVM case, the globalized robust counterpart can be expressed as:

$$\min_{w,b} \sum_{i=1}^{m} (1 - y_i(z_i^T w + b))_+ \; : \; \sqrt{m}\|\theta\|_2 \leq \alpha,$$

  which is a classical form of SVM.

- For l_p-norm regression with m data points, the globalized robust counterpart takes the form

$$\min_{\theta} \|X^T\theta - y\|_p \; : \; \kappa(m,p)\|\theta\|_2 \leq \alpha$$

  where $\kappa(m,1) = \sqrt{m}$, $\kappa(m,2) = \kappa(m,\infty) = 1$.

Chance constraints

Theory can address problems with “chance constraints”

$$\min_{\theta} \max_{p \in \mathcal{P}} \mathbf{E}_p \mathcal{L}(Z(\delta)^T \theta)$$

where δ follows distribution p, and $\mathcal{P}$ is a class of distributions

- Results are more limited, focused on upper bounds.
- Convex relaxations are available, but more expensive.
- Approach uses Bernstein approximations (Nemirovski & Ben-tal, 2006).

Robust regression with chance constraints: an example

$$\phi_p := \min_{\theta} \max_{x \sim (\hat{x}, X)} \mathbf{E}_x \|A(x)\theta - b(x)\|_p$$

- Regression variable is $\theta \in \mathbf{R}^n$
- $x \in \mathbf{R}^q$ is an uncertainty vector that enters affinely in the problem matrices: $[A(x), b(x)] = [A_0, b_0] + \sum_i x_i [A_i, b_i]$.
- The distribution of uncertainty vector x is unknown, except for its mean $\hat{x}$ and covariance X.
- Objective is worst-case (over distributions) expected value of l_p-norm residual ($p = 1, 2$).

Main result

(Assume $\hat{x} = 0$, $X = I$ WLOG)
For $p = 2$, the problem reduces to least-squares:

$$\phi_2^2 = \min_\theta \sum_{i=0}^{q} \|A_i\theta - b_i\|_2^2$$

For $p = 1$, we have $(2/\pi)\psi_1 \le \phi_1 \le \psi_1$, with

$$\psi_1 = \min_\theta \sum_{i=0}^{q} \|A_i\theta - b_i\|_2$$

Example: robust median

As a special case, consider the median problem:

$$\min_{\theta} \sum_{i=1}^{q} |\theta - x_i|$$

Now assume that vector x is random, with mean $\hat{x}$ and covariance X, and consider the robust version:

$$\phi_1 := \min_{\theta} \max_{x \sim (\hat{x}, X)} \mathbf{E}_x \sum_{i=1}^{q} |\theta - x_i|$$

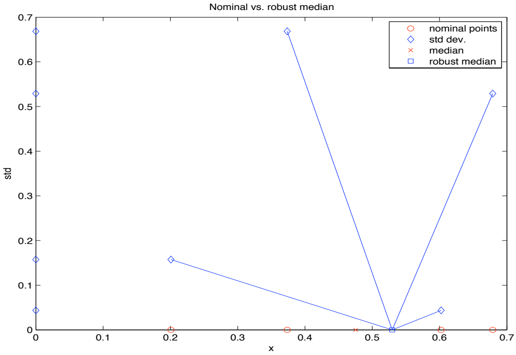
Approximate solution

We have $(2/\pi)\psi_1 \leq \phi_1 \leq \psi_1$, with

$$\psi_1 := \sum_{i=1}^{n} \sqrt{(\theta - \hat{x}_i)^2 + X_{ii}}$$

Amounts to find the minimum distance sum (a very simple SOCP).

Geometry of robust median problem



Outline

References

A. Bental, L. El Ghaoui, and A. Nemirovski.
Robust Optimization.
Princeton Series in Applied Mathematics. Princeton University Press, October 2009.

C. Caramanis, S. Mannor, and H. Xu.
Robust optimization in machine learning.
In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*, chapter 14. MIT Press, 2011.