Learning high-dimensional data

Michel Verleysen
Université catholique de Louvain (Louvain-la-Neuve, Belgium)
Electricity department

January 2002

Acknowledgements

Part of this work has been realized in collaboration with PhD students:
- Philippe Thissen
- Jean-Luc Voz
- Amaury Lendasse
- John Lee

Some ideas and figures come from
Data mining

\[\begin{array}{cccccc}
1.2 & 7.5 & -1.9 & 2 & \cdots & 1.9 \\
-7.6 & 12 & 17.2 & 2.4 & \cdots & 1.5 \\
-8.5 & 13 & 14 & 8.5 & \cdots & -1.9 \\
9 & 5.4 & -5.2 & 8.2 & \cdots & 9.4 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
25.1 & 5.2 & -9.1 & -8.5 & \cdots & 5.4 \\
\end{array}\]

\(N\) lines
(number of observations)

\(D\) columns (dimension of space)

\(D\) and/or \(N\)

Data mining: find information in large databases

High-Dimensional spaces

Inputs = High-Dimension (HD) vectors \((D\) is large)
- \(D\) many parameters in the model
- \(D\) local minima
- \(D\) slow convergence

The questions

- Learning algorithms in HD spaces ?
- Local learning or not ?

The arguments

- real data are seldom HD (concept of intrinsic dimension)
- local learning is not worse than global learning…
Contents

- High-dimensional data
  - Surprising results
  - Intrinsic dimension

- Local learning
  - Use of distance measures

- Dimension reduction
  - Non-linear projection
  - Application to time-series forecasting
Data mining: large databases

\(\text{(biomedical) signals}\)

![Diagram of biomedical signals]

Data mining: large databases

\(\text{financial data}\)

![Diagram of financial data]
Data mining: large databases

- imagery

![Image of faces]

- recordings of consumers’ habits (credit cards)
- bio- data (human genome, etc.)
- satellite images
- hyperspectral images
- …
John Wilder Tukey


« Analyze data rather than prove theorems… »

- In other words:
  - data are here
  - they will be coming more and more in the future
  - we must analyze them
  - with very humble means
  - insistence on mathematics will distract us from fundamental points

Empty space phenomenon

- Necessity to fill space with learning points
- # learning points exponential with dimension
Example: Silvermann’s result

- How to approximate a Gaussian distribution with Gaussian kernels
- Desired accuracy: 90%

![Graph showing the relationship between # points and dim]

Surprising phenomena in HD spaces

- Sphere volume
- Sphere volume / cube volume
- Embedded spheres (radius ratio = 0.9)
Gaussian kernels

\[ \text{1-D Gaussian} \]

\[ \text{% points inside a sphere of radius 1.65} \]

Concentration of measure phenomenon

\[ \text{Take all pairwise distances in random data} \]
\[ \text{Compute the average } A \text{ and the variance } V \text{ of these distances} \]
\[ \text{If } D \text{ increases then} \]
\[ \text{ } V \text{ remains fixed} \]
\[ \text{ } A \text{ increases} \]
\[ \text{All distances seem to concentrate !!!} \]

\[ \text{Example: Euclidean norm of samples} \]
\[ \text{average } A \text{ increases with } (D)^{0.5} \]
\[ \text{variance } V \text{ remains fixed} \]
\[ \text{samples seem to be normalized !} \]
Intrinsic dimension

- No definition
- Example:


Estimation of intrinsic dimensions 1/4

<table>
<thead>
<tr>
<th>method</th>
<th>description</th>
</tr>
</thead>
<tbody>
<tr>
<td>local PCA</td>
<td>a posteriori estimation</td>
</tr>
<tr>
<td>box counting</td>
<td>Grassberger-Proccacia</td>
</tr>
</tbody>
</table>

- small region $\rightarrow$ approximately plane
- PCA applied on small regions
Estimation of intrinsic dimensions 2/4

\begin{center}
\begin{tabular}{|c|c|}
\hline
local PCA & a posteriori estimation \\
\hline
box counting & Grassberger-Proccacia \\
\hline
\end{tabular}
\end{center}

From P. De martines, op. cit.

Estimation of intrinsic dimensions 3/4

\begin{center}
\begin{tabular}{|c|c|}
\hline
local PCA & a posteriori estimation \\
\hline
box counting & Grassberger-Proccacia \\
\hline
\end{tabular}
\end{center}

\begin{itemize}
\item Similar to box counting
\item Mutual distances between pairs of points
\item Advantage: $N(N-1)/2$ distances for $N$ points
\item $\log N - \log r$ graph: number $N$ of pairs of points closer than $r$
\end{itemize}
Estimation of intrinsic dimensions 4/4

- local PCA
- box counting
- Grassberger-Proccacia

![Diagram showing dimension reduction and neural network with performances OK?]

Example: forecasting problem

Intrinsic dimension:
limitation of the concept

- Seen from very close
- Seen from very far
Contents

- High-dimensional data
  - Surprising results
  - Intrinsic dimension
- Local learning
  - Use of distance measures
- Dimension reduction
  - Non-linear projection
  - Application to time-series forecasting

Local learning

- "Local": by means of local functions
- Typical example: Gaussian kernels
- Radial Basis Function Networks

$$f(x) = \sum_{j=1}^{p} w_j K_j(x)$$

$$K_j(x) = \exp\left(-\frac{||x - c_j||^2}{\sigma_j^2}\right)$$

- Local and global: ANN are interpolators, and do not extrapolate to regions without learning data!
Radial-Basis Function Networks (RBFN) for approximation

\[ f(x) = \sum_{j=1}^{P} w_j K_j(x) \]

\[ K_j(x) = \exp \left( -\frac{|x-c_j|^2}{h_j^2} \right) \]

\(\nabla\) Advantages (over MLP, ...):
- splitted computation of centres \( c_j \)
- widths \( h_j \)
- multiplying factors \( w_j \)
- easier learning

Local learning 2/2

\(\nabla\) Sum of sigmoids = Gaussian!
Problem with (local ?) learning

- Most ANN use distances between input and weight vectors:
  - RBFN: as argument to radial kernels
  - VQ and SOM: to choose the « winner »
  - first layer in MLP: \( \mathbf{x} \mathbf{w} \)
- In high-dimensional spaces: all these distances seem identical !
  (concentration of measure phenomenon)
- → need for
  - other neural networks
  - same neural networks, with other distance measures

Using new distance measures

- Example: RBF
  \[
  t(\mathbf{x}) = \sum_{j=1}^{P} w_j K_j(\mathbf{x}) \quad K_j(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{c}_j||^2}{\sigma_j^2}\right)
  \]
- Use of « super-Gaussian » kernels:
  \[
  K'_j(\mathbf{x}) = \exp\left(-\frac{||\mathbf{x} - \mathbf{c}_j||}{\sigma'_j}\right)
  \]
Not assuming evidence...

◇ widths of kernels in RBF are not necessarily equal to the STDV of samples in the Voronoi zone!

\[
\sigma_j = \text{STDV(points in Voronoi zone)}
\]

\[
K_j(x) = \exp\left(-\frac{|x - c_j|^2}{\sigma_j^2} \right)
\]

Contents

◇ High-dimensional data
◇ Surprising results
◇ Intrinsic dimension

◇ Local learning
◇ Use of distance measures

◇ Dimension reduction
◇ Non-linear projection
◇ Application to time-series forecasting
Dimension reduction

- reduced # inputs
- reduced # model parameters
- reduced overfitting

Dimension reduction & intrinsic dimension

Questions:
- intrinsic dimension is unknown
- non-linear submanifolds
Kohonen maps

- Based on topology preservation

Distance-preservation methods

- Many non-linear projection methods: based on distance preservation between input and output pairs

- Examples
  - Multi-dimensional scaling (MDS)
  - Sammon’s mapping
  - Curvilinear Component Analysis

- Principle: place points in the output space, so that pairwise distances are as equal as possible to the corresponding pairwise distances in the input space
  - Impossible to respect all pairwise distances → insist on small ones (local pairs)
Curvilinear Component Analysis

未经授权，不得以任何方式复制或转载。
Application of dimension reduction to forecasting tasks

- Regressor: - past values $x(t-i)$
  - exogenous data $in(j)$

- Forecasting
  $x(t+1) = f(x(t), x(t-1), ..., x(t-k), in(1), in(2), ..., in(l))$

- Non-linear forecasting:
  1. optimise regressor on linear predictor
  2. use the same regressor with non-linear predictor $f$
  3. trials and errors (computational load !)
Forecasting: selection of input variables

- Starting with many input variables, then reduce their number
- Two options:
  1. selection of input variables
     - interpretability
     - limited to existing variables
  2. projection of input variables
     - linear: PCA
     - non-linear: CCA, Kohonen, etc.

Forecasting: Taken’s theorem 1/2

- Takens’ theorem:
  \( q \leq \text{size of regressor} \leq 2q+1 \) (AR model)
Forecasting: Taken’s theorem 2/2

- Takens’ theorem:
  \[ q \leq \text{size of regressor} \leq 2q+1 \]

- In the \(2q+1\) space, there exists a \(q\)-surface without intersection points

- Projection from \(2q+1\) to \(q\) possible!

Forecasting: 1st example 1/2

- Artificial series
  \[ x(t+1) = ax(t)^2 + bx(t - 2) + \varepsilon(t) \]
  Two past values!

- Linear AR model

Sum of errors (on 1000 samples)

AR order
Forecasting: 1st example 2/2

- Non-linear AR model
  - initial regressor: size=6
  - intrinsic dimension: 2
  - CCA from dim=6 to dim=2
  - MLP on 2-dim data

![Graph showing sum of errors for AR order]

Forecasting: 2nd example 1/2

- Daily returns of BEL20 index

- 42 indicators from inputs and exogenous variables:
  - returns: $x_t, x_{t-10}, x_{t-20}, x_{t-40}, \ldots, y_t, y_{t-10}, \ldots$
  - differences of returns: $x_t-x_{t-5}, x_{t-5}-x_{t-10}, \ldots, y_t-y_{t-5}$
  - oscillators: $K(20), K(40), \ldots$
  - moving averages: $MM(10), MM(50), \ldots$
  - exponential moving averages: $MME(10), MME(50), \ldots$
  - etc
Forecasting: 2nd example 2/2

- Method:
  - 42 indicators
  - PCA → 25 variables
  - Grassberger-Proccacia: intrinsic dimension = 9
  - CCA → 9 variables
  - RBF → forecasting
- Result: % of correct approximations of sign (90-days average)
- In average: 57.2% on test set

Conclusion

- High dimensions: > 3 !
- NN in general: also difficulties in high dimensions
- Common problems:
  - Euclidean distance
  - empty space phenomenon
- Towards solutions…
  1. Local NN (RBFN, etc.): easier learning
  2. Generic methods for non-linear projections: reduce dimension
- Open perspectives:
  - using dimension reduction techniques based on topology (and not on distances)
  - study the possibility of non-Euclidean distances and non-Gaussian kernels