Regularization frontier in machine learning

G. Gasso

LITIS EA 4108
INSA de Rouen, France

ECML PKDD 2008

September 19th, 2008

Antwerp, Belgium
Roadmap

1. Introduction
   - Framework
   - Model selection

2. Regularization path and pareto frontier

3. Efficient regularization path running
   - Piecewise linear regularization path

4. Two examples of regularization path
   - Lasso path
   - SVM path

5. Regularization path and sparsity

6. Extensions and efficiency evaluation
Learning problem

Framework

- Set of data \( \mathcal{D} = \{x_i, y_i\}_{i=1,\ldots,n} \) with \((x, y) \in \mathcal{X} \times \mathcal{Y} \)
  \((X, Y) \sim P_{X,Y} \) with \( P_{X,Y} \) the unknown joint distribution

- Supervised learning
  - Binary classification \( \mathcal{Y} = \{-1, +1\} \)
  - Regression \( \mathcal{Y} = \mathbb{R} \)

- Task: find a predictive model \( f \)

\[ f : \mathcal{X} \rightarrow \mathcal{Y} \]
\[ x \mapsto \hat{y} = f(x) \]

- \( f \) belongs to an hypothesis space \( \mathcal{H} \)
Learning problem

**Framework (ct’d)**

- A non-negative loss function $\ell$
- Expected risk minimization $f^* = \text{argmin}_{f \in \mathcal{H}} E_{X,Y} (\ell(f(X), Y))$
- **Empirical loss** minimization

$$
\hat{f} = \text{argmin}_{f \in \mathcal{H}} L(f)
$$

with $L(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i))$

- To avoid overtraining, some constraints (smoothness, sparsity, robustness, …) are enforced on $f$ by using a Regularizer or penalty term $P(f)$
- **Regularized optimization problem**

$$
\hat{f} = \text{argmin}_{f \in \mathcal{H}} L(f) + \lambda P(f)
$$

$\lambda \in \mathbb{R}^+$ is a trade-off or regularization parameter
Model selection

Tuning of $\lambda$

Identify the appropriate value $\lambda^*$ associated to the best solution $\hat{f}^*$

Illustration: non-linear ridge regression

Overfitting

Small $\lambda$

Some irregular behavior

Convenient solution

Mid value of $\lambda$

Underfitting

High $\lambda$

Too smooth
Determination of $\lambda$

- Compute the decision function $\hat{f}_\lambda$ for different values of $\lambda$
- Select the best solution according to some generalization performance
Model selection

Determination of $\lambda$

- Compute the decision function $\hat{f}_\lambda$ for different values of $\lambda$
- Select the best solution according to some generalization performance

Two approaches

1. Grid search over predefined set $\{\lambda_1, \ldots, \lambda_K\}$
2. Compute the regularization path

Gasso (LITIS, EA 4108) Regularization path and machine learning Antwerp, 19/09/2008
Model selection

Determination of $\lambda$

- Compute the decision function $\hat{f}_\lambda$ for different values of $\lambda$
- Select the best solution according to some generalization performance

Two approaches

1. Grid search over predefined set $\{\lambda_1, \ldots, \lambda_K\}$

- Values specified by the user
- Retained solution $\hat{f}^*(x)$ depends highly on the grid resolution
Model selection

**Determination of \( \lambda \)**

- Compute the decision function \( \hat{f}_\lambda \) for different values of \( \lambda \)
- Select the best solution according to some generalization performance

**Two approaches**

1. Grid search over predefined set \( \{\lambda_1, \ldots, \lambda_K\} \)
2. Compute the regularization path

- No values specified by the user
- Find automatically all solutions \( \hat{f}_\lambda(x) \)

**Regularization path**

The set of all solutions \( \hat{f}_\lambda(x) \) i.e. \( \mathcal{R} = \{ \hat{f}_\lambda(x) \mid \lambda \in [0, \infty) \} \)
Introduction

2. Regularization path and pareto frontier

3. Efficient regularization path running

4. Two examples of regularization path

5. Regularization path and sparsity

6. Extensions and efficiency evaluation
Case study

Linear ridge regression

- Model: \( f(x) = x^T \beta \) with \( \beta \in \mathbb{R}^d \)
- Problem:
  \[
  \min_{\beta \in \mathbb{R}^d} \|y - X\beta\|^2 + \lambda \|\beta\|^2
  \]
- Solution:
  \[
  \hat{\beta}(\lambda) = (X^TX + \lambda I)^{-1} X^T y
  \]
  \( I \): identity matrix

Regularization path

- \( R = \{ \hat{\beta}(\lambda) \mid \lambda \in [0, \infty) \} \)
- \( \lambda = 0, \hat{\beta}_{LS} = (X^TX)^{-1} X^T y \) (least squares solution)
- \( \lambda \to \infty, \hat{\beta} = 0 \)
The Loss $L$ as a function of the regularizer $P$

\[
\begin{align*}
L(\beta) &= \sum_{i=1}^{n} (x_i \beta - y_i)^2 \\
P(\beta) &= \beta^2
\end{align*}
\]

It holds that

\[
\begin{align*}
L(P) &= aP \pm b\sqrt{P} + c \\
a, b \text{ and } c \in \mathbb{R}
\end{align*}
\]
Notion of Dominance and Pareto frontier

\[
\begin{align*}
L(\beta) &= ||y - X\beta||^2 \\
P(\beta) &= ||\beta||^2
\end{align*}
\]

Dominance

A vector $\beta_1$ dominates another vector $\beta_2$ if $L(\beta_1) \leq L(\beta_2)$ and $P(\beta_1) \leq P(\beta_2)$

Pareto frontier

Pareto frontier is the set of all non dominated solutions

Fig.: dominated point (red), non dominated point (purple) and Pareto frontier (blue).

Pareto frontier $\Leftrightarrow$ Reg. path
3 equivalent formulations

It works for CONVEX criteria!

Formulation 1: Lagrangian

(Linear combination of $L$ and $P$)

$$
\min_\beta \|y - X\beta\|^2 + \lambda \|\beta\|^2
$$

![Graph showing the relationship between loss and penalty](Image)
3 equivalent formulations

It works for CONVEX criteria!

Formulation 1

$$\min_{\beta} \| y - X\beta \|^2 + \lambda \|\beta\|^2$$

Formulation 2

$$\begin{cases} 
\min_{\beta} \| y - X\beta \|^2 \\
\text{s.t.} \quad \|\beta\|^2 \leq C
\end{cases}$$
3 equivalent formulations

It works for CONVEX criteria!

Formulation 1

$$\min_\beta \|y - X\beta\|^2 + \lambda \|\beta\|^2$$

Formulation 2

$$\begin{cases} 
\min_\beta \|y - X\beta\|^2 \\
\text{s.t. } \|\beta\|^2 \leq C 
\end{cases}$$

Formulation 3

$$\begin{cases} 
\min_\beta \|\beta\|^2 \\
\text{s.t. } \|y - X\beta\|^2 \leq C' 
\end{cases}$$
The importance of convexity

Non convex case

The 3 formulations are not equivalent
so far...

- learning is a multi objective problem
- the regularization path is the Pareto frontier
- beware the non convex case
- it works for more than 2 criteria

What for?

To tune (efficiently) the regularization parameter $\lambda$
Roadmap

1. Introduction

2. Regularization path and pareto frontier

3. Efficient regularization path running

4. Two examples of regularization path

5. Regularization path and sparsity

6. Extensions and efficiency evaluation
Tuning the regularization parameter $\lambda$

Ridge regression example

$$\min_{\beta \in \mathbb{R}^d} \| y - X\beta \|^2 + \lambda \| \beta \|^2$$
Tuning the regularization parameter $\lambda$

Ridge regression example: $\min_{\beta \in \mathbb{R}^d} ||y - X\beta||^2 + \lambda ||\beta||^2$

**Grid Search**

- for each $\lambda_1 < \lambda_2 < ... < \lambda_t < ... < \lambda_K$
- compute $\beta_t = (X^\top X + \lambda_t I)^{-1} X^\top y$, $t = 1, \ldots, K$

$O(Kd^3)$
Tuning the regularization parameter $\lambda$

Ridge regression example

$$\min_{\beta \in \mathbb{R}^d} \| y - X\beta \|^2 + \lambda \| \beta \|^2$$

- **Grid Search**
  
  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_t < \ldots < \lambda_K$
  
  compute $\beta_t = (X^\top X + \lambda_t I)^{-1} X^\top y$, $t = 1, \ldots, K$

  $O(Kd^3)$

- **Warm start**
  
  $\beta_t = \Phi(\beta_{t-1})$ (using $\ell$ conjugate gradient iterations)

  $O(K\ell d^2)$
Tuning the regularization parameter $\lambda$

Ridge regression example \[ \min_{\beta \in \mathbb{R}^d} \| y - X\beta \|^2 + \lambda \| \beta \|^2 \]

- **Grid Search**
  
  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_t < \ldots < \lambda_K$
  
  compute $\beta_t = (X^\top X + \lambda_t I)^{-1} X^\top y$, \( t = 1, \ldots, K \)
  
  $O(Kd^3)$

- **Warm start**
  
  $\beta_t = \Phi(\beta_{t-1})$  \((using \ell \ conjugate \ gradient \ iterations)\)
  
  $O(K\ell d^2)$

- **Warm start + prediction step**
  
  $\beta_t^{(p)} = \beta_{t-1} + \rho \nabla \beta (L(\beta_{t-1}) + \lambda_t P(\beta_{t-1}))$  \( \text{prediction step} \)

  $\beta_t = \Phi(\beta_t^{(p)})$  \( \text{correction step using conjugate gradient} \)

  $O(K\ell d^2)$

Use only the prediction step!

To do so the regularization path has to be piecewise linear

$O(Kd^2)$
Tuning the regularization parameter $\lambda$

Ridge regression example: $\min_{\beta \in \mathbb{R}^d} \|y - X\beta\|^2 + \lambda \|\beta\|^2$

- **Grid Search**
  
  for each $\lambda_1 < \lambda_2 < \ldots < \lambda_t < \ldots < \lambda_K$
  
  compute $\beta_t = (X^\top X + \lambda_t I)^{-1} X^\top y$, $t = 1, \ldots, K$
  
  $O(Kd^3)$

- **Warm start**
  
  $\beta_t = \Phi(\beta_{t-1})$ (using $\ell$ conjugate gradient iterations)
  
  $O(K\ell d^2)$

- **Warm start + prediction step**
  
  $\beta_t^{(p)} = \beta_{t-1} + \rho \nabla_\beta(L(\beta_{t-1}) + \lambda_t P(\beta_{t-1}))$ (prediction step)
  
  $\beta_t = \Phi(\beta_t^{(p)})$ (correction step using conjugate gradient)
  
  $O(K\ell'd^2)$

- **Use only the prediction step!**
  
  $\beta_t = \beta_{t-1} + \lambda_t \Psi(\beta_{t-1})$ (prediction step)
  
  to do so the regularization path has to be piecewise linear
  
  $O(Kd^2)$
Piecewise linearity conditions

$$\min_{\beta \in \mathbb{R}^d} L(\beta) + \lambda P(\beta)$$

**How to choose $L$ and $P$ to get linear reg. path?**

Solution path is linear $\iff$ one cost is piecewise quadratic and the other one piecewise linear

convex case [Rosset & Zhu, 07]

Piecewise linearity can be interpreted as

$$\lim_{\varepsilon \to 0} \frac{\beta(\lambda + \varepsilon) - \beta(\lambda)}{\varepsilon} = \text{constant}$$
Piecewise linearity conditions: proof

1. Optimality conditions

For $\lambda \rightarrow$ 
$$\nabla L(\beta(\lambda)) + \lambda \nabla P(\beta(\lambda)) = 0$$

For $\lambda + \varepsilon \rightarrow$ 
$$\nabla L(\beta(\lambda + \varepsilon)) + (\lambda + \varepsilon) \nabla P(\beta(\lambda + \varepsilon)) = 0$$

2. Use first order Taylor expansion around $\beta(\lambda)$

$$\nabla L(\beta(\lambda)) + \nabla^2 L(\beta(\lambda)) [\beta(\lambda + \varepsilon) - \beta(\lambda)]$$
$$+ \lambda \nabla P(\beta(\lambda)) + \lambda \nabla^2 P(\beta(\lambda)) [\beta(\lambda + \varepsilon) - \beta(\lambda)] + \varepsilon \nabla P(\beta(\lambda)) + O(\varepsilon^2) = 0$$

3. Variation of $\beta$ according to $\lambda$

$$\lim_{\varepsilon \to 0} \frac{\beta(\lambda + \varepsilon) - \beta(\lambda)}{\varepsilon} = -\left[\nabla^2 L(\beta(\lambda)) + \lambda \nabla^2 P(\beta(\lambda))\right]^{-1} \nabla P(\beta(\lambda))$$

$$\nabla^2 L(\beta(\lambda)) = \text{constant} \quad \text{and} \quad \nabla^2 P(\beta(\lambda)) = 0$$
Examples of Loss and Penalty

- 0/1 loss
- hinge
- hinge$^2$
- logistic

Loss $L$ vs $yf(x)$

- $\varepsilon$ insistent
- L1
- L2
- Huber

Loss $L$ vs $f(x) - y$
### Tab.: example of piecewise linear regularization path algorithms.

<table>
<thead>
<tr>
<th>$L$</th>
<th>$P$</th>
<th><strong>regression</strong></th>
<th><strong>classification</strong></th>
<th><strong>clustering</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_2$</td>
<td>$L_1$</td>
<td>Lasso/LARS</td>
<td>L1 L2 SVM</td>
<td>OneClass SVM</td>
</tr>
<tr>
<td>$L_1$</td>
<td>$L_2$</td>
<td>SVR</td>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>$L_1$</td>
<td>$L_1$</td>
<td>L1 least absolute deviation</td>
<td>L1 SVM</td>
<td></td>
</tr>
</tbody>
</table>

\[
P : \quad L_p = \sum_{j=1}^{d} |\beta_j|^p
\]

\[
L : \quad L_p : |f(x) - y|^p \quad \text{hinge} \ (yf(x) - 1)_+^p
\]

\[
\varepsilon\text{-insensitive} \quad \begin{cases} 
0 & \text{if } |f(x) - y| < \varepsilon \\
|f(x) - y| - \varepsilon & \text{otherwise}
\end{cases}
\]

\[
\text{Huber's loss :} \quad \begin{cases} 
|f(x) - y|^2 & \text{if } |f(x) - y| < t \\
2t|f(x) - y| - t^2 & \text{otherwise}
\end{cases}
\]
**Piecewise regularization path**

- the problem

\[
\min_{\beta \in \mathbb{R}^d} L(\beta) + \lambda P(\beta) \iff \{ \beta(\lambda) \mid \lambda \in [0, \infty] \}
\]

- \( L \) and \( P \) are convex
- efficient computation

\[ \Rightarrow \text{piecewise linearity} \]

\[ \beta_{t+1} = \beta_t + (\lambda_{t+1} - \lambda_t)w \]

- piecewise linearity

\[ \Rightarrow \text{either } L \text{ or } P \text{ is } L_1 \text{ type} \]
An old result revisited

- **Portfolio management (Markovitz, 1952)**
  
  Gain vs. risk
  
  \[
  \min_\beta \; \frac{1}{2} \beta^\top Q \beta \\
  \text{with} \quad e^\top \beta = C
  \]

  *efficiency frontier*: piecewise linearity (*Critical path Algo.*)

- **Sensitivity analysis (Heller, 1954)**
  

  \[
  \min_\beta \; \frac{1}{2} \beta^\top Q \beta + (c + \lambda \Delta c)^\top \beta \\
  \text{avec} \quad A\beta = b + \mu \Delta b
  \]

- **Parametric programming (Gal 1968)**
  
  Parametric Linear Programming is piecewise linear
  
  PQP piecewise quadratic
  
  Multiparametric programming...
Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation
Lasso (Basis pursuit) problem

\[ \min_{\beta \in \mathbb{R}^d} \| y - X\beta \|_2^2 + \lambda \sum_{i=1}^d |\beta_j| \iff \left\{ \begin{array}{l}
\min_{\beta \in \mathbb{R}^d} \| y - X\beta \|_2^2 \\
\text{with} \sum_{i=1}^d |\beta_j| \leq C
\end{array} \right. \]

- Variables \( x_j = X(:,j) \) for \( j = 1, \cdots, d \)
- Assume all \( x_j \) and \( y \) are centered and normalized

Illustration on two dimensional example

Small \( C \) leads to \( \beta_1 = \beta_2 = 0 \)

High \( C \) produces the least squares solution
Lasso regularization path

\[ \min_{\beta} \frac{1}{2} \| y - X\beta \|^2 + \lambda \sum_{i=1}^{d} |\beta_i| \]

Optimality condition for variable \( x_j \)

\[ -x_j^T (X\beta - y) + \lambda \partial(|\beta_j|) = 0 \]

Subdifferential

\[ \partial(|\beta_j|) = \begin{cases} \text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\ \alpha_j \in [-1, 1] & \text{if } \beta_j = 0 \end{cases} \]
Lasso regularization path

\[
\min_\beta \frac{1}{2} \|y - X\beta\|^2 + \lambda \sum_{i=1}^{d} |\beta_i|
\]

Optimality condition for variable \(x_j\)

\[
-x_j^T (X\beta - y) + \lambda \partial(|\beta_j|) = 0
\]

Subdifferential

\[
\partial(|\beta_j|) = \begin{cases} 
\text{sign}(\beta_j) & \text{if } \beta_j \neq 0 \\
\alpha_j \in [-1, 1] & \text{if } \beta_j = 0
\end{cases}
\]

Active set: \(l_\beta = \{ \beta_j \mid \beta_j \neq 0 \} \)

\[
|x_j^T (X\beta - y)| = \lambda, \quad \beta_j \in l_\beta
\]

Inactive set: \(l_0 = \{ \beta_j \mid \beta_j = 0 \} \)

\[
|x_j^T (X\beta - y)| \leq \lambda, \quad \beta_j \in l_0
\]
Lasso regularization path

\[
\min_{\beta} \frac{1}{2} \| y - X\beta \|^2 + \lambda \sum_{i=1}^{d} |\beta_j| 
\]

- Let \( \beta_{(l)} = \beta(l_{\beta}) \) and \( X_{\beta} = X_{\beta}(; l_{\beta}) \)
- optimality conditions become

\[
-X_{\beta}^T (X_{\beta}\beta_{(l)} - y) + \lambda \text{sign}(\beta_{(l)}) = 0
\]
Lasso regularization path

- Let $\beta_{\beta} = \beta(I_{\beta})$ and $X_{\beta} = X_{\beta}(:, I_{\beta})$
- Optimality conditions become

$$-X_{\beta}^T (X_{\beta}\beta_{\beta} - y) + \lambda \text{ sign}(\beta_{\beta}) = 0$$

- For $\lambda_t$, assume the solution $\beta_{\beta}^t$ and the corresponding set $I_{\beta}^t$
- Assume $\lambda = \lambda_t + \gamma$ such that $I_{\beta}^t$ and $\text{sign}(\beta_{\beta}^t)$ remain unchanged

\[
\begin{align*}
X_{\beta}^T (X_{\beta}\beta_{\beta}^t - y) &= \lambda_t \text{ sign}(\beta_{\beta}^t) \\
X_{\beta}^T (X_{\beta}\beta_{\beta} - y) &= \lambda \text{ sign}(\beta_{\beta}) \\
\hline
X_{\beta}^T X_{\beta} (\beta_{\beta} - \beta_{\beta}^t) &= (\lambda - \lambda_t) \text{ sign}(\beta_{\beta})
\end{align*}
\]

$$\beta_{\beta} = \beta_{\beta}^t + (\lambda - \lambda_t)w = \beta_{\beta}^t + \gamma w$$

Descent direction $w = (X_{\beta}^T X_{\beta})^{-1} \text{ sign}(\beta_{\beta})$
The linear variation holds until the set $I^t_\beta$ changes \implies detect events

**Event detection**

- $\beta_\ell \in I_\beta$ moves to $I_0$
  - Compute the step size $\gamma$ such as $0 = \beta^t_\ell + \gamma w_j$

- $\beta_j \in I_0$ moves to $I_\beta$
  - Recall $|x_\ell^T (X_{\beta_\beta} - y)| = \lambda$, $\beta_\ell \in I_\beta$ and $|x_j^T (X_{\beta_\beta} - y)| \leq \lambda$, $\beta_j \in I_0$
  - Compute $\gamma$ to obtain the correlation $|x_j^T (X_{\beta_\beta} - y)| = \lambda_t + \gamma$
  - Choose $\beta_j$ as the most correlated variable to the residual i.e. $j = \arg\max_{j \in I_0} |x_j^T (X_{\beta_\beta}^t - y)|$
Algorithm 1 Lasso solution path

Set $t = 0$, $\beta^0 = 0$, $l_\beta = \emptyset$ and $l_0 = \{1, \cdots, d\}$

Find $\beta_j$ to add to $l_\beta$: $j = \arg\max_{j \in l_0} |x_j^\top y|$, $j \in l_0$ (max of correlation)

repeat

Compute the descent direction $w$

Compute the step size $\gamma$

Update the sets $l_\beta$ and $l_0$ according to the event detected ($l_0 \rightarrow l_\beta$ or $l_\beta \rightarrow l_0$)

$t = t + 1$

until termination
the solution in the $X$ space

starting point: all the $\beta$ are set to 0

residual: $\mathbf{R} = X\mathbf{\beta} - y = y$
Interpretation of Lasso path (V. Guigue)

the solution in the $X$ space

projection of the residual error on the active variable
Interpretation of Lasso path (V. Guigue)

The solution in the $X$ space

Stepsize computation, same correlation of residual errors
the solution in the $X$ space

projection of the residual error on the active variable
Interpretation of Lasso path (V. Guigue)

the solution in the $X$ space

stepsize computation
Example (provided by A. Rakotomamonjy)

- Diabetes data set: 10 variables, 442 observations
Linear SVM

Model: \( f(x) = \langle \omega, x \rangle \) (to easy the presentation)

Problem: \[
\min_{\omega} \sum_{i=1}^{n} \max(1 - y_i f(x_i), 0) + \frac{\lambda}{2} \|\omega\|^2
\]

Optimality condition

\[
- \sum_{i=1}^{n} \alpha_i y_i x_i^\top + \lambda \omega = 0
\]

\[
\begin{cases}
\alpha_i = 1 & \text{if } y_i f(x_i) < 1 \\
\alpha_i = 0 & \text{if } y_i f(x_i) > 1 \\
\alpha_i \in [0, 1] & \text{if } y_i f(x_i) = 1
\end{cases}
\]

Sets

\[
l_0 = \{x_i \mid y_i f(x_i) > 1\}, \quad l_1 = \{x_i \mid y_i f(x_i) < 1\}, \quad l_{\alpha} = \{x_i \mid y_i f(x_i) = 1\}
\]
Optimality condition

\[ \sum_{\alpha} \alpha_i y_i x_i^\top + \sum_1 y_i x_i^\top = \lambda \omega \quad \text{with} \quad \alpha_i \in [0, 1] \]
Optimality condition

\[ \sum_{i=1}^{l_\alpha} \alpha_i y_i x_i^\top + \sum_{i=1}^{l_1} y_i x_i^\top = \lambda \omega \quad \text{with} \quad \alpha_i \in [0, 1] \]

Path derivation

- Let \( \lambda_t \rightarrow \text{solution} \alpha_i^t, \quad i \in l_\alpha \), the sets \( l_\alpha, l_0, l_1 \)
- \( \lambda = \lambda_t + \gamma \), such that the sets remain unchanged
- Hence \( \forall x_j \in l_\alpha, \quad f(x_j) = \langle \omega, x_j \rangle = y_j \) \quad (margin points)
Optimality condition
\[ \sum_{i \in I} \alpha_i y_i x_i^\top + \sum_{i \in I_1} y_i x_i^\top = \lambda \omega \quad \text{with } \alpha_i \in [0, 1] \]

Path derivation
- Let \( \lambda_t \to \) solution \( \alpha_i^t, \ i \in I_\alpha \), the sets \( I_\alpha, I_0, I_1 \)
- \( \lambda = \lambda_t + \gamma \) such that the sets remain unchanged
- Hence \( \forall x_j \in I_\alpha, \ f(x_j) = \langle \omega, x_j \rangle = y_j \quad \text{(margin points)} \)

\[
\begin{align*}
\sum_{i \in I} \alpha_i^t y_i k(x_i, x_j) + \sum_{i \in I_1} y_i k(x_i, x_j) &= \lambda_t y_j \quad \text{with } k(x_i, x_j) = \langle x_i, x_j \rangle \\
\sum_{i \in I} \alpha_i y_i k(x_i, x_j) + \sum_{i \in I_1} y_i k(x_i, x_j) &= \lambda y_j \\
G(\alpha - \alpha^t) &= (\lambda - \lambda_t) y_{\alpha} \quad \text{with } G_{ij} = y_i k(x_i, x_j) \\
\alpha &= \alpha^t + (\lambda - \lambda_t) w \\
w &= G^{-1} y_{\alpha}
\end{align*}
\]
The variation holds until the sets change

Event detection
- \( x_i \in l_\alpha \rightarrow l_0 \cup l_1 \)
- \( \alpha_i \) goes to 0 or 1
- \( x_i \in l_0 \cup l_1 \rightarrow l_\alpha \)
- \( y_i f(x_i) \) becomes 1

Linear variation
\[
\alpha = \alpha^t + (\lambda - \lambda_t) w
\]
**SVM regularization path**

**Linear variation**

\[ \alpha = \alpha^t + (\lambda - \lambda^t)w \]

The variation holds until the sets change.

**Event detection**

- \( x_i \in l_\alpha \rightarrow l_0 \cup l_1 \)
- \( \alpha_i \) goes to 0 or 1
- \( x_i \in l_0 \cup l_1 \rightarrow l_\alpha \)
- \( y_if(x_i) \) becomes 1

**Algorithm**

Similar to the algorithm of lasso path

**Remark**

- Nonlinear case: \( \min_{f \in \mathcal{H}} \sum_{i=1}^{n} \max(1 - y_if(x_i), 0) + \frac{1}{2} \|f\|_\mathcal{H}^2 \)
- Use the reproducing property \( \langle f(\cdot), k(x, \cdot) \rangle \) to derive the previous results
SVM regularization path

Dealing with the bias term of SVM model

- SVM model: \( f(x) = \langle \omega, x \rangle + b \)
- Problem: \( \min_{\omega, b} \sum_{i=1}^{n} \max(1 - y_i f(x_i), 0) + \frac{\lambda}{2} \|\omega\|^2 \)

Optimality conditions

- For \( \omega \): \( \sum_{I_\alpha} \alpha_i y_i x_i^\top + \sum_{I_1} y_i x_i^\top = \lambda \omega \) with \( \alpha_i \in [0, 1] \)
- For \( b \): \( \sum_{I_\alpha} \alpha_i y_i + \sum_{I_1} y_i = 0 \) with \( \alpha_i \in [0, 1] \)

Piecewise linear variation

Let \( \alpha_0 = \lambda b \). Using the previous analysis, one gets

\[
\begin{bmatrix}
\alpha \\
\alpha_0
\end{bmatrix} =
\begin{bmatrix}
\alpha^t \\
\alpha_0^t
\end{bmatrix} + (\lambda - \lambda_t) \begin{bmatrix}
G & 1^\top \\
1 & 0
\end{bmatrix}^{-1} \begin{bmatrix}
y_\alpha \\
0
\end{bmatrix}
\]
Nonlinear SVM with gaussian kernel
Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation
## Common points between lasso and SVM path

<table>
<thead>
<tr>
<th>LASSO</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta = 0 )</td>
<td>Initialize ( \alpha )</td>
</tr>
<tr>
<td>( \text{While } l_0 \neq \emptyset ) Move a variable ( x_j ) ( I_0 \leftrightarrow I_\beta ) compute ( w ) compute ( \gamma ) ( \beta = \beta^t + (\lambda - \lambda_t)w )</td>
<td>( \text{While } l_1 \neq \emptyset ) Move a point ( x_j ) ( I_0 \leftrightarrow I_\alpha \leftrightarrow I_1 ) compute ( w ) compute ( \gamma ) ( \alpha = \alpha^t + (\lambda - \lambda_t)w )</td>
</tr>
</tbody>
</table>

### Lesson

Running the path, we select the “good” variables or points and set the other parameters to zero.

### Why this behavior of sparsity?
Definition: strong homogeneity set (variables)

\[ I_0 = \{ j \in \{1, \ldots, d\} \mid \beta_j = 0 \} \]

Theorem

Regular if \( L(\beta) + \lambda P(\beta) \) differentiable and if \( I_0(y) \neq \emptyset \)

\[ \forall \varepsilon > 0, \exists y' \in B(y, \varepsilon) \text{ such that } I_0(y') \neq I_0(y) \]

Singular if \( L(\beta) + \lambda P(\beta) \) NON differentiable and if \( I_0(y) \neq \emptyset \)

\[ \exists \varepsilon > 0, \forall y' \in B(y, \varepsilon) \text{ then } I_0(y') = I_0(y) \]

Singular criteria \( \implies \) sparsity

\( L_1 \) criteria are singular in 0

Singurality provides sparsity

Nikolova, 2000
Roadmap

1. Introduction
2. Regularization path and pareto frontier
3. Efficient regularization path running
4. Two examples of regularization path
5. Regularization path and sparsity
6. Extensions and efficiency evaluation
Extensions of piecewise linear path algorithm

Lasso type

**Seminal paper:** LAR algorithm [Efron et al. 2004]

- Elastic net (double penalization $L_1$ and $L_2$) [Zhou and Hastie, 2005]
- Fused Lasso ($L_1$ and total variation penalizations) [Tibshirani et al. 2005]
- Grouped Lasso [Yuan and Lin, 2006]
- Least absolute deviation regression ($L_1$ loss and penalization) [Wang et al. 2007]
- Non negative garrotte [Yuan and Lin, 2007]
- $L_1$ penalization in infinite dimension [Rosset et al. 2007]
- Graph data and Lasso [Tsuda, 2007]
- ...
Extensions of piecewise linear path algorithm

SVM type

Seminal paper: SVM path [Efron et al. 2004]

- 1-norm SVM (SVM with $L_1$ penalty) [Zhou et al. 2003]
- Asymmetric cost SVM [Bach et al. 2005]
- Doubly regularized SVM [Wang et al. 2006]
- $\nu$-SVM [Loosli et al. 2007]
- SVR [Gunter and Zhu, 2005], [Wang et al. 2006], [Gasso et al., 2007]
- Laplacian Semi-supervised SVM [Wang et al. 2006], [Gasso et al. 2007]
- Oneclass SVM [Rakotomamonjy and Davy 2007]
- Ranking SVM [Zapien et al. 2008]
- ...
Empirical efficiency evaluation

$\nu$-SVR [Gasso et al. 07]

$$\min_{f, \epsilon} \frac{1}{n} \sum_{i=1}^{n} \max(0, |y_i - f(x_i)| - \epsilon) + \nu \epsilon + \frac{\lambda}{2} \| f \|^2 \quad \text{s.t.} \quad \epsilon \geq 0$$

- Two hyperparameters: $\nu$ and $\epsilon$
- $\epsilon$ insensitive tube with $\epsilon$: tube width

Residuals: $r = y - f(x)$

Data
True function
Tube

Gasso (LITIS, EA 4108)  Regularization path and machine learning  Antwerp, 19/09/2008  40 / 44
Empirical efficiency evaluation

\( \nu \text{-SVR [Gasso et al. 07]} \)

\[
\min_{f, \nu} \frac{1}{n} \sum_{i=1}^{n} \max(0, |y_i - f(x_i)| - \epsilon) + \nu \epsilon + \frac{\lambda}{2} \|f\|^2 \\
\text{s.t. } \epsilon \geq 0
\]

- Two hyperparameters: \( \nu \) and \( \epsilon \)
- \( \epsilon \) insensitive tube with \( \epsilon \): tube width

Toy problem

- Gaussian kernel with bandwidth \( \sigma = 0.05 \)
- Run the \( \lambda \)-path for different values of \( \nu \)
- Average over 10 trials
Empirical efficiency evaluation

\( \nu\text{-SVR} \) [Gasso et al. 07]

\[
\min_{f, \epsilon} \frac{1}{n} \sum_{i=1}^{n} \max(0, |y_i - f(x_i)| - \epsilon) + \nu \epsilon + \frac{\lambda}{2} \| f \|^2 \quad \text{s.t.} \quad \epsilon \geq 0
\]

- Two hyperparameters: \( \nu \) and \( \epsilon \)
- \( \epsilon \) insensitive tube with \( \epsilon \): tube width

\( N = 1500 \) samples - Computational time (sec)

<table>
<thead>
<tr>
<th></th>
<th>( \nu = 0.01 )</th>
<th>( \nu = 0.5 )</th>
<th>( \nu = 0.75 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda )-path</td>
<td>1.70 ( \pm ) 0.076</td>
<td>1.95 ( \pm ) 0.03</td>
<td>2 ( \pm ) 0.031</td>
</tr>
<tr>
<td>( \nu )-SVR with warm restart</td>
<td>4.30 ( \pm ) 0.053</td>
<td>21.8 ( \pm ) 0.15</td>
<td>21.15 ( \pm ) 0.12</td>
</tr>
</tbody>
</table>

Computational gain up to 11
Empirical efficiency evaluation

Efficiency of the algorithm: Boston Housing data (UCI repository)

- Multidimensional regression ($x \in \mathbb{R}^{13}$), 506 points
- $N = 406$ samples for training
- Gaussian kernel with different bandwidths $\sigma$
- Run the $\lambda$-path for different values of $\nu$
- Average over 10 trials (random data selection)
Empirical efficiency evaluation

Efficiency of the algorithm: Boston Housing data (UCI repository)

- Multidimensional regression \((x \in \mathbb{R}^{13})\), 506 points
- \(N = 406\) samples for training
- Gaussian kernel with different bandwidths \(\sigma\)
- Run the \(\lambda\)-path for different values of \(\nu\)
- Average over 10 trials (random data selection)

<table>
<thead>
<tr>
<th>Bandwidth (\sigma = 1) - Computational time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\lambda)-path</td>
</tr>
<tr>
<td>(\nu)-SVR with warm restart</td>
</tr>
</tbody>
</table>

Computational gain up to 9
Empirical efficiency evaluation

Efficiency of the algorithm: Boston Housing data (UCI repository)

- Multidimensional regression ($x \in \mathbb{R}^{13}$), 506 points
- $N = 406$ samples for training
- Gaussian kernel with different bandwidths $\sigma$
- Run the $\lambda$-path for different values of $\nu$
- Average over 10 trials (random data selection)

$\sigma = 0.1$ - Computational time (sec)

<table>
<thead>
<tr>
<th></th>
<th>$\nu = 0.01$</th>
<th>$\nu = 0.5$</th>
<th>$\nu = 0.75$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$-path</td>
<td>12.31 $\pm$ 0.34</td>
<td>12.29 $\pm$ 0.44</td>
<td>12.27 $\pm$ 0.38</td>
</tr>
<tr>
<td>$\nu$-SVR with warm restart</td>
<td>51.44 $\pm$ 0.78</td>
<td>51.63 $\pm$ 1.24</td>
<td>51.32 $\pm$ 0.95</td>
</tr>
</tbody>
</table>

Computational gain up to 4
Empirical efficiency evaluation

One-class SVM [Rakotomamonjy et al., 07]

- Level set estimation

\[
\begin{align*}
\min_{\beta, \rho, \xi_i} & \quad \frac{\lambda}{2} \| \beta \|^2 + \sum_{i=1}^{n} \xi_i - \lambda \rho \\
\text{st} & \quad \xi_i \geq 0, \quad x_i^\top \beta \geq \rho - \xi_i \quad \forall i = 1, \ldots, n
\end{align*}
\]
Empirical efficiency evaluation

One-class SVM [Rakotomamonjy et al., 07]

Tab.: Comparing computational time in seconds of alpha seeding and a regularization path approach for computing several level sets

<table>
<thead>
<tr>
<th>Datasets</th>
<th># examples</th>
<th>σ</th>
<th>Alpha Seeding</th>
<th>Reg. Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>credit</td>
<td>653</td>
<td>1</td>
<td>18.1</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>21.4</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>15.8</td>
<td>4.4</td>
</tr>
<tr>
<td>pima</td>
<td>768</td>
<td>1</td>
<td>54.3</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>39.8</td>
<td>20.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>25.5</td>
<td>11.2</td>
</tr>
<tr>
<td>yeast-cyt</td>
<td>1484</td>
<td>1</td>
<td>42.9</td>
<td>49.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>42.6</td>
<td>51.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>42.5</td>
<td>38.9</td>
</tr>
<tr>
<td>spamdata</td>
<td>4601</td>
<td>1</td>
<td>18220</td>
<td>7460</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>2265</td>
<td>1446</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10</td>
<td>1114</td>
<td>1039</td>
</tr>
</tbody>
</table>
Concluding remarks

Summary

- Linear combination of convex criteria $\rightarrow$ Pareto frontier $\equiv$ Regularization path
- Efficient computation of the path and sparsity
- Practical for small and medium data set

Extensions

- Large scale data
- Non convex case
- Stopping on the path especially for more than two criteria
References

- F. Bach, D. Heckerman, and E. Horvitz, On the path to an ideal ROC Curve: considering cost asymmetry in learning classifiers, AISTATS, 2005
- G. Gasso, K. Zapien and S. Canu, Computing and stopping the solution paths for \( \nu \)-SVR, ESANN 2007
- G. Gasso, K. Zapien and S. Canu, Sparsity regularization path for Semi-Supervised SVM, ICMLA 2007
- G. Loosli, G. Gasso and S. Canu, Regularization paths for \( \nu \)-SVM and \( \nu \)-SVR, ISNN 2007
- A. Rakotomamonjy, M. Davy, One-class SVM regularization path and comparison with alpha seeding, ESANN 2007
- S. Rosset, G. Swirszcz1, N. Srebro and Ji Zhu, \( L_1 \) Regularization in Infinite Dimensional Feature Spaces, Colt 2007
R. Tibshirani, S. Rosset, Ji Zhu and K. Knight, Sparsity and Smoothness via the Fused Lasso. JRSSB, 2005.

K. Tsuda, Entire Regularization Paths for Graph Data, ICML 2007


L. Wang, M. Gordon and J. Zhu, Regularized Least Absolute Deviations Regression and an Efficient Algorithm for Parameter Tuning, ICDM’06


M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, JRSSB, 2006.


