Methods of Model Selection and Dimensionality Reduction

Vadim STRIJOV

Computing Center of the Russian Academy of Sciences

Ankara, October 06, 2009
Institute of Applied Mathematics, METU
**Dimensionality reduction and index construction problem**

- There is a set of objects, i.e. power plants:
  - Beckjord
  - East Bend
  - Miami Fort
  - Zimmer

- The **index** is a measure of an object’s quality. It is a scalar, corresponded to an object.

- Expert estimation of an object’s quality could be an index, too.
## Examples

<table>
<thead>
<tr>
<th>Index name</th>
<th>Objects</th>
<th>Features</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOEFL</td>
<td>Students</td>
<td>Tests</td>
<td>Sum of scores</td>
</tr>
<tr>
<td>Eurovision</td>
<td>Singers</td>
<td>Televotes, Jury votes</td>
<td>Linear (weighted sum)</td>
</tr>
<tr>
<td>S&amp;P500, NASDAQ</td>
<td>Time ticks</td>
<td>Shares (prices, volumes)</td>
<td>Non-linear</td>
</tr>
<tr>
<td>Bank ratings</td>
<td>Banks</td>
<td>Requirements</td>
<td>By an expert commission</td>
</tr>
<tr>
<td>Kyoto-index</td>
<td>Power plants</td>
<td>Greenhouse gases</td>
<td>Linear</td>
</tr>
</tbody>
</table>
There are lots of ways to construct indices. However, when algorithms are chosen and some results obtained, the following question arises:

- **How to show adequacy of the calculated indices?**

To answer the question analysts invite experts. The experts express their opinion and then the second question arises:

- **How to show that expert estimations are valid?**
How to construct an index?

- Assign a comparison criterion.
- Gather a set of comparable objects.
- Gather features of the objects.
- Make a data table: objects/features, i.e.

<table>
<thead>
<tr>
<th>#</th>
<th>Plant Name</th>
<th>Plant Type</th>
<th>Total Net Generation</th>
<th>CO2 emission</th>
<th>NOx emission</th>
<th>SOx emission</th>
<th>Population density</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>10^6 KWHours</td>
<td>Short tons per month</td>
<td>Short tons per month</td>
<td>Short tons per month</td>
<td>Qty per sqmile</td>
</tr>
<tr>
<td>1</td>
<td>Beckjord</td>
<td>Coal</td>
<td>458505</td>
<td>191</td>
<td>16</td>
<td>45</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>East Bend</td>
<td>Coal</td>
<td>356124</td>
<td>147</td>
<td>16</td>
<td>43</td>
<td>34</td>
</tr>
<tr>
<td>4</td>
<td>Miami Fort</td>
<td>Coal</td>
<td>484590</td>
<td>204</td>
<td>6</td>
<td>23</td>
<td>45</td>
</tr>
<tr>
<td>5</td>
<td>Dark Creek</td>
<td>Coal</td>
<td>818435</td>
<td>329</td>
<td>5</td>
<td>64</td>
<td>34</td>
</tr>
</tbody>
</table>

Optimal value: max min min min min min

The criterion could be: Ecological footprint of a plant
Notations

\[ A = \{a_{ij}\} \quad - \quad (n \times m) \text{ real matrix, data set,} \]
\[ q = [q_1, \ldots, q_m]^T \quad - \quad \text{vector of object indices,} \]
\[ w = [w_1, \ldots, w_n]^T \quad - \quad \text{vector of feature importance weights,} \]
\[ q_0, w_0 \quad - \quad \text{expert estimations of indices and weights.} \]

Usually, data prepared so that
1. the minimum of each feature equals 0, while the maximum equals 1;
2. the bigger value of each implies better quality of the index.
The first method, **Pareto slicing**

An easiest method to obtain indices in ordinal scales is to find non-dominated objects at each slicing level.

The object $a$ is non-dominated if there is no $b_i$ such that $b_{ij} \geq a_i$ for all features $j$. 
Supervised way-1, the Weighted sum

\[ q_1 = A w_0. \]
Unsupervised way,

**Principal Components Analysis**

\[ Q = AW, \quad \text{where } W \text{—rotation matrix of the principal components.} \]

\[ q_{PCA} = Aw_{1PC}, \quad \text{where } w_{1PC} \text{ is the 1st column vector of } W. \]

PCA gives minimal mean square error between objects and their projections.
Unsupervised way, useful tool for PCA

\[ A = ULW^T \]

\[ A^T A = WLU^T ULW^T \]

\[ A^T AW = WL^2 \]
Supervised way-2, the Expert-Statistical Technique

\[ w_1 = \arg \min ||q_0 - A w||^2, \]

least squares, \[ w_1 = (A^T A)^{-1} A^T q_0. \]
The problem of specification

- We have the data table $A$, expert estimations $q_0, w_0$, calculated weights and indices $q_1, w_1$.

- Contradiction

Calculated indices are not the same as the expert estimations for the indices; as well, calculated weights are not the same as the expert estimations of the weights:

in general,
Linear specification

\[ w_\alpha = \alpha A^+ q_0 + (1-\alpha) w_0, \quad q_\alpha = (1-\alpha) A w_0 + \alpha q_0. \]

Parameter \( \alpha \) is in \([0,1]\).

\( \alpha = 0 \), we trust expert estimations of the weights,
\( \alpha = 1 \), we trust expert estimations of the indices.
Quadratic specification

If parameter $\gamma^2$ is 0, then we trust expert estimations of the indices.
Comparison of the methods, what is the difference?
**Ordinal specification**

\[ \mathbf{w}_0 = [w_1 \geq w_2 \geq \ldots \geq w_n \geq 0]^T, \mathbf{q}_0 = [q_1 \geq q_2 \geq \ldots \geq q_m \geq 0]^T. \]
Rank-scaled expert estimations

\[ w_0 = [w_1 \geq w_2 \geq ... \geq w_n \geq 0]^T, q_0 = [q_1 \geq q_2 \geq ... \geq q_m \geq 0]^T. \]

\[ Q_0 = \{q_0 \mid J_m q_0 \geq 0\}, \]
\[ W_0 = \{w_0 \mid J_n w_0 \geq 0\}. \]

\[ J = \begin{pmatrix} 1 & -1 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \]
The cones intersection exists

\[ q_1 \in AW_0 \cap Q_0, \]

or not, then specify

\[ q_\alpha = (1 - \alpha)Aw' + \alpha q', \quad \text{where} \]

\[ w', q' = \arg \min_{w \in W_0, \|w\|^2 = 1} \|A w - q\|^2. \]

\[ q \in Q_0, \|q\|^2 = 1 \]
Pair-wise comparison

If an object in a row is better than the other one in a column then put “+”, otherwise “-”.

Make a graph, row + column means row ——— column. Find the top and remove extra nodes.
The results of the specification are

- adequate indices,
- reasoned expert estimations.

We know why our expert valued each object and what contribution each feature makes to the index.
Problem statement

Model generation

Model selection

---

Model selection for (generalized) linear models

Let there be given

1. Sample set:
   \( \{(x_i, y_i) | i = 1, \ldots, \ell \} \), where \( x_i \in \mathbb{R}^P, y_i \in \mathbb{R}^1, P = |N| \)
   - and \( N \subset \mathbb{N} \).

2. Linear model:
   
   \[
   y = f(w, x) + \varepsilon, \\
   y = \langle w, x \rangle + \varepsilon.
   \]

3. Data generation hypothesis:
   distribution of the random variable \( \varepsilon_i \) is in the exponential family.

4. Target function:
   minimum of the residual vector norm
   
   \[
   SSE = \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \rightarrow \min.
   \]
find a subset $\mathcal{A} \subset N$ of the indices $\hat{x} = \{x_j^i | j \in \mathcal{A}\}$, such that the model $f(w, \hat{x})$ brings the optimum to the given criterion.

For example to the Colin Mallows’ $C_P$:

$$C_P = \frac{SSE_P}{RMS} - \ell + 2P,$$

where

$$RMS = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(w, x_i))^2.$$

Or to another criterion from the following list.
Criterions for model selection

1. Information criterions
   - Akaike Information Criterion, $AIC = 2P - 2\ln(S)$
   - Bayesian Information Criterion, $BIC = P \ln(\ell) - 2\ln(S)$

2. Cross-validation
   - Retrospective Forecasting
   - Leave One Out
   - Random Split, etc.

3. Model Comparison
   - Bayesian Comparison
   - Minimum Description Length
Let there be given

- $\Xi = \{\xi^u\}_{u=1}^U$ — set of measured (nongenerated) independent variables,

- $G = \text{id} \cup \{g_v\}_{v=2}^V$ — finite set of primitive functions.

Consider Cartesian product $G \times \Xi$. An element $(g_v, \xi^u)$ corresponds to the superposition $g_v(\xi^u)$ and defined by indices $v, u$. Denote $s_\iota = g_v(\xi^u)$, where the index $\iota = (v - 1)U + u$.

Consider $S \times S \times \ldots \times S$ — Cartesian product $\mathcal{N}$ of the sets $S = \{s_\iota\}$. Each element of $\mathcal{N}$ bijectively corresponds to the superposition $a_i = s_\iota^1 \circ s_\iota^2 \circ \ldots \circ s_\iota^N$. 
The basic model of the feature generation is

\[
y = w_0 + \sum_{i=1}^{UV} w_i a_i + \sum_{i=1}^{UV} \sum_{j=1}^{UV} w_{ij} a_i a_j + \cdots + \sum_{i=1}^{UV} \cdots \sum_{z=1}^{UV} w_{i\ldots z} a_i \ldots a_z,
\]

where the coefficients

\[
w = (w_0, w_i, w_{ij}, \ldots, w_{i\ldots z})_{i,j,\ldots,z=1,\ldots,UV}.
\]

Represent this series as

\[
y = \sum_{j \in N} w_j x^j.
\]

The variables \(\{x^j\}\) bijectively correspond to monomials of the polynomial.
The model selection algorithms

Exhaustive search and modifications

1. Exhaustive search of $2^P$ models
2. Method of group data handling, $K \cdot C_P^2$ models
3. Genetic algorithms
4. Add (append a feature), $P(P - 1)/2$ models
5. Del (eliminate a feature)
6. Add-del or stepwise regression, $\sim P^2$ models

Parameter space analysis

1. Least angle regression (LARS), Lasso
2. Optimal brain surgery
The basic linear model includes all independent variables

\[ y = w_0 + \alpha_1 w_1 x_1 + \alpha_2 w_2 x_2 + \ldots + \alpha_R w_P x_P. \]

The hyperparameter \( \alpha \in \{0, 1\} \) is included for the model. The exhaustive search

\[
\begin{array}{cccc}
\alpha_1 & \alpha_2 & \ldots & \alpha_P \\
1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 1 & \ldots & 1 \\
\end{array}
\]
Add (append a feature)

Step 0.
The active set $\mathcal{A}_0 = \emptyset$, and $N$ is the set if feature indices, $P = |N|$.

Step $k = 1, \ldots, P$.
Select the next best feature index

$$\hat{j} = \arg \min_{j \in P \setminus \mathcal{A}_k} \min_{w \in \mathbb{W}_k} \| (X_{\mathcal{A}_k} : \mathbf{x}_j)w - y \|_2^2,$$

then

$$\mathcal{A}_{k+1} = \mathcal{A}_k \cup \hat{j}.$$
Assume the following

The column vectors

\[ x^j = \{ x_i^j | i \in 1, \ldots, \ell \} \quad \text{and} \quad y = \{ y_i | i \in 1, \ldots, \ell \}. \]

The model

\[ y = w_1 x^j + \ldots + w_P x^P + \epsilon, \]

in the other words,

\[ y = Xw + \epsilon. \]

Assume for all \( j \in N \)

\[ \| x^j \|_1 = 0, \quad \| x^j \|_2 = 1 \quad \text{and} \quad \| y \|_1 = 0, \quad \| y \|_2 = 1. \]

For all \( j, k \in N, j \neq k \) the vectors \( x^j, x^k \) are linear independent.

Then the vector of correlation coefficients

\[ c = X^T y. \]
**Fast orthogonal search**

**Step 0.**
The residuals $\varepsilon_0 = 0$, the active set $\mathcal{A}_0 = \emptyset$.

**Step $k = 1, \ldots, P$.**

$$\mathcal{A}_k = \mathcal{A}_{k-1} \cup \hat{j},$$

where $\hat{j}$ — feature, which has maximum correlation with $\varepsilon_k$:

$$\hat{j} = \arg \max_{j \in \{\mathcal{N} \setminus \mathcal{A}_k\}} \frac{\langle w, x^j \rangle}{\|x\| \|\varepsilon_k\|},$$

and

$$\varepsilon_k = X_A w_A - \varepsilon_{k-1}.$$
Fast orthogonal search

\[ Xw - y = \varepsilon \]

\[ w_1x^1 - y = \varepsilon_1 \]
Least angle regression, LARS

Denote $\mu = Xw$.

Step 0.
$\mu_0 = 0$, residual vector $\varepsilon_0 = y - \mu_0$.

Step 1.
Let $y$ has greater correlation with $x^1$ than with $x^2$. Then the new value of $\mu_1 = \mu_0 + w_1x^1$, where $w_1$ is chosen so, that the vector $y_2 - \mu$ — is a bisector for the vectors $x^1, x^2$.

Step 2.
For the unit bisector $u_2$ calculate $w_2$:

$$\mu_2 = \mu_1 + w_2u_2 = y_2 \quad \text{for P=2}.$$
Least angle regression, LARS
Lasso

Minimize the error

$$\|Xw - y\|_2^2 \rightarrow \min,$$

subject to

$$\sum_{j \in N} \|w_j\|_1 \leq T.$$ 

Theorem (Efron et al., 2004).
Assuming the «one at time» condition, the LARS algorithm yields all Lasso solutions.
Lasso and LARS
Optimal brain surgery

- Approximate $SSE = S(w)$:
  $$S(w + \Delta w) = S(w) + g^T(w)\Delta w + \frac{1}{2} \Delta w^T H \Delta w + o(\|w\|^3).$$
- Elimination a feature is equivalent to $e_i^T \Delta w + w_i = 0$.
- Minimize the quadratic form $\Delta w^T H \Delta w$ subject to $e_i^T + w_i = 0$, for all $i$.
- The index of the eliminated feature is $i = \arg \min_i (\min_{\Delta w} (\Delta w^T H \Delta w | e_i^T + w_i = 0))$.
- Introduce Lagrange function $S = \Delta w^T H \Delta w - \lambda (e_i^T + w_i)$.
- For all $i$ $\Delta w = -\frac{w_i}{[H^{-1}]_{ii}} H^{-1} e_i$.
- The salience of the target function is $L_i = \frac{w_i^2}{2[H^{-1}]_{ii}}$. 
Optimal brain surgery
Optimal brain surgery
Ventia non sunt multiplicanda praeter necessitatem

Occam’s razor: entities (model elements) must not be multiplied beyond necessity

William of Ockham
1285-1349