Information theoretic feature selection for non-standard data

Michel Verleysen

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

Thanks to

- PhD and post-doc and other colleagues (in and out UCL), in particular

Catherine Krier

Damien François

Amaury Lendasse

Gauthier Doquire

Fabrice Rossi

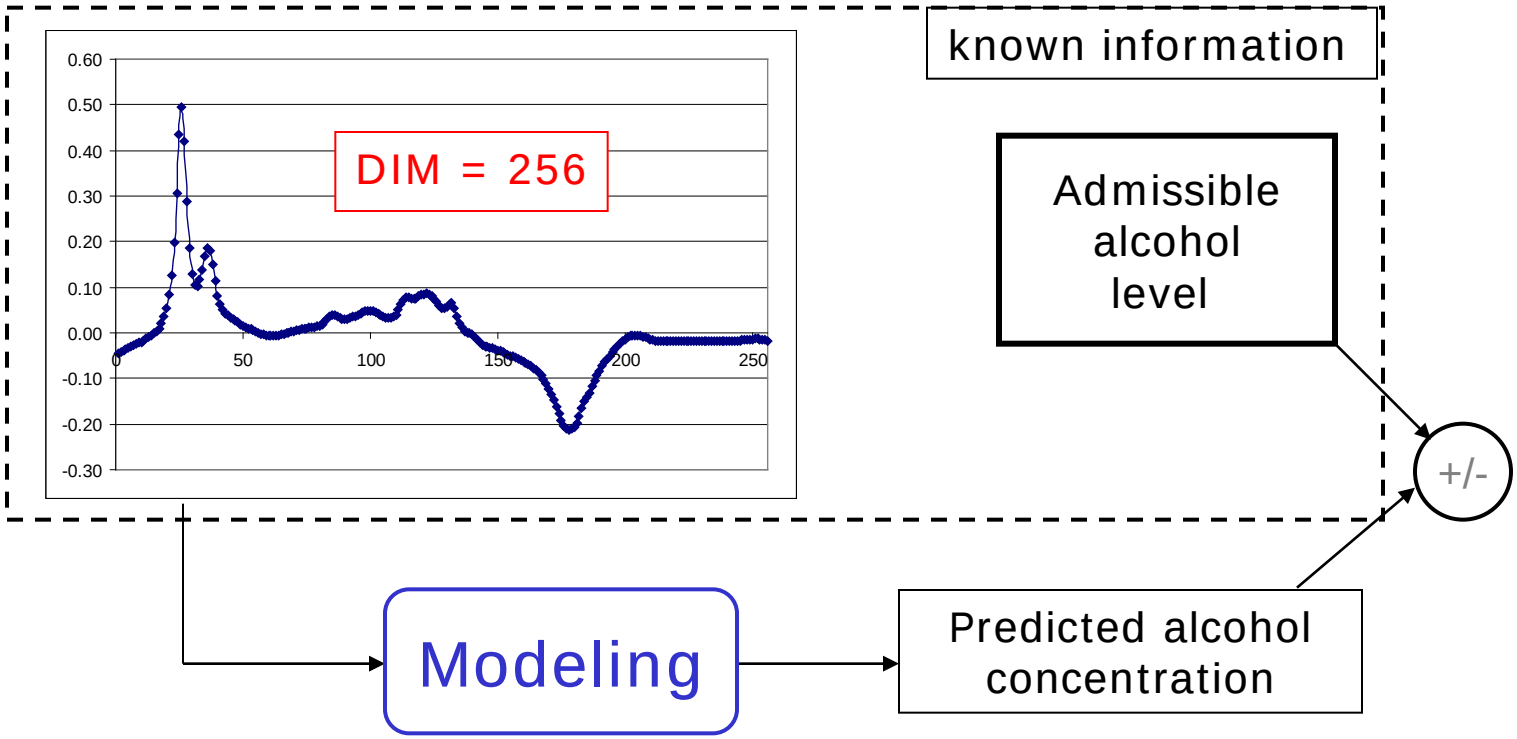
Frederico Coelho

Outline

- Motivation
- Feature selection in a nutshell
- Relevance criterion
- Mutual information
- Structured data
- Case studies
  - MI with missing data
  - MI with mixed data
  - MI for multi-label data
  - semi-supervised feature selection

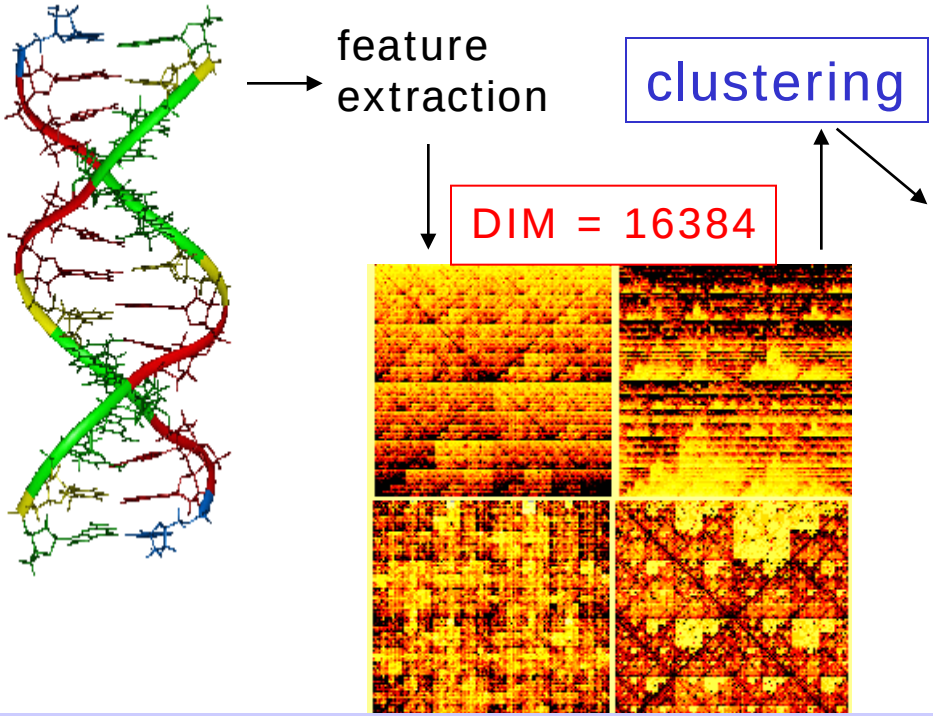
HD data are everywhere

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



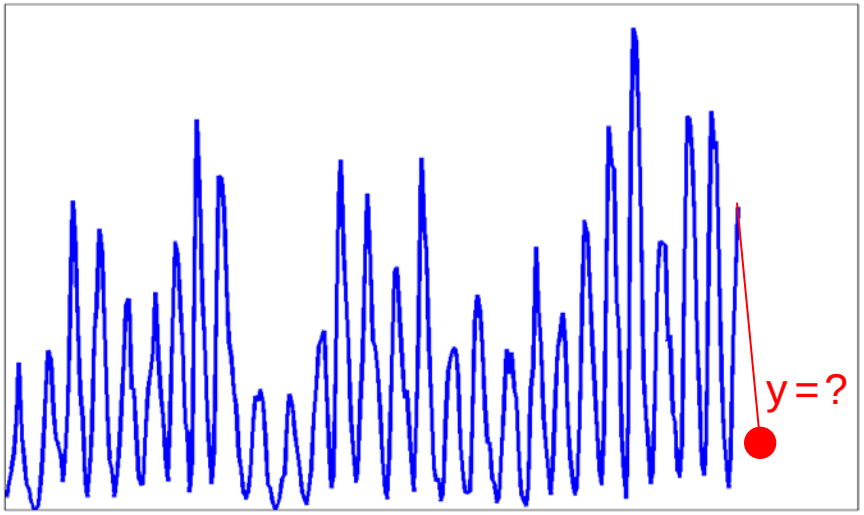
HD data are everywhere

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



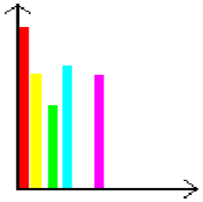
HD data are everywhere

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



$$y = f\left(x_{t-\mathrm{DIM}+1}, \ldots x_{t-1}, x_t\right)$$

Generic data analysis



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

When	1
Times	1
Trouble	1
Let	65
wisdom	1

number of
Variables or features

Data
matrix

number of
observations

Analysis

Models

The big challenge

- What is the problem with many features ?
  - Computational complexity ?

The big challenge

- What is the problem with many features ?
  - Computational complexity ? Not really

The big challenge

- What is the problem with many features ?
  - Computational complexity ? Not really
  - Models stuck in local minima?

The big challenge

- What is the problem with many features ?
  - Computational complexity ? Not really
  - Models stuck in local minima? Not so much

The big challenge

- What is the problem with many features ?
  - Computational complexity ? Not really
  - Models stuck in local minima? Not so much
  - Concentration of distances ?

The big challenge

- What is the problem with many features ?
  - Computational complexity ? Not really
  - Models stuck in local minima? Not so much
  - Concentration of distances ? Yes

The big challenge

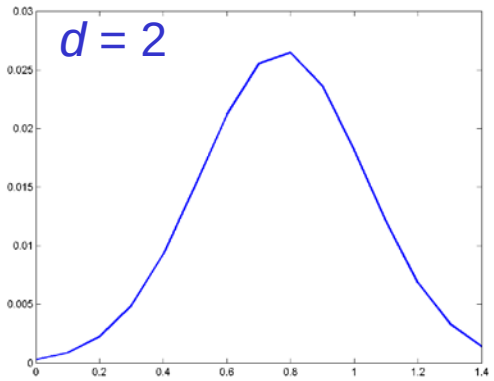
- What is the problem with many features ?
  - Computational complexity ? Not really
  - Models stuck in local minima? Not so much
  - Concentration of distances ? Yes
  - Poor estimations ?

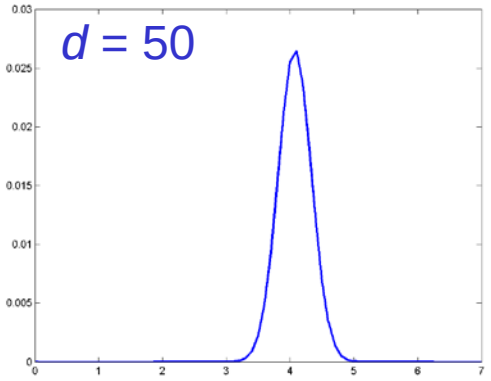
The big challenge

- What is the problem with many features ?
  - Computational complexity ? Not really
  - Models stuck in local minima? Not so much
  - Concentration of distances ? Yes
  - Poor estimations ? Yes

Concentration of the Euclidean norm

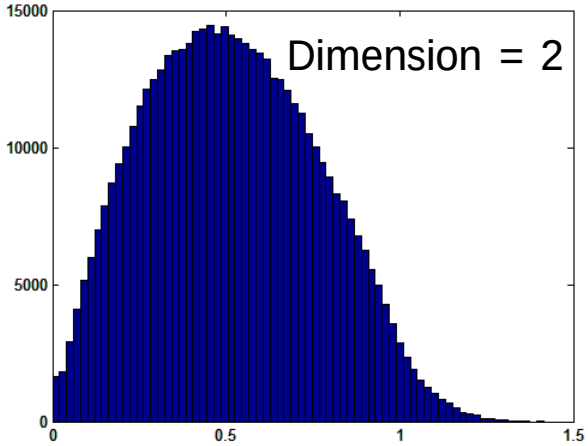
- Distribution of the norm of random vectors
  - i.i.d. components in [0,1]
  - norms in $[0, \sqrt{d}]$ as

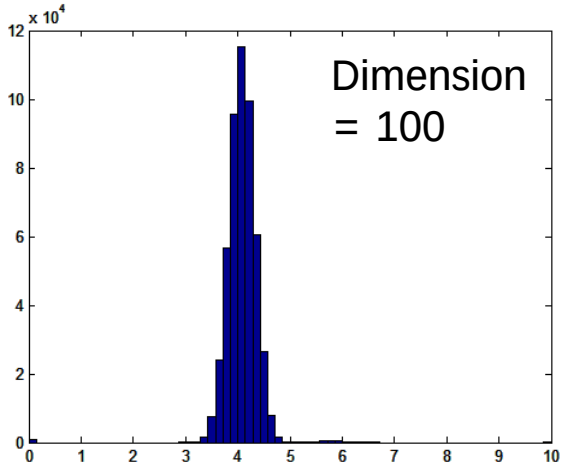




- Norms concentrate around their expectation
- They don't discriminate anymore !

Distances also concentrate





Pairwise distances seem nearly equal for all points

Relative contrast vanishes as the dimension increases

If $\lim_{d\to\infty} \frac{\sqrt{\mathrm{Var}\left(\|X\|_2\right)}}{\mathrm{E}\left(\|X\|_2\right)} = 0$ then $\frac{DMAX_d - DMIN_d}{DMIN_d} \to_p 0$

when $d \to \infty$

[Beyer]

The estimation problem

- An example of linear method: Principal component analysis (PCA)

  Based on covariance matrix
  - huge (DIM x DIM)
  - poorly estimated with low/finite number of data

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

  Similar problems!

Nonlinear tools

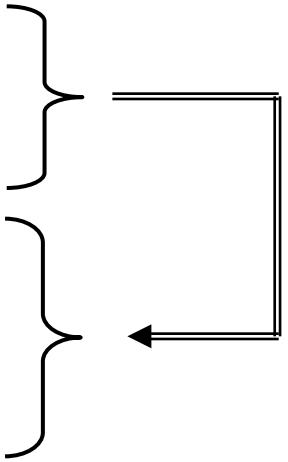
Nonlinear models

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

- If d ↗↗, size(θ) ↗↗

  θ 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), kernel machines, self-organizing maps, etc.

Why reducing the dimensionality ?

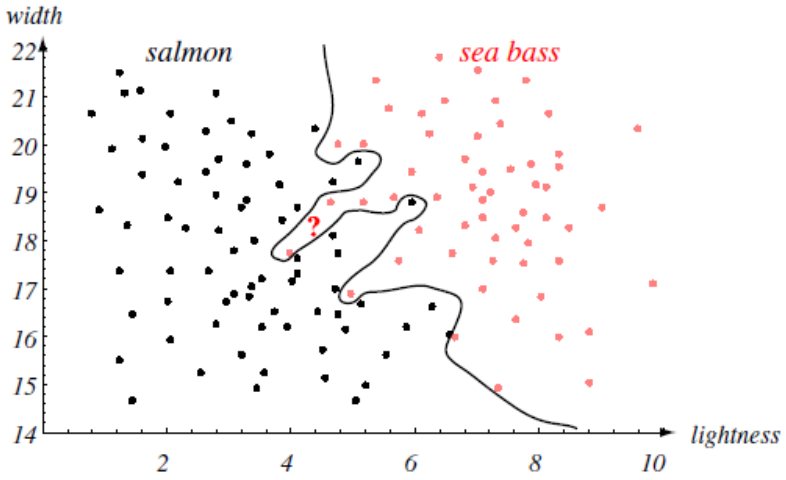
- Not useful in theory:
  - More information means easier task
  - Models can ignore irrelevant features
    (e.g. set weights to zero)

- But…
  - Lot of inputs means …
    Lots of parameters & Large input space

- Curse of dimensionality and risks of overfitting !

Overfitting

Model-dependent

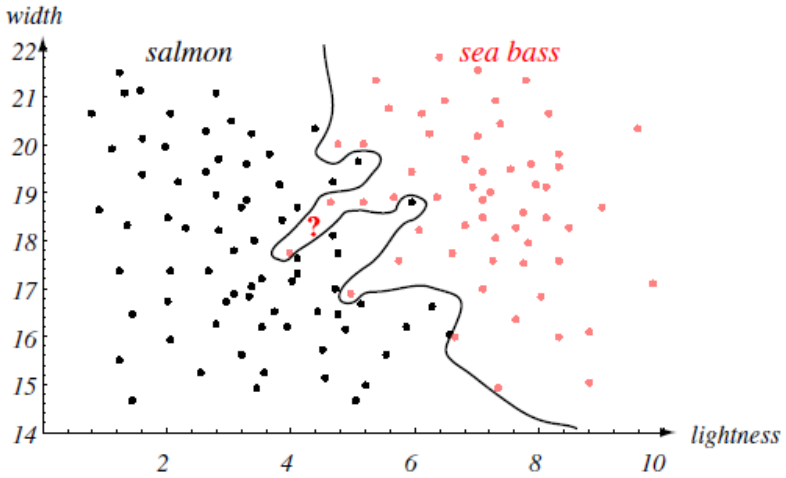


- Use regularization

From: Duda et al., Pattern Classification, 2nd ed., Wiley, 2001

Overfitting

Model-dependent



- Use regularization

From: Duda et al., Pattern Classification, 2nd ed., Wiley, 2001

Data-dependent

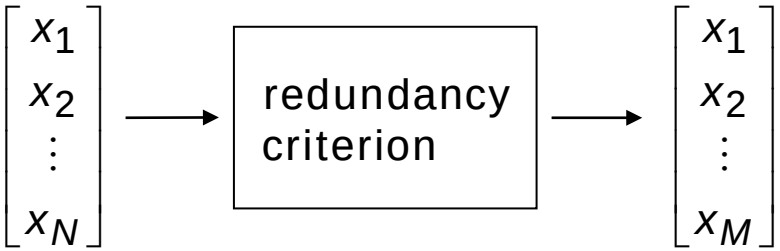
- *D* points to fit the simplest (linear) model in a *D*-dim space
- (perfect) fitting → approximation: much more than *D* points!
- What is much less than *D* points are available?

Outline

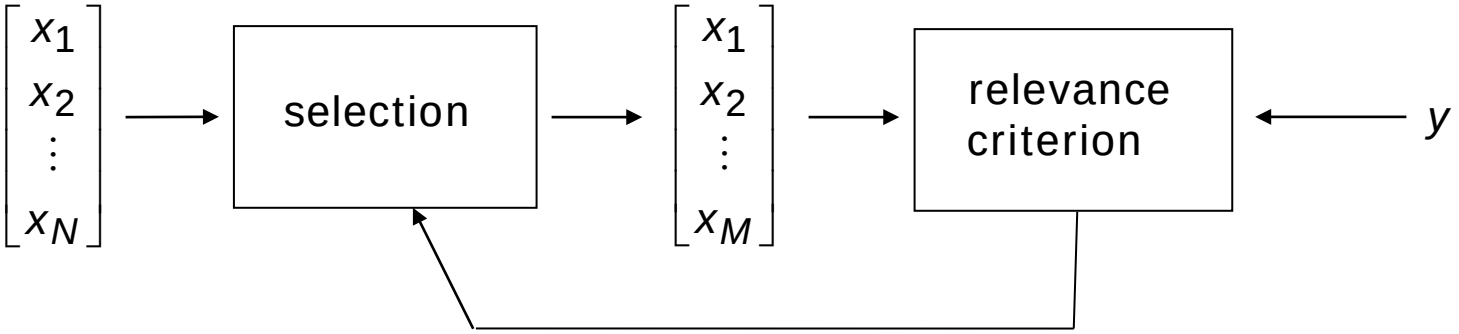
- Motivation
- Feature selection in a nutshell
- Relevance criterion
- Mutual information
- Structured data
- Case studies
  - MI with missing data
  - MI with mixed data
  - MI for multi-label data
  - semi-supervised feature selection

Feature selection in a nutshell

- 1001 ways (and more…) to perform feature selection
- The challenges:
  - Unsupervised

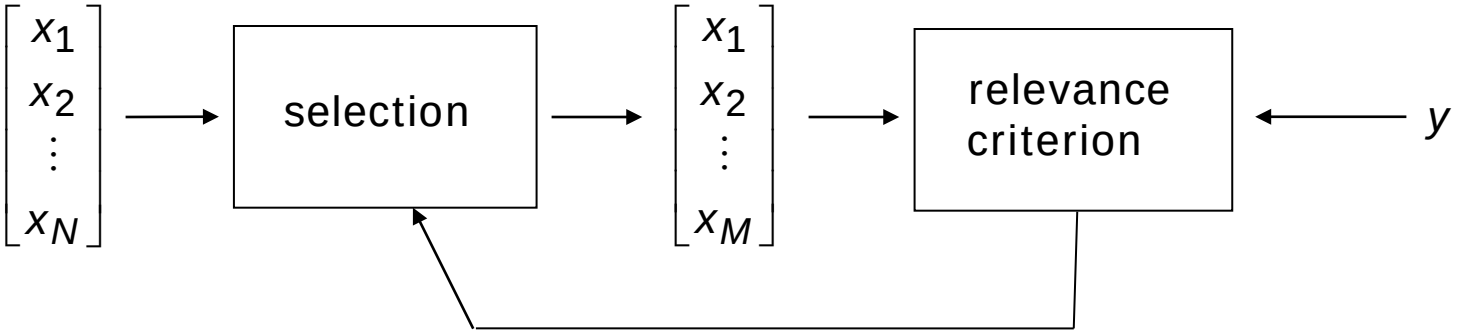


Supervised

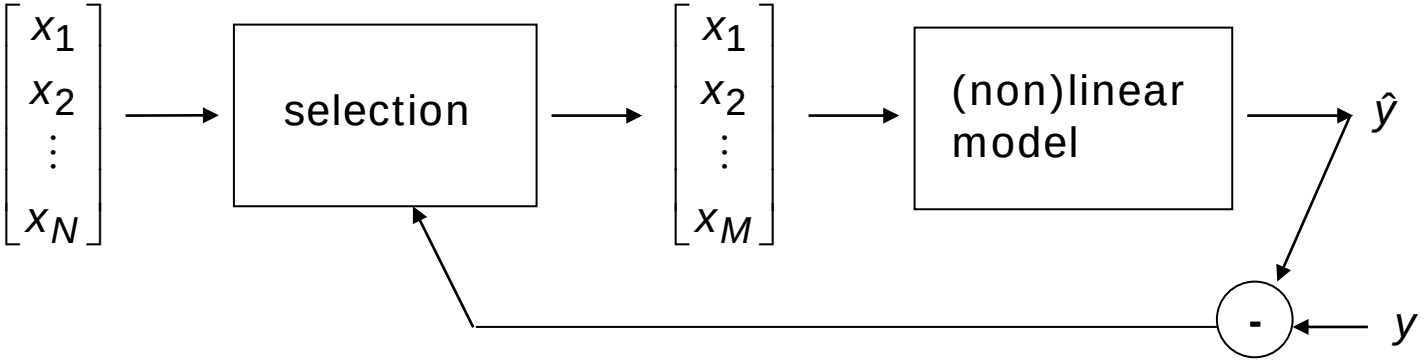


Feature selection in a nutshell

- 1001 ways (and more…) to perform feature selection
- The challenges:
  - Unsupervised – supervised
  - Filter

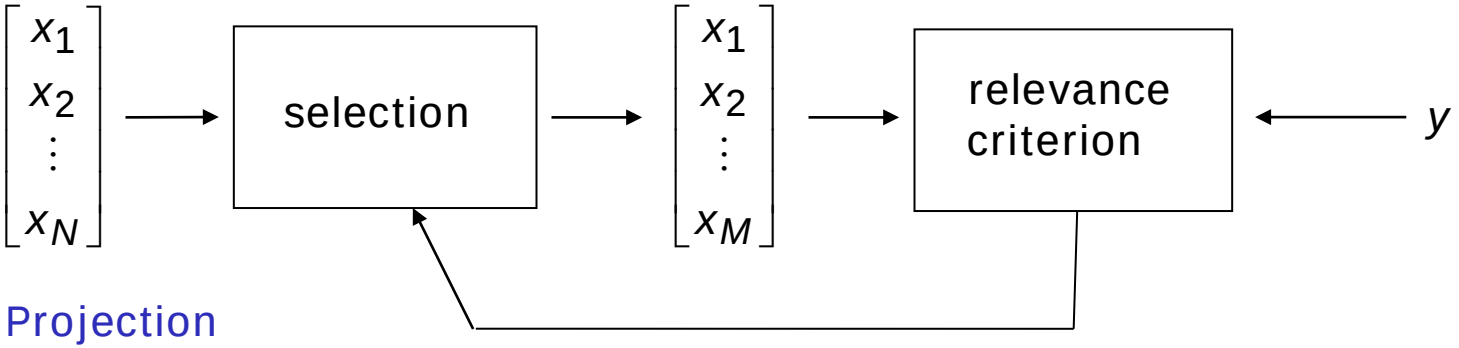


Wrapper

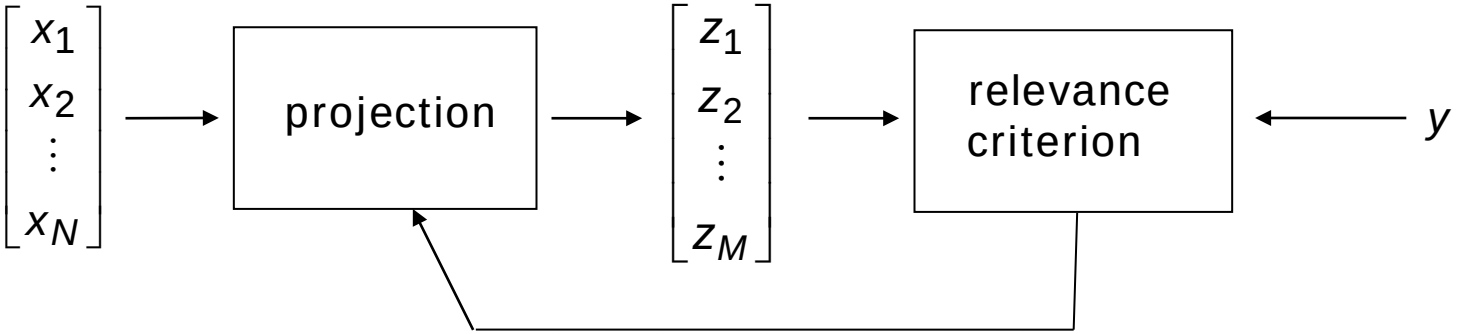


Feature selection in a nutshell

- 1001 ways (and more…) to perform feature selection
- The challenges:
  - Unsupervised – supervised
  - Filter – wrapper
  - Selection



Projection



Feature selection in a nutshell

- 1001 ways (and more…) to perform feature selection
- The challenges:
  - Unsupervised – supervised
  - Filter – wrapper
  - Selection – Projection
  - Linear
    - Straightworward, easy
    - No tuning parameter
    - No estimation problem
    - But obviously doesn't capture nonlinear relationships…

  Nonlinear
    - Less intuitive (interpretability)
    - Less straightworward (bounds,…)
    - Estimation difficulties

Feature selection in a nutshell

- 1001 ways (and more…) to perform feature selection
- The challenges:
  - Unsupervised – supervised
  - Filter – wrapper
  - Selection – Projection
  - Linear – nonlinear
  - Greedy approach

  For D features, there exist 2^D-1 possible feature subsets
  - Not possible to test all of them -> greedy approaches
  - Start with 1 feature, then add (forward search)
  - Start with all, then remove (backward search)
  - Start with 1 feature, then add, but possibility to remove (forward-backward)
  - Genetic algorithms
  - …

Feature selection in a nutshell

- 1001 ways (and more…) to perform feature selection
- The challenges:

Choice	# of possibilities
Unsupervised - supervised	2
Filter – wrapper	2
Selection - projection	2
Linear – non-linear criterion	> 5
Greedy approach	> 5

- At least 150 very good feature selection methods!

One choice among others...

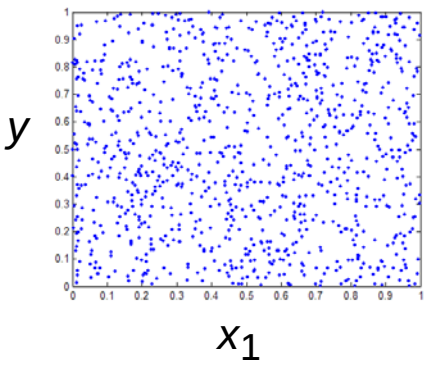
- Selection
  - To keep the interpretability of features
- Supervised
  - to take as much information as possible into account
- Nonlinear
  - To tackle a large class of problems
- Filter
  - To avoid the computational burden of models
- Greedy approach
  - Ad-hoc, according to problem and computational constraints

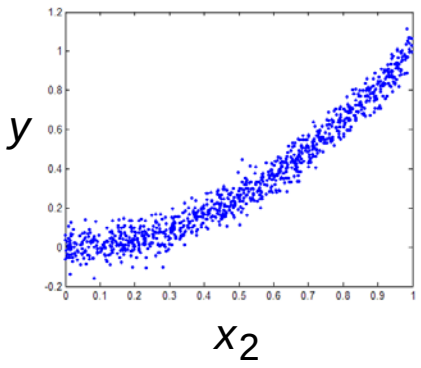
Outline

- Motivation
- Feature selection in a nutshell
- Relevance criterion
- Mutual information
- Structured data
- Case studies
  - MI with missing data
  - MI with mixed data
  - MI for multi-label data
  - semi-supervised feature selection

Relevance criterion

- Is x_1 relevant to predict y? What about x_2?



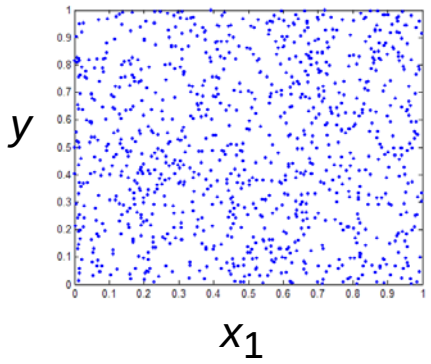


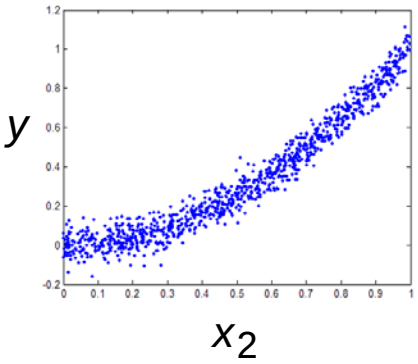
- Relevance
  - easy intuitive concept
  - difficult to define

Relevance criterion

- Nonparametric approach
  - Model free
  - “filter”
  - a variable (or set of) is relevant if it is statistically dependent on *y*

$$P(y \mid x_i) \neq P(y)$$





  - But needs probability density estimations

- Parametric approach
  - Uses prediction model *f*
  - “wrapper”
  - a variable (or set of) is relevant if the model built on it shows good performances

$$\min_f (y - f(x_i))^2 \approx 0$$

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)

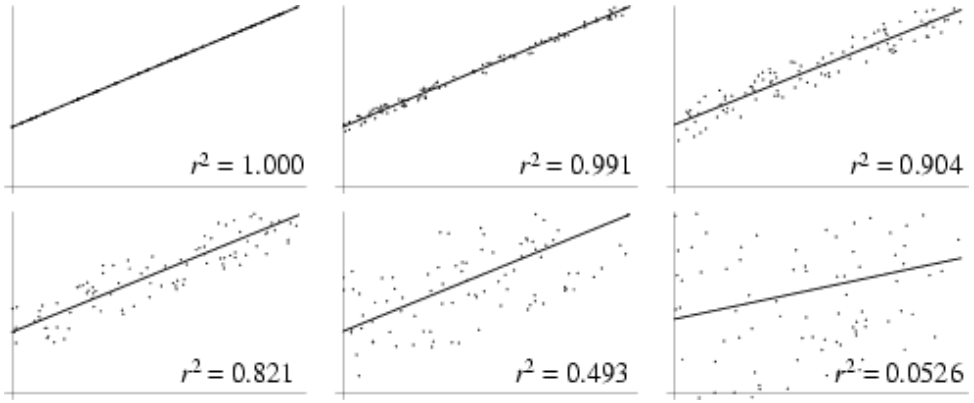
$$r = \frac{\sum_{j=1}^{N} \left((x^j - \overline{x}) \cdot (y^j - \overline{y}) \right)}{\sqrt{\sum_{j=1}^{N} \left((x^j - \overline{x})^2 \right) \cdot \sum_{j=1}^{N} \left((y^j - \overline{y})^2 \right)}}$$

- Measures linear dependencies
  - Always comprised between -1 and +1
  - 0 indicates decorrelation (no linear relation)

Correlation, a linear filter

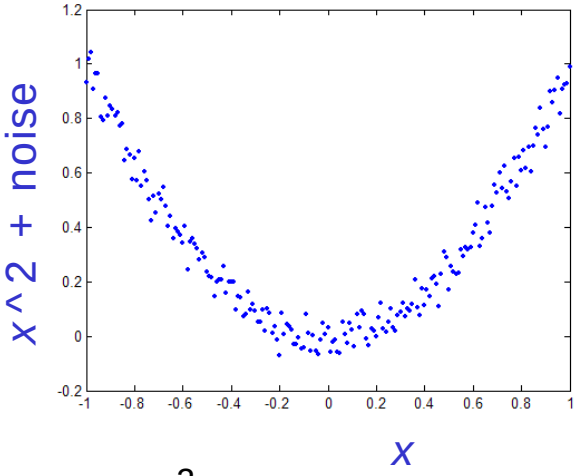
- Linear dependancy

Strong correlation



Weak correlation

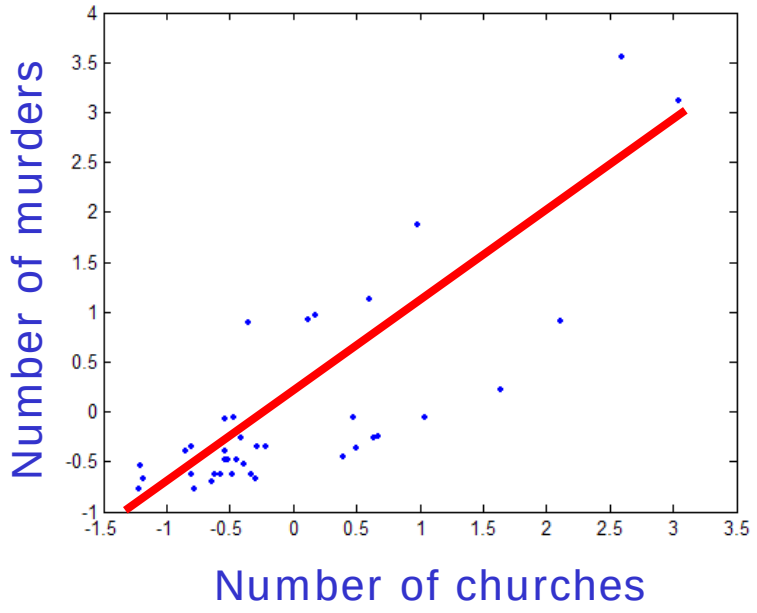
- Nonlinear dependancy



$r^2 \approx 0$

Correlation ≠ causality

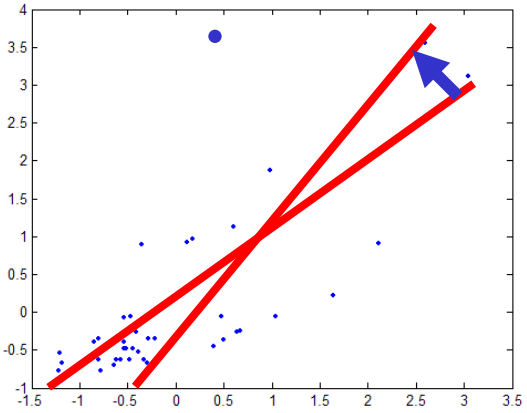
- High correlation does not mean causality
  - Number of murders in a city highly correlated (0.80) with number of churches
  - Simply because both murders and number of churches increase with population density



	christian chruches	murders 2002
Albuquerque	211	61
Atlanta	1500	152
Austin	353	25
Baltimore	466	253
Boston	370	60
Charlotte	505	67
Cleveland	980	80
Colorado Springs	400	25
Colombus	436	81
Denver	859	51
Detroit	1165	402
El paso	320	14
Fresno	450	42
Honolulu	39	18
Houston	1750	256
Indianapolis	1191	112
Jacksonville	21	3
Kansas city	1001	83
Long beach	236	67
Los Angeles	2000	654
Miami	911	65
milwaukee	411	111
Minneapolis	419	47
New Orleans	712	258
New York	2233	587
oakland	374	108
Oklahoma City	25	38
Omaha	236	26
philadelphia	963	288
Portland	498	20
St Louis	900	111
San Diego	373	47
San Francisco	540	68
San Jose	403	26
Seattle	482	26
Tucson	382	47
Tulsa	330	26
Virginia Beach	248	3
Washington	742	264

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 - NMSE)



Outline

- Motivation
- Feature selection in a nutshell
- Relevance criterion
- Mutual information
- Structured data
- Case studies
  - MI with missing data
  - MI with mixed data
  - MI for multi-label data
  - semi-supervised feature selection

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...

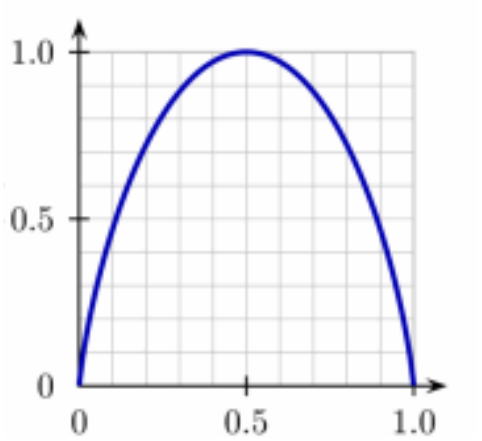
Entropy = uncertainty

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

- Can be interpreted as the average number of bits needed to describe y

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

- An example:
  Entropy of a binary variable y,
  $P[y=1] = p$ and $P[Y=0] = 1-p$



Conditional entropy

- Conditional entropy H(y |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 $\min(H(y), H(x))$

$$I(y;x) = \int\int p_{x,y}(x,y) \log \frac{p_{x,y}(x,y)}{p_x(x) p_y(y)} dx\, dy$$

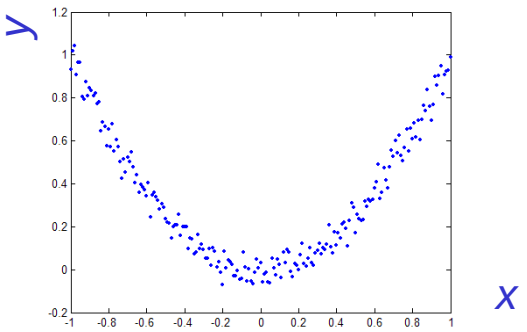
Nonlinear dependencies with MI

- Mutual information identifies nonlinear relationships between variables

- Example:
  - x uniformly distributed over [-1 1]
  - $y = x^2$ + noise

  x and y are dependant
  - z uniformly distributed over [-1 1]

  z and y are independent



- Results:

1000 samples	y,y	x,y	z,y
Correlation	1	0.0460	0.0522
Mutual information	2.2582	1.1996	0.0030

High-dimensional mutual information

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

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

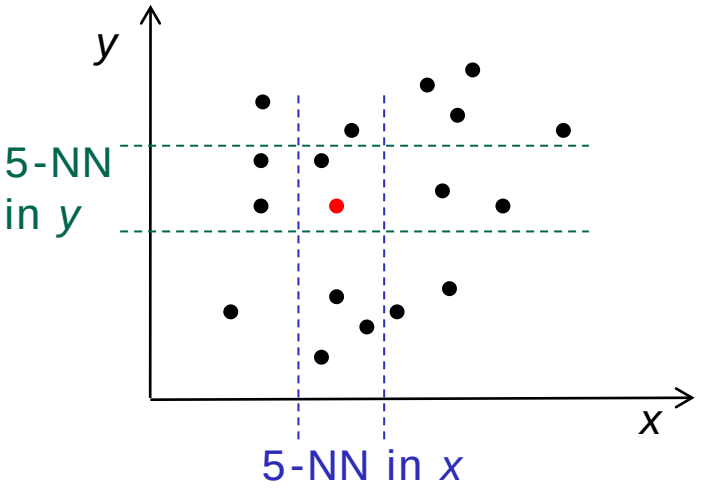
- x and y may be vectors!
- If x is a subset of features, its relevance may still be evaluated
- Evaluating subsets is the right issue!

The difficulty is in the *estimation* of $I(y;x)$

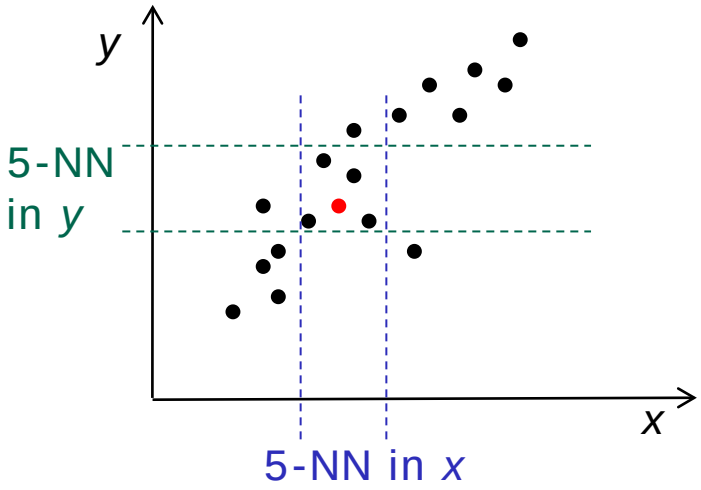
$$I(y;x) = \int\int p_{x,y}(x,y) \log \frac{p_{x,y}(x,y)}{p_x(x) p_y(y)} dx\, dy$$

- Need to estimate probability densities
  - In high-dimensional spaces if x are
- Histograms, kernels and splines suffer from the curse of dimensionality!
  - OK in dimension 2 (see mRmR in a few minutes…)
  - For large-dimensional spaces: k-NN based estimators are the (almost only) solution

K-NN to estimate MI



Only 1neighbor in common

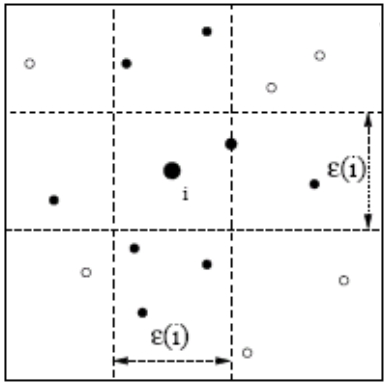


4 neighbors in common

- Why nearest neighbors?
  - More robust to curse of dimensionality
  - Do not suffer from the concentration of distances
  - But still hardly convincing what is a neighbor in a 20000-dim space!

K-NN to estimate MI

- Kraskov MI estimator
  - Based on Kozachenko-Leonenko estimator of entropy



$$\hat{I}(y;x) = \psi(N) + \psi(k) - \frac{1}{K} - \frac{1}{N}\sum_{n=1}^{N}\left(\psi(\tau_x(n)) + \psi(\tau_y(n))\right)$$

Permutation test

- Mutual information:
  - > 0
  - < entropy(features) (not known)

- Estimated mutual information:
  - can be slightly less than 0…
  - difficult to know if value is significant

$$I(y;x) = \int\int p_{x,y}(x,y) \log \frac{p_{x,y}(x,y)}{p_x(x) p_y(y)} dx\, dy$$

- Use permutation test!
  - permute the *y* but not the *x* in the learning set
  - marginals remain identical, but MI should drop to 0
  - repeat permutation to get a distribution of non-significant MI
  - compare the non-permuted one to the distribution → statistical test

Outline

- Motivation
- Feature selection in a nutshell
- Relevance criterion
- Mutual information
- Structured data
- Case studies
  - MI with missing data
  - MI with mixed data
  - MI for multi-label data
  - semi-supervised feature selection

But what about data?

- Traditionally analyzed data are like this:

X								Y
1	4	5	…	8	11	3	7	3
6	3	2	…	7	9	14	4	4
5	6	3	…	5	15	3	8	11
12	1	7	…	2	9	7	30	2
⋮	⋮	⋮		⋮	⋮	⋮	⋮	15
3	9	8	…	5	12	11	14	17
7	3	1	…	4	2	10	11	7

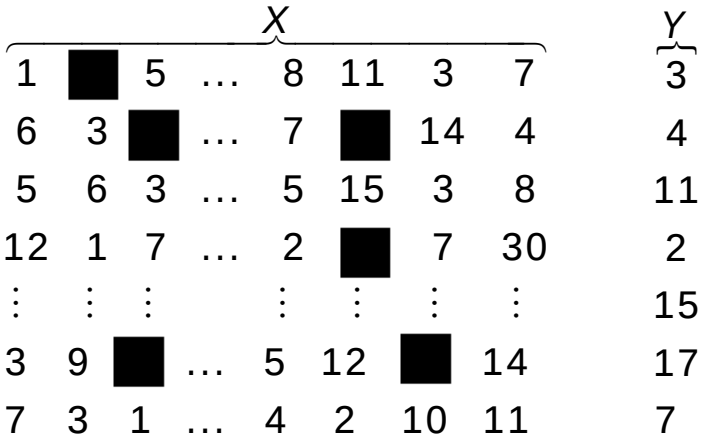
- But modern data analysis concerns *structured* data

Structured data

- Uncomplete
  - randomly missing data
  - different sizes of vectors
  - semi-supervsied data
- Complex
  - mixed discrete and real-valued data
- Non-conventional
  - possibilistic data, data known with some degree of certitude
  - data belonging to several classes
  - Data expressed as trees, graphs, etc.

Missing data

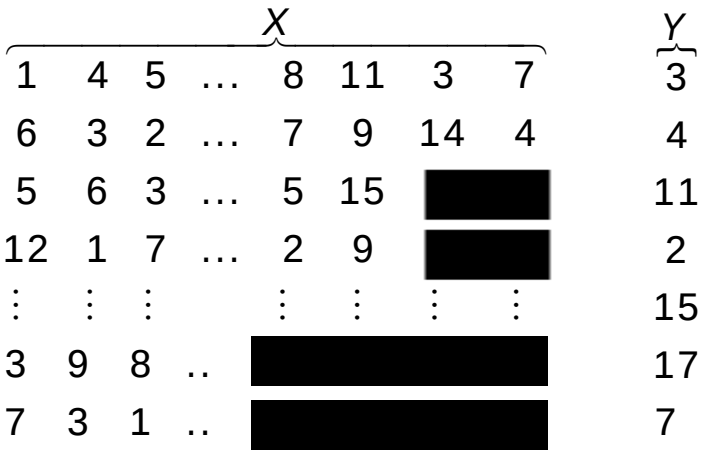
- Randomly missing data



  - measurement equipment failure
  - not answered questions in surveys
  - wrong data that to be removes
  - etc.

Missing data

- Different sizes of vectors



  - patient data in hospitals
  - etc.

Missing data

- Semi-supervised data

X								Y
1	4	5	…	8	11	3	7	3
6	3	2	…	7	9	14	4	■
5	6	3	…	5	15	3	8	11
12	1	7	…	2	9	7	30	■
⋮	⋮	⋮		⋮	⋮	⋮	⋮	■
3	9	8	…	5	12	11	14	17
7	3	1	…	4	2	10	11	7

  - some desired outputs are not known (labelling too expensive, experts not available, etc.)

Data in non-matrix form

- Graphs (social networks, phone call networks,…)
  - Classical question: clustering according to distances betwee nodes
  - But information on nodes is also available → multiobjective problem
  - Which information?

Wall | Info | Photos | Boxes

Basic Information

Birthday:	December 25, 1971
Hometown:	Tournai, Belgium
Relationship Status:	In a Relationship with

Personal Information

Interests:	voyages, déco intérieure, lecture

Contact Information

Email:	
Current City:	Dampremy, Belgium

Education and Work

Colleges:	Université Catholique de Louvain '95 pharmacie FUNDP '91 pharmacie
High School:	institut des ursulines tournai '89
Employer:	eupharma
Position:	pharmacien
Time Period:	September 1999 - Present
Location:	Châtelet, Belgium

Outline

- Motivation
- Feature selection in a nutshell
- Relevance criterion
- Mutual information
- Structured data
- Case studies
  - MI with missing data
  - MI with mixed data
  - MI for multi-label data
  - semi-supervised feature selection

MI with missing data

$$\hat{I}(y;x) = \psi(N) + \psi(k) - \frac{1}{K} - \frac{1}{N}\sum_{n=1}^{N}\left(\psi(\tau_x(n)) + \psi(\tau_y(n))\right)$$

- Just define the neighbours according to the known features
- Ex: $dist([3 \bullet 8 \; 7 \; 2],[1 \; 9 \bullet 3 \; 4]) = \sqrt{\frac{(3-1)^2 + (7-3)^2 + (2-4)^2}{3}}$
- Experiments:
  - 1 to 20% randomly chosen missing values
  - classical way: imputation before feature selection, compared to proposed way: feature selection (then imputation for regression)
  - imputation by k-NN or regularized EM
  - forward selection with MI, feature vectors of increasing size

MI with missing data

- Results
  - Delve census dataset

%	EM before	EM after	KNN before	KNN after
1	1,937 ± 0,090	1,892 ± 0,054	1,597 ± 0,111	1,902 ± 0,110
5	2,148 ± 0,052	1,911 ± 0,073	2,700 ± 0,282	2,272 ± 0,098
10	2,416 ± 0,392	1,972 ± 0,113	3,451 ± 0,146	2,674 ± 0,252
20	2,600 ± 0,126	2,174 ± 0,126	4,504 ± 0,257	2,872 ± 0,143

  - All improvements are significant from 5% of missing data
  - Other results available from

    G. Doquire, M. Verleysen, Mutual information for feature selection with missing data, to be presented at ESANN'2011

MI with mixed data

- Difficulties with mixed data
  - comparision between MI values for continuous and discrete features is hardly convincing
  - high-dimensional MI with discrete features is not very effective

- Solutions
  - use mRmR approach:

$$Score(x_i) = \underbrace{I(y; x_i)}_{\max \text{ Relevance}} - \underbrace{\frac{1}{|S|} \sum_{s \in S} I(x_i; x_s)}_{\min \text{ Redundancy}}$$

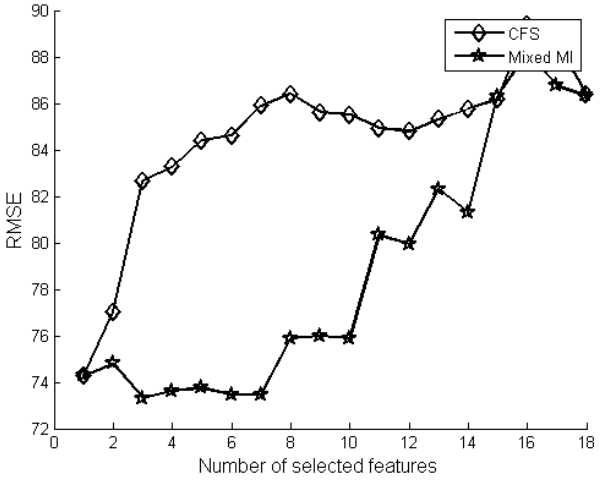
Restricted to 2-dimensional estimation (but approximation of *I*...)

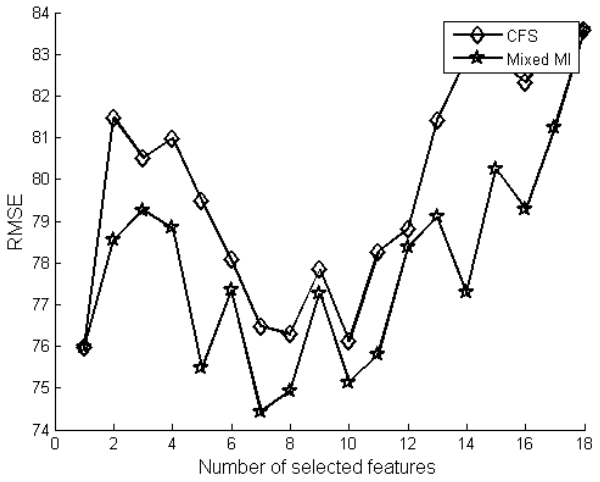
  - Keep the best *Score* for continuous and the best *Score* for discrete features
  - Decide (forward principle) by a wrapper (only 2 models to evaluate)

MI with mixed data

- Results
  - PCB dataset
    - 10 continuous and 8 categorical features
    - prediction by m5 regression tree and 5-NN
    - compared to CFS algorithm (mRmR approach based on correlation)

m5 tree





5-NN

  - Other results available from

    G. Doquire, M. Verleysen, Mutual information based feature selection for mixed data, to be presented at ESANN'2011

MI with multi-label data

- Multi-label: each instance can belong to several classes
- If each class is learned separately: loss of crucial information

- Standard procedure: Pruned Problem Transformation (PPT)
  - each unique set of labels is considered as a class
  - classes with too few instances are discarded

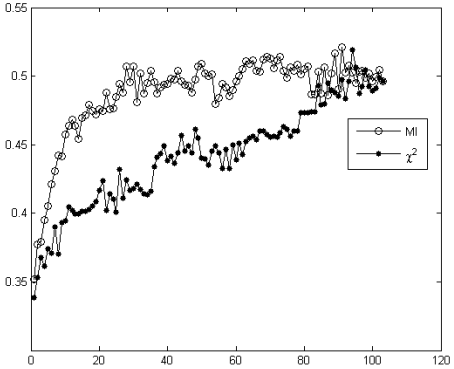
- Here MI necesitates *k* nearest neighbors → keep minimum *k*

MI with multi-label data

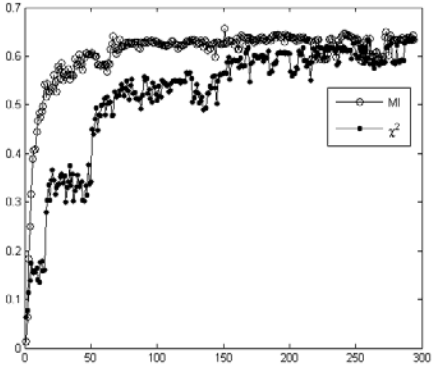
- Experiments:
  - Yeast dataset: 103 features and 14 possible labels
  - Scene dataset: 294features and 6 possible labels

  - Multi-label k-NN algorithm [Zhang and Zhou] used for evaluation
  - forward selection by MI
  - evaluation: accuracy as defined by $Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i \cap \hat{Y}_i\right|}{\left|Y_i \cup \hat{Y}_i\right|}$

Yeast: accuracy



# of features



# of features

Scene: accuracy

  - Other results available from

    G. Doquire, M. Verleysen, Feature selection for multi-label classification problems, to be presented at IWANN 2011

Semi-supervised learning

- Output labels are known for some instances only

- mRmR approach:

$$Score(x_i) = \underbrace{I(y; x_i)}_{\max \text{ Relevance}} - \underbrace{\frac{1}{|S|} \sum_{s \in S} I(x_i; x_s)}_{\min \text{ Redundancy}}$$

- Exploiting all the information:
  - redundandcy with all instances
  - relevance with labeled instances only

Laplacian score

- Laplacian score is used for unsupervised features selection
- Let x^n be a data point, and x_i^n its ith feature

- Unsupervised graph matrix: $S_{n,m}^{\text{uns}} = e^{-\frac{\left\| x^n - x^m \right\|}{t}}$
- Graph Laplacian: $L^{\text{uns}} = D^{\text{uns}} - S^{\text{uns}}$

- Laplacian score for each feature x_i (after centering):

$$L_i = \frac{x_i^T L^{\text{uns}} x_i}{x_i^T D^{\text{uns}} x_i} = \frac{\sum_{n,m} \left(x_i^n - x_i^m \right)^2 S_{n,m}^{\text{uns}}}{\text{var}(x_i)}$$

Laplacian score

- How to take supervised data into account?

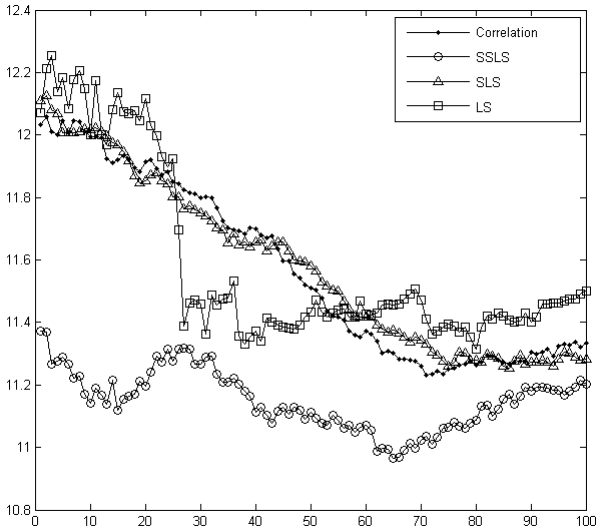
$$S_{n,m}^{sup} = e^{-\frac{\left\| y^n - y^m \right\|}{t}}$$

- Semi-supervised?
  - Use S^{sup} when both outputs are known
  - Use S^{unsup} otherwise
  - Apply some weighting (hyperparameter) between S^{sup} and S^{unsup}

- Results: juice dataset

  - Other results available from
    G. Doquire, M. Verleysen, Graph Laplacian for semi-supervised feature selection in regression problems, to be presented at IWANN 2011

RMSE



selected features

Conclusions

- Mutual information: the right concept to measure information from a (set of) feature(s)
- But it remains difficult to estimate in HD-spaces
  - there are effective (approximate) solutions: mRmR,...
- Big advantages
  - MI is multimensional by nature
  - MI can be easily extended to *structured* data
- Feature selection scheme depends on problem
  - linear or not
  - many features or not
  - computationally intensive model or not
  - ...