Secondo incontro sul tema:

“Prospettive di ricerca allo IASI in Fisiopatologia, Bioinformatica e Biomatematica - Parte II”

Giovanni Felici

Zero-norm regularization of linear regression: could it be useful in the analysis of biological data?

Main results based on paper:

*MIP-BOOST: Efficient and Effective $L_0$ Feature Selection for Linear Regression*,

Linear regression.... Simple, yet...

\[ Y = b_0 + b_1 X_1 + e \]

\[ Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_p X_p + e \]
Linear regression.... Simple, yet...

\[
\text{Loss}(h_w) = \sum_{i=1}^{n} [y_i - (w_0 + w_1 x_i)]^2
\]

\[
\frac{\partial}{\partial w_0} \sum_{i=1}^{n} [y_i - (w_0 + w_1 x_i)]^2 = 0
\]

\[
\frac{\partial}{\partial w_1} \sum_{i=1}^{n} [y_i - (w_0 + w_1 x_i)]^2 = 0
\]

- \( p \) variables, or features, \( n \) observations / samples

\[
Y = X^t \beta + \varepsilon
\]

\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X^t \beta\|_2
\]

\[
\hat{\beta} = (X^t X)^{-1} X^t Y
\]

- To invert covar matrix,

\[
P \geq p
\]

- What if \( p \) is large, \( p >> n \)?

- **Sparsification / Feature Selection**: use few terms
- **Regularization**: force coefficients to be “nice” in size
In biological applications

• $p >> n$? May happen...

• Gene expression data
• CNV data
• Clinical data
• Proteomics, omics in general

-> very often, the variables are many many, the samples are few few (they have indeed a large cost!)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + ... + b_p X_p + e$$

• Do we need all $p$ of them for a "good" regression?
• Are some variables "correlated" (sort of, equal)
• Don’t we overfit with all these variables?
• Is our model robust enough with all these variables? Noise is always around the corner...
Regularization = sparisification?

$$\min_{\beta} \| Y - X^t \beta \|_q$$

s.t. $\| \beta \|_r \leq s$

$\Rightarrow$

$$\min_{\beta} \| Y - X^t \beta \|_q + \lambda \| \beta \|_r$$

Error  
Penalty

$r = 2 \quad \rightarrow \textit{Ridge Regression} \quad (\text{Tikhonov regularization})$

$r = 1 \quad \rightarrow \textit{Lasso}$

$R = 1 \& 2 \quad \rightarrow \textit{Elastic Net}$


Do RR or LASSO really *sparsify* the model?

- Going from 2-norm to 1-norm cuts out portions of the feasible region where all coefficients have similar values.
- The solution is pushed towards the vertices of the feasible region.
- Sparsification.
\[ \left\| \beta \right\|_r \leq s \]
\[ \sqrt{\sum_i \beta_i^2} \leq 1 \]
\[ \left| \sum_i \beta_i \right| \leq 1 \]
\[ \left( \sum_i \beta_i^{1/2} \right)^2 \leq 1 \]

\[ r = r' < 1 \]
\[ r < r' < 1 \]
\[ r = 0 \]

\[ \left\| \beta \right\|_0 = \text{Number of nonzero elements} \]
“0-norm” as a Mixed Integer Linear Problem

✧ Optimize under 0-norm constraints: EASY ? NO, exponential WC complexity
✧ IMPOSSIBLE ? NEITHER...
✧ Similar to subset selection: what is the subset of variables of given dimension k that minimizes my error?

Much easier using 1-norm for the OF (error)

Least Absolute Deviation (LAD), Robustness, etc...

\[
\begin{align*}
\text{min} & \quad \|Y - X^t \beta\|_2 \\
\text{s.t.} & \quad \|\beta\|_0 \leq s
\end{align*}
\]

\[
\begin{align*}
\text{min} & \quad \|Y - X^t \beta\|_1 \\
\text{s.t.} & \quad \|\beta\|_0 \leq s
\end{align*}
\]
“norm-0” as a MIQP

\[
\begin{align*}
\min & \quad \sum_{i=1,n} e_i^2 \\
\text{s.t.} & \quad Y_i - \sum_{j=1,p} \beta_j X_{ij} \leq e_i, \quad i = 1, n \\
& \quad Y_i - \sum_{j=1,p} \beta_j X_{ij} \geq -e_i, \quad i = 1, n \\
& \quad \beta_j \geq -M z_j, \quad j = 1, p \\
& \quad \beta_j \leq M z_j, \quad j = 1, p \\
& \quad \beta_j \in R, \ z_j \in \{0,1\}, \quad j = 1, p \\
\sum_{j=1,p} z_j & \leq k
\end{align*}
\]

**CONS**
- Use binary variables to indicate if coefficients are used (1) or not (0)
- Numerically unstable constraints to link the choice and the value
- NP complexity class

**PROS**
- Extremely flexible
- Direct control of sparsity
- Many extensions
- Powerful optimization software to solve it efficiently
- Powerful heuristics
min $\sum_{i=1,n} e_i^2$

s.t. 

$Y_i - \sum_{j=1,p} \beta_j X_{ij} \leq e_i, \ i = 1, n$

$Y_i - \sum_{j=1,p} \beta_j X_{ij} \geq -e_i, \ i = 1, n$

$\beta_j \geq -Mz_j, \ j = 1, p$

$\beta_j \leq Mz_j, \ j = 1, p$

$\beta_j \in R, z_j \in \{0,1\}, \ j = 1, p$

$\sum_{j=1,p} z_j \leq k$

Not a new idea

- TSato, Takano, Miyashiro, Yoshise,(2016) Feature subset selection for logistic regression via mixed integer optimization, Computational Optimization and Applications 64 (3)
What we get for free:

\[ \sum_{j=1,p} z_j \leq k \]

\[ \sum_{j \in G_1} z_j \leq k_1, \sum_{j \in G_2} z_j \leq k_2 \]

\[ z_i + z_j \leq 1 \]

\[ z_i \leq z_j \]

\[ \sum_{j \in G} z_j \geq k' \]

Well, not for free: at computational cost

\[ \sum_{j \in G} c_j z_j \leq B \]
Would not it be nice?

$p = 100, (\rho, \omega) = (0.5, 0.4)$ with SNR of 1

Prospettive di ricerca allo IASI in Fisiopatologia, Bioinformatica e Biomatematica - Parte II, 16.06.2020
3 tricks

1. Bisection with feelers and randomly added variables
2. Integrated cross-validation
3. Whitening for feature selection
Sprser and Better solutions

Prospettive di ricerca allo IASI in Fisiopatologia, Bioinformatica e Biomatematica - Parte II, 16.06.2020
Time

Low Dimension

<table>
<thead>
<tr>
<th>Total Time (Hours)</th>
<th>CV10m</th>
<th>CV3m</th>
<th>M-B</th>
<th>M-B+</th>
<th>M-B++</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>29.94</td>
<td>13.03</td>
<td>2.61</td>
<td>1.23</td>
<td>0.15</td>
</tr>
</tbody>
</table>

High Dimension

<table>
<thead>
<tr>
<th>Total Time (Hours)</th>
<th>CV20m</th>
<th>CV3m</th>
<th>M-B</th>
<th>M-B+</th>
<th>M-B++</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>150.61</td>
<td>24.61</td>
<td>25.47</td>
<td>13.15</td>
<td>10.52</td>
</tr>
</tbody>
</table>

Number of Features

<table>
<thead>
<tr>
<th>Number of Features</th>
<th>CV10m</th>
<th>CV3m</th>
<th>M-B</th>
<th>M-B+</th>
<th>M-B++</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prospettive di ricerca allo IASI in Fisiopatologia, Bioinformatica e Biomatematica - Parte II, 16.06.2020
### Applications: Diabetes (Efron et al. (2004)), Bodyfat (Penrose et al. (1985))

**Diabetes**
- \( n = 442 \) diabetes patients
- \( 10 \) baseline predictors (age, body mass index, average blood pressure, six blood serum measurements and sex)
- \( p = 64 \) features including 45 pair-wise products and 9 squared terms
- \( Y = \) disease progression

**Bodyfat**
- \( n = 252 \) patients
- \( 13 \) predictors (age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrists circumference)
- \( Y = \% \text{ of body fat (based on density...)} \)
- Inclusion of interactions on poly terms
- \( p = 559 / 679 \)

**Lamotrigine (LTG)**
- Body Mass Index
- Mean Arterial Pressure

---

<table>
<thead>
<tr>
<th>Procedure</th>
<th>Diabetes</th>
<th>Bodyfat</th>
<th>Bodyfat with Density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{k}_0 )</td>
<td>VALMSE</td>
<td>( \hat{k}_0 )</td>
</tr>
<tr>
<td>LMN</td>
<td>19</td>
<td>0.488</td>
<td>15</td>
</tr>
<tr>
<td>LSD</td>
<td>9</td>
<td>0.501</td>
<td>4</td>
</tr>
<tr>
<td>RL</td>
<td>19</td>
<td>0.488</td>
<td>15</td>
</tr>
<tr>
<td>FS</td>
<td>7</td>
<td>*0.477</td>
<td>2</td>
</tr>
<tr>
<td>MIP-BOOST</td>
<td>3</td>
<td>0.496</td>
<td>13</td>
</tr>
</tbody>
</table>

*: minimum Validation MSE.

- LMN lasso with min.
- LSD lasso with parsim.
- RL relaxed lasso
- FS forward selection
- 25% validation

**Without density:** some combination of age, weight, height, abdomen, and wrist circumference
Thank you