A short tutorial on Feature Selection

Michel Verleysen
Machine Learning Group
Université catholique de Louvain
Louvain-la-Neuve, Belgium
michel.verleysen@uclouvain.be
Outline

• Motivation: High-dimensional data, concentration of the norm

• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods

• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
HD data exist!

- Enhanced data acquisition possibilities → many HD data!
  - classification - clustering - regression

Motivation

Modeling → Admissible alcohol level

Known information:

- Predicted alcohol concentration

DIM = 256
HD data exist!

- Enhanced data acquisition possibilities
  → many HD data!
  classification - clustering - regression

DIM = 16384

HD data exist!

- Enhanced data acquisition possibilities
  → many HD data!
  classification - clustering - regression

Motivation

\[ y = f(x_{t-DIM+1}, \ldots, x_{t-1}, x_t) \]
Motivation

Generic data analysis

When I find myself in times of trouble
Mother Mary comes to me
Speaking words of wisdom, let it be...

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>When</td>
<td>1</td>
</tr>
<tr>
<td>Times</td>
<td>1</td>
</tr>
<tr>
<td>Trouble</td>
<td>1</td>
</tr>
<tr>
<td>Let</td>
<td>65</td>
</tr>
<tr>
<td>wisdom</td>
<td>1</td>
</tr>
</tbody>
</table>

DNA sequence:

- CC GC CG GG
- AC TC AG TG
- CA GA CT GT
- AA TA AT TT

- Number of variables or features
- Number of observations

Analysis
Models
High-dimensional data

\[
data = \begin{pmatrix} 0.32 & 2.5 & -0.01 & -3.7 & \ldots & 12.1 \end{pmatrix}^T \in \mathbb{R}^d
\]

- Data are described in a normed vector space (Euclidean space)
- Tools derived from algebra & geometry

Principal Component Analysis

k-Nearest Neighbours
Motivation

High-dimensional data

- Situations we can imagine, represent, draw
- Strong intuition of how the tools behave
- Consider cases where #observations $>> d$
High-dimensional data

- Situations we can imagine, represent, draw
- No representation

- Strong intuition of how the tools behave
- No intuition

- Consider cases where \#observations \gg e^D
- Often \#observations >> e^D
Linear tools

- Principal component analysis (PCA):
  - based on covariance matrix
  - huge (DIM x DIM)
  - poorly estimated with finite number of data

- Other methods:
  - Linear discriminant analysis (LDA)
  - Partial least squares (PLS)
  - ...

Similar problems!
Nonlinear tools

\[ y = f(x_1, x_2, \ldots, x_d, \theta) \]

- If \( d \rightarrow \), size(\( \theta \) \( \rightarrow \))
- \( \theta \) results from the minimization of a non-convex cost function
  - local minima
  - numerical problems (flats, high slopes)
  - convergence
  - etc
- Ex: Multi-layer perceptrons, Gaussian mixtures (RBF), self-organizing maps, etc.
Curse of dimensionality

Motivation

- Model OK
- Model KO

Graphs showing the difference between high and low dimensionality conditions.
Curse of dimensionality

- Number of points on a grid increases exponentially with DIM

- In high DIM:
  - never enough data
  - never sure to \textit{interpolate}
Curse of dimensionality

- Example: Silverman (1986)
  Number of Gaussian kernels necessary to approximate a (Gaussian) distribution in DIM

Motivation
Empty space phenomenon

• “Statistical” view of the same problem:

To estimate parameters

   ex: covariance matrix of Gaussian distribution in DIM

   one needs (too...) many data

• Example: histograms in DIM

   - compromise between accuracy and
     # of data in each bin
Surprising facts: sphere

- Volume of sphere of constant radius (=1) in dimension DIM

- \( DIM=1 \)
  - 1
  - \( \text{vol} = 2 \)

- \( DIM=2 \)
  - \( \text{vol} = \pi \)

Motivation

\[
V(d) = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} r^d
\]
- Ratio volume sphere / cube

in HD, all points are here and not here
Surprising facts: spheres

- Volume ratio of embedded spheres

![Graph showing volume ratio of embedded spheres](image)
Motivation

Surprising facts: Gaussians

- Multi-DIM Gaussian distributions

- % points inside a sphere of radius 1.65
Surprising facts: Gaussians

- Another view of high-DIM Gaussian distributions:
  - Probability to find a point at distance $r$ from the center of a DIM-dimensional multinormal distribution
Concentration of the Euclidean norm

- Distribution of the norm of random vectors
  - i.i.d. components in [0,1]
  - norms in $[0, \sqrt{d}]$ as

- Norms concentrate around their expectation
- They don’t discriminate anymore!

Motivation
Motivation

Distances also concentrate

Pairwise distances seem nearly equal for all points

Relative contrast vanishes as the dimension increases

If

$$\lim_{d \to \infty} \frac{\sqrt{\text{Var}(\|X\|_2)}}{\text{E}(\|X\|_2)} = 0$$

then

$$\frac{D_{\text{MAX}} - D_{\text{MIN}}}{D_{\text{MIN}} d} \to_p 0$$

when

$$d \to \infty$$

[Beyer]
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
Reducing (the curse of) dimensionality

Feature selection

\[
\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{bmatrix} \quad \xrightarrow{\text{dimension reduction}} \quad \begin{bmatrix} x'_1 \\ x'_2 \\ x'_p \end{bmatrix}
\]

\( d > p \)
Why reducing the dimensionality?

- Theoretically not useful:
  - More information means easier task
  - Models can ignore irrelevant features
    (e.g. set weights to zero)

  « In theory, practice and theory are the same. But in practice, they're not »

- Lot of inputs means ...
  - Lots of parameters & Large input space

⇒ Curse of dimensionality and risks of overfitting!
Why feature selection?

- Feature selection is often as important as the model itself!
- Industrial example

Road Temperature
Air Temperature
Wind Direction
Dew Temperature

Presence of ice on the road

which feature (input, variable, attribute, predictor, ...) should we feed our model with??
Why feature selection?

- To visualize data

\[
x = \begin{pmatrix} x_1 \\ \vdots \\ x_N \end{pmatrix} \quad \x' = \begin{pmatrix} x'_1 \\ x'_2 \end{pmatrix}
\]

- To get insight about the causes of the problem

If air temperature is selected and road temperature is not

Then water temperature is closer to air temperature than to road temperature.

- Reduce data collection time/cost, computation time, etc.
### Feature selection reduces dimensionality

#### Unsupervised

<table>
<thead>
<tr>
<th>Selection</th>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation between inputs</td>
<td>Mutual information between inputs</td>
</tr>
<tr>
<td>Projection</td>
<td>Principal Component Analysis</td>
<td>Sammon’s Mapping, Kohonen maps</td>
</tr>
</tbody>
</table>

#### Supervised

<table>
<thead>
<tr>
<th>Selection</th>
<th>Linear</th>
<th>Nonlinear</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Correlation between inputs and output</td>
<td>Mutual information between inputs and outputs, Greedy algorithms, Genetic algorithms</td>
</tr>
<tr>
<td>Projection</td>
<td>Linear Discriminant Analysis, Partial Least Squares</td>
<td>Projection pursuit</td>
</tr>
</tbody>
</table>
Supervised selection: filter versus wrapper

- Supervising does not necessarily mean to use the model!

**FILTER**

\[
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{bmatrix}
\xrightarrow{\text{selection}}

\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_M
\end{bmatrix}
\xrightarrow{\text{relevance criterion}}

Not the final quality criterion!

**WRAPPER**

\[
\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_N
\end{bmatrix}
\xrightarrow{\text{selection}}

\begin{bmatrix}
    x_1 \\
    x_2 \\
    \vdots \\
    x_M
\end{bmatrix}
\xrightarrow{(\text{non})\text{linear model}}

Many (non)linear models to design!
The ingredients of feature selection

- **Key Element 1**: *Subset relevance assessment*
  - Among all $2^{d-1}$ possible subsets, which is the best one?

- **Key Element 2**: *Optimal Subset search*
  - How not to consider all $2^{d-1}$ possible subsets?

- **Filters**
  - Correlation
  - Mutual information

- **Wrappers**
  - Greedy search
  - Genetic algorithms
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
Is $x_1$ relevant to predict $y$? What about $x_2$?

Relevance is difficult to define

Filter approach (model free): $P(y \mid x_i) \neq P(y)$
- a variable (or set of ) is relevant if it is statistically dependent on $y$

Wrapper approach (uses model $f$): $\min_{f} (y - f(x_i))^2 \approx 0$
- a variable (or set of ) is relevant if the model built on it shows good performances
Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series
Correlation, a linear filter

- **Definition**: correlation between random variable \( X \) and random variable \( Y \) \((E[.]\) is the expectation operator\):

\[
\rho_{xy} = \frac{E[(x - E[x]) \cdot (y - E[y])]}{\sqrt{E[(x - E[x])^2] \cdot E[(y - E[y])^2]}}
\]

- **Estimation**: when one has a dataset \( \{x_j, y_j\} \) (\( \overline{x} \) means the average of \( x_i \))

\[
r = \frac{\sum_{j=1}^{N} ((x_j - \overline{x}) \cdot (y_j - \overline{y}))}{\sqrt{\sum_{j=1}^{N} (x_j - \overline{x})^2} \cdot \sqrt{\sum_{j=1}^{N} (y_j - \overline{y})^2}}
\]

- **Measures linear dependencies**
  - Always comprised between -1 and +1
  - 0 indicates decorrelation (no linear relation)
Correlation, a linear filter

- Examples

**Strong correlation**

- $r^2 = 1.000$
- $r^2 = 0.991$
- $r^2 = 0.904$
- $r^2 = 0.821$
- $r^2 = 0.493$
- $r^2 = 0.0526$

**Weak correlation**
Correlation does not measure nonlinear relations

- Low correlation does not mean absence of relationship

Feature selection $\rightarrow$ Subset relevance assessment $\rightarrow$ Correlation

$\rho_{x^2} \approx 0$
Correlation does not mean causality

- High correlation does not mean causality
  - Number of murders in a city highly correlated (0.80) with number of churches
  - Simply because both murders and number of churches increase with population density
Limitations of correlation

- Correlation
  - is linear
  - is parametric (it makes the hypothesis of a ...linear model)
  - does not *explain*
  - is almost impossible to define between more than 2 variables
  - is sensitive to outliers ($R^2 = 1 - \text{NMSE}$)
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
Mutual information

- Relevance of a subset $X_S$: mutual information $I(X_S; y)$ between this subset and the target variable $Y$

- What is the mutual information?

- Mutual information between random variable $x$ and random variable $y$ measures how the uncertainty on $y$ is reduced when $x$ is known. (and vice versa)

- Let's begin by the entropy...
Entropy

- The entropy of a random variable is a measure on its uncertainty

\[ H(y) = -\mathbb{E}[\log(P[y])] \]
\[ = -\sum_{y \in \Omega} \log(P[y])P[y] \quad \text{when } Y \text{ is discrete} \]
\[ = -\int \log(P[y])dy \quad \text{when } Y \text{ is continuous} \]

- Can be interpreted as the average number of bits needed to describe \( y \)
Example of entropy

- Entropy of a discrete variable $y$ taking values in $\Omega = \{0,1\}$, with $P[y=1] = p$ and $P[Y=0] = 1-p$

$$H(y) = - \sum_{y \in \Omega} P[y] \log(P[y])$$

- Uncertainty is maximal when both events have the same probability

Feature selection $\rightarrow$ Subset relevance assessment $\rightarrow$ Mutual information
Conditional entropy

- Conditional entropy $H(y \mid x)$ measures the uncertainty on $y$ when $x$ is known.

$$H(y \mid x) = H(y, x) - H(x)$$

- If $Y$ and $X$ are independent,

$$H(y \mid x) = H(y)$$

the uncertainty on $Y$ is the same if we know $X$ as if we don’t!
Mutual information

- Mutual information between $x$ and $y$

$$I(y; x) = H(y) - H(y \mid x) = H(x) - H(x \mid y)$$

- Difference between entropy of $y$ and entropy of $y$ when $x$ is known

- Some properties:
  - If $x$ and $y$ are independent, $I(y; x) = 0$
  - $I(y; y) = H(y)$
  - $I(y; x)$ is always non negative and less than $\min(H(y), H(x))$
Nonlinear dependencies with MI

- Mutual information identifies nonlinear relationships between variables

- Examples:
  - $x$ uniformly distributed over $[-1, 1]$
  - $y = x^2 + \text{noise}$
  - $z$ uniformly distributed over $[-1, 1]$
  - $z$ and $x$ are independent

- Results:

<table>
<thead>
<tr>
<th>1000 samples</th>
<th>$y, y$</th>
<th>$x, y$</th>
<th>$z, y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>1</td>
<td>0.0460</td>
<td>0.0522</td>
</tr>
<tr>
<td>Mutual info.</td>
<td>2.2582</td>
<td>1.1996</td>
<td>0.0030</td>
</tr>
</tbody>
</table>
High-dimensional mutual information

• What about the relevance of a set of features?
• Reminder:

\[ I(x, y) = H(y) - H(y | x) = H(x) - H(x | y) \]

• \( x \) and \( y \) may be vectors!
• If \( X \) is a subset of features, its relevance may still be evaluated
• Evaluating subsets is the right issue!

• The difficulty is in the estimation of \( I(y;x) \):
  - histograms, kernels and splines suffer from the curse of dimensionality!
  - k-NN based estimators are the (almost only) solution
Estimators for mutual information

• All estimators suffer from the curse of dimensionality!

• Histograms, kernels (Parzen windows, etc.) are the worst...

• $k$-NN based estimators (Kraskov) are more robust
Relevance and redundancy

- In practice: many methods based on the estimation of the mutual information on a limited subset of features (2, 3, few...)

- Relevance: $I(x_i, y)$

- Redundancy: $I(x_i, x_j)$

- Principle when dealing with bivariate mutual information only: Max. relevance and min. redundancy together!
  - this is a multi-objective criterion
  - depends on the weighting between the two criteria

- ∃ extensions to tri-variate mutual information (interaction measure), etc.
Stopping criterion

- In theory, the mutual information (of a set) can only increase when adding variables.

- When to stop then?
  - Often: when the estimation of the MI decreases (bad idea!)
  - Better: evaluate the statistical relevance (hypothesis test) of the addition of a new variable (ex: permutation test)
  - For external reasons: fix the max. number of variables
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
Wrappers

Many (non)linear models to design!

- Just build the models, and evaluate them...
- Problems come when the models are "computationally intensive"
  - need to train each of them
  - need to set the hyper-parameters (cross-validation, etc.)
  - need to evaluate the performances (double cross-validation, etc.)
Wrappers versus filters

<table>
<thead>
<tr>
<th>FILTERS</th>
<th>WRAPPERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ <strong>Fast</strong> : build only one model</td>
<td>+ Relevance criterion <strong>easy to estimate</strong></td>
</tr>
<tr>
<td>+ <strong>Intuitive</strong> : identifies statistical dependency</td>
<td>+ <strong>Model-aware</strong> : identifies optimal subset to build optimal model</td>
</tr>
<tr>
<td>- Relevance criterion <strong>hard to estimate</strong></td>
<td>- <strong>Slow</strong> : must build lots of models</td>
</tr>
<tr>
<td>- <strong>Model-ignorant</strong> : most relevant subset might not be optimal for subsequent model</td>
<td>- <strong>Not intuitive</strong> : features for best model might not actually be most explanatory variables</td>
</tr>
</tbody>
</table>
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
Optimal subset search

- **In theory:** just try $2^d$ subsets and evaluate them...
  - Evaluation: mutual information (filters), or model itself (wrapper)
  - just imagine for $d=200$ 😊

- **In practice:** greedy procedure
  - Define an initial subset
  - Choose a strategy to update subset
  - Decide when to stop
Greedy subset search

- Example with 4 features

- Makes hypothesis that best subset can be constructed iteratively
Forward search

- Define an initial subset
  - begin with empty set

- Choose a strategy to update subset
  - Filter: add feature that increases the most the relevance/redundancy compromise
  - Wrapper: add feature that increases the most the performances of the model

- Decide when to stop
  - Filter: needs a supplementary criterion
  - Wrapper: stop when adding a feature increases the generalization error
Backward search

- Define an initial subset
  - begin with the full set

- Choose a strategy to update subset
  - Filter: remove feature that increases most the relevance/redundancy compromise
  - Wrapper: remove feature that increases the most the performances of the model

- Decide when to stop
  - Filter: needs a supplementary criterion
  - Wrapper: stop when removing a feature increases the generalization error
Variants

- **Forward-backward**
  - At each step, consider all additions and removals of one variable, and select the best result
  - Wrapper: this makes sense
  - Filter: in theory, the mutual information cannot increase with less variables
Genetic algorithms

- “clever” random exploration of the space of subsets
  1. Draw initial population (candidate subsets)
  2. Select individuals
  3. Apply cross-overs and mutations on individuals
  4. Repeat from 2 until a new population is generated
  5. Select best individuals and repeat from 1

Feature selection → Greedy search methods
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
Embedded methods

- Idea: to build the model and restrict the number of features used by the model together

- Appealing idea! Optimizing the whole can only be better than separating into 2 parts (feature selection and model)

- Example: LASSO
  - Linear model $y = W.X$
  - Criterion: $\min_w \frac{1}{N} \sum_{j=1}^{N} (y^j - Wx^j)^2 + \mu \|W\|_1$
    - Training error
    - Regularization
  - Usually (ridge regression): regularization term is a 2-norm
  - Here a 1-norm (this reduces the effective number of features used by the model, both in theory and in practice)

- Many interesting directions for embedded methods!
Outline

- **Motivation**: High-dimensional data, concentration of the norm
- **Feature selection**
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- **Examples**
  - Housing
  - Business plans
  - Tecator
  - Time series
Illustrative examples

- Housing (a traditional benchmark)
  - to show that it can work

- Business plan classification
  - to understand

- IR Spectra
  - to analyze data

- Time series
  - how to choose the regressor?
Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series
Housing

- Concerns housing values in suburbs of Boston
- Attributes:
  1. CRIM  per capita crime rate by town
  2. ZN    proportion of residential land zoned for lots over 25,000 sq.ft.
  3. INDUS proportion of non-retail business acres per town
  4. CHAS  Charles River dummy variable (= 1 if tract bounds river, 0 otherw.)
  5. NOX   nitric oxides concentration (parts per 10 million)
  6. RM    average number of rooms per dwelling
  7. AGE   proportion of owner-occupied units built prior to 1940
  8. DIS   weighted distances to five Boston employment centres
  9. RAD   index of accessibility to radial highways
  10. TAX  full-value property-tax rate per $10,000
  11. PTRATIO pupil-teacher ratio by town
  12. B    $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
  13. LSTAT% lower status of the population
  14. MEDV Median value of owner-occupied homes in $1000's
- Forward selection by mutual information between output and set of attributes
Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series

Examples → Business plans
Business plan classification

- Context: a competition of business plans evaluated by experts
- The question: is the success of a company in relation with the scores given by the experts?
- 161 business plans
- 7 criteria:
  - interest for investor
  - content of the business plan
  - usefulness for clients
  - differentiation with other products
  - size of market
  - competitors
  - global rating
- To predict: success of company after two years
Business plan classification

• Important variables are evaluated by:
  – correlation
  – difference of means
  – mutual information

• Results

<table>
<thead>
<tr>
<th></th>
<th>high</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>correlation</td>
<td>3,4</td>
<td>7,1,2,5,6,7</td>
</tr>
<tr>
<td>difference of means</td>
<td>3,4,7</td>
<td>1,2,5,6</td>
</tr>
<tr>
<td>mutual information</td>
<td>3,4,5</td>
<td>7,1,2,6,7</td>
</tr>
</tbody>
</table>
Outline

- Motivation: High-dimensional data, concentration of the norm
- Feature selection
  - Subset relevance assessment
    - Correlation
    - Mutual information
    - Wrappers
  - Greedy search methods
  - Embedded methods
- Examples
  - Housing
  - Business plans
  - Tecator
  - Time series
Infrared spectra (Tecator)

- Meat spectra (Tecator)
- Prediction of fat content
- 100 wavelengths
- 172 training samples, 43 test samples
Infrared spectra (Tecator)

- Original methodology (as an example, not a rule...)
- 100 wavelengths (sometimes 1000, 2000...) -> very high-dimensional data
- estimation of MI becomes hard
- Two solutions:
  - ranking: $I(x_i;i)$
    does not take relations between $X_i$ into account
  - forward selection: $I((x_i, x_j, x_k, ...);y)$
    curse of dimensionality in the estimation
- Take both!
  - set of features selected by forward (until decrease)
  - set of features selected by ranking (until the union of both sets reaches $N$ elements)
  - try the $2^N$ IM evaluations (filter) or the $2^N$ models (wrapper)

Examples $\rightarrow$ Tecator
Infrared spectra (Tecator)

- Only 7 variables are selected
- In two ranges
- Probably information about first and second derivatives
Infrared spectra (Tecator)

<table>
<thead>
<tr>
<th>model</th>
<th>variables</th>
<th>NMSEt</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>1.64 e-2</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1.36 e-2</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>4.6 e-3</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>1.78 e-2</td>
</tr>
<tr>
<td>5</td>
<td>42</td>
<td>7.8 e-2</td>
</tr>
<tr>
<td>6</td>
<td>42</td>
<td>6.08 e-2</td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>8.27 e-3</td>
</tr>
<tr>
<td>8</td>
<td>20</td>
<td>5.2 e-3</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>1.35 e-2</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>1.12 e-1</td>
</tr>
<tr>
<td>11</td>
<td>7</td>
<td>6.60 e-3</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>2.70 e-3</td>
</tr>
<tr>
<td>13</td>
<td>7</td>
<td>2.93 e-2</td>
</tr>
</tbody>
</table>
Outline

• Motivation: High-dimensional data, concentration of the norm
• Feature selection
  – Subset relevance assessment
    • Correlation
    • Mutual information
    • Wrappers
  – Greedy search methods
  – Embedded methods
• Examples
  – Housing
  – Business plans
  – Tecator
  – Time series
• Model: \( x(t + 1) = f(x(t), x(t - 1), x(t - 2), x(t - 3), x(t - 4), \ldots) \)
Time series

- Three questions
  - Sampling frequency
    \[ x(t + 1) = f(x(t), x(t - 1), x(t - 2), x(t - 3), x(t - 4), ...) \]
    or \[ x(t + 1) = f(x(t), x(t - 1), x(t - 4), x(t - 7), x(t - 10), ...) \]
  - Size of regressor
    \[ x(t + 1) = f(x(t), x(t - 1), x(t - 2), x(t - 3), x(t - 4), ...) \]
    or \[ x(t + 1) = f(x(t), x(t - 1), x(t - 2), x(t - 3)) \]
  - Non-contiguous values
    \[ x(t + 1) = f(x(t), x(t - 1), x(t - 2), x(t - 3), x(t - 4), ...) \]
    or \[ x(t + 1) = f(x(t), x(t - 1), x(t - 7), x(t - 8), x(t - 14), ...) \]
A digression about the size of regressors

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

Examples → Time series
A digression about the size of regressors

- Forecasting: Taken’s 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!

- In practice
  - Take many input variables
  - Select 2q+1 by feature selection
  - If needed, project them on q new variables

Examples → Time series
Selection / projection

• Selection
  (choosing among the original features which ones to keep)
    + easy
    + interpretability of the features

• Projection
  (creating new features from the original ones)
    + more general → possibly more efficient
    - more difficult
    - features not interpretable

  The book:
  • Nonlinear Dimensionality Reduction
    Springer, Series: Information Science and Statistics
    John A. Lee, Michel Verleyesen, 2007
    300 pp., ISBN: 978-0-387-39350-6
Conclusions

- Feature selection = two ingredients
  - Subset evaluation criterion
  - Greedy search in the space of subsets

- Mutual information is a good “filter” criterion
  - But difficult to evaluate
  - Many options to make a compromise between relevance and redundancy

- Wrapper approach is better (for prediction), but computationally (too) intensive
Some good questions (and not so good answers…)

• Is most relevant feature necessarily optimal for prediction?
  – No for example if the model is not powerful enough

• Is most useful feature for prediction necessarily relevant?
  – No for example bias terms are irrelevant but can help for prediction

• Are two redundant features really useless together?
  – Perfectly redundant features are useless
  – Highly correlated features can be useful if the model is able to use the ‘information’ contained in their differences

• Can two features be useless alone and useful together?
  – Yes, think of the XOR problem