Science is to test crazy ideas – Engineering is to put these ideas into Business

Andreas Holzinger
VO 709.049 Medical Informatics
07.12.2016 11:15-12:45

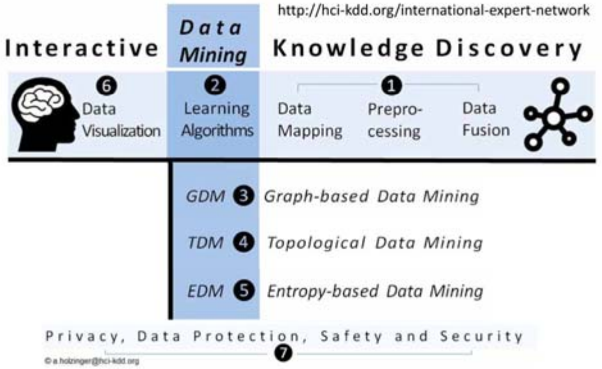
Lecture 07 Dimensionality Reduction and Subspace Clustering with the Doctor-in-the-Loop

a.holzinger@tugraz.at
Tutor: markus.plass@student.tugraz.at

http://hci-kdd.org/biomedical-informatics-big-data

ML needs a concerted effort fostering integrated research

http://hci-kdd.org/international-expert-network


Interactive
Data Mining
Knowledge Discovery
6 Data Visualization
2 Learning Algorithms
Data Mapping
1 Preprocessing
Data Fusion
GDM 3 Graph-based Data Mining
TDM 4 Topological Data Mining
EDM 5 Entropy-based Data Mining
Privacy, Data Protection, Safety and Security
7
© a.holzinger@hci-kdd.org

Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Keywords

- Classification
- Clustering
- Curse of dimensionality
- Dimensionality reduction
- Interestingness
- Feature extraction
- Feature selection
- Mapping
- Subspace analysis
- Subspace clustering

Advance Organizer (1/2)

- **Artificial neural network (ANN)** = a computational adaptive model (inspired by biological neural networks), consisting of interconnected groups of artificial neurons; processes information using a connectionist approach.
- **Association rule learning** = a set of techniques for discovering interesting relationships, i.e., "association rules," among variables in large databases used for data mining;
- **Classification** = a set of techniques to identify the categories in which new data points belong, based on a training set containing data points that have already been categorized; these techniques are often described as supervised learning because of the existence of a training set; they stand in contrast to cluster analysis, a type of unsupervised learning; used e.g. for data mining;
- **Cluster analysis** = statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance; a type of unsupervised learning because training data are not used - in contrast to classification; used for data mining.
- **Data mining** = a set of techniques to extract patterns from large data by combining methods from statistics and machine learning with database management (e.g. association rule learning, cluster analysis, classification, regression, etc.);
- **Knowledge Discovery (KD)** = process of identifying valid, novel, useful and understandable patterns out of large volumes of data

Advance Organizer (2/2)

- **Deep Learning** = class of machine learning algorithms using layers of non-linear processing units for feature extraction (remember: features are key for learning and understanding) - learning representations from data;
- **Knowledge Extraction** = is the creation of knowledge from structured (relational databases, XML) and unstructured (text, documents, images) sources;
- **Multimedia** = several data of different modalities are processed at the same time, i.e. encompassing audio data (sound, speech), image data (b/w and colour images), video data (time-aligned sequences of images), electronic ink (sequences of time aligned 2D and 3D coordinates of a stylus, pen, data gloves etc.)
- **Principal Component Analysis (PCA)** = statistical technique for finding patterns in high-dimensional data;
- Sparse Data =
- **Supervised learning** = inferring a function from supervised training data on the basis of training data which consist of a set of training examples, the input objects (typically vectors) and a desired output value (also called the supervisory signal).
- **Supervised learning algorithm** = analyzes the training data and produces an inferred function, called a classifier (if the output is discrete) or a regression function (if the output is continuous); the algorithm generalizes from the training data to unseen situations.
- **Support vector machine (SVM)** = concept for a set of related supervised learning methods to analyze data and recognize patterns, used for classification and regression analysis.
- **Unsupervised learning** = establishes clusters in data, where the class labels of training data is unknown.

Glossary

- ANN = Artificial Neural Network
- ANOVA = Analysis of Variance
- AUC - area under the curve
- CDT = Clinical Decision Tree
- DM = Data Mining
- KDD = Knowledge Discovery from Data(bases)
- LLE = Locally Linear Embedding
- MDS = Multi Dimensional Scaling
- MELD = model for end-stage liver disease
- MM = Multimedia
- NLP = Natural Language Processing
- PCA = Principal Components Analysis
- ROC = Receiver Operating Characteristic
- SVM = Support Vector Machine

Learning Goals: At the end of this lecture you ...

- ... know the differences between **classification** and **clustering** and why it is important for health;
- ... are aware that **features** are key to learning and understanding;
- ... understand the **curse of dimensionality**;
- ... have an idea of **dimensionality reduction**;
- ... recognize the value of **subspace clustering** and analysis with the **doctor-in-the-loop**;
- Understand why the question **"What is interesting?"** is not easy to answer;

Key Challenges

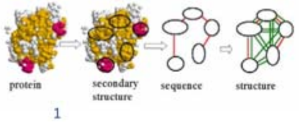
- Uncertainty
- Validation
- Curse of Dimensionality
- Large spaces gets sparse
- Distance Measures get useless
- Patterns occur in different subspaces
- Pressing question: "What is interesting?"

Agenda for today

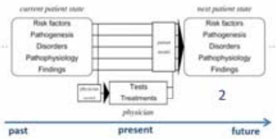
- **00 Reflection – follow-up from last lecture**
- **01 Classification vs. Clustering**
- **02 Feature Engineering**
- **03 Curse of Dimensionality**
- **04 Dimensionality Reduction**
- **05 Subspace Clustering and Analysis**
- **06 Projection Pursuit**
- **07 Conclusion and Future Challenges**

00 Reflection

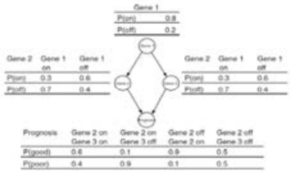
Warm-up Quiz



1



2



Gene 1	
P(on)	0.8
P(off)	0.2

Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

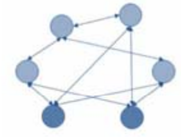
Gene 2	Gene 1 on	Gene 1 off
P(on)	0.3	0.6
P(off)	0.7	0.4

Prognosis	Gene 2 on Gene 3 on	Gene 2 on Gene 3 off	Gene 2 off Gene 2 on	Gene 2 off Gene 3 off
P(good)	0.6	0.1	0.9	0.5
P(poor)	0.4	0.9	0.1	0.5

3

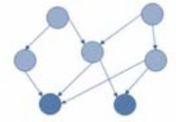
4

Three types of Probabilistic Graphical Models



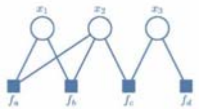
Undirected: Markov random fields, useful e.g. for computer vision (Details: Murphy 19)

$$P(X) = \frac{1}{Z}\exp\left(\sum_{ij} W_{ij}\, x_i x_j + \sum_i x_i b_i\right)$$



Directed: Bayes Nets, useful for designing models (Details: Murphy 10)

$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

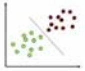


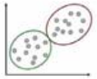
Factored: useful for inference/learning

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s)$$

Tutorial on Factor Graphs http://deepdive.stanford.edu/inference

01 Classification vs. Clustering





Classification (A) vs. Clustering (C) – Intro Quiz

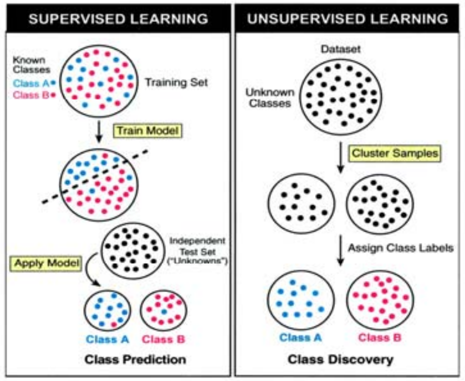
1) The data is not labeled (clA/Clu)?
2) Identify structure/patterns (clA/Clu)?
3) Predicting an item set, identify to which set of categories a new observation belongs (clA/Clu)?
4) Assigning a set of objects into groups (clA/Clu)?
5) Having many labelled data points (clA/Clu)
6) Using the concept of supervised learning (clA/Clu)?
7) Grouping data items close to each other (clA/Clu)?
8) Used to explore data sets (clA/Clu)?

Classification vs. Clustering

- Classification
  - Supervised learning, Pattern Recognition, Prediction, …)
  - Supervision = the training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations;
  - New data is **classified based on the training set**
  - Important for clinical decision making
  - Example: Benign/Malign Classification of Tumors
- Clustering
  - Unsupervised learning, class discovery, …
  - The class labels of training data is unknown
  - Given a set of measurements, observations, etc. with the aim of **establishing the existence of clusters** in the data;
  - Example: K-Means Algorithm for disease clustering

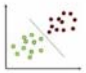
15

Discovery vs. Prediction

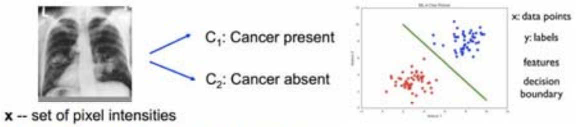


Ramaswamy, S. & Golub, T. R. (2002) DNA Microarrays in Clinical Oncology. *Journal of Clinical Oncology, 20, 7, 1932-1941.*

Why do we need Classification in Health Informatics?

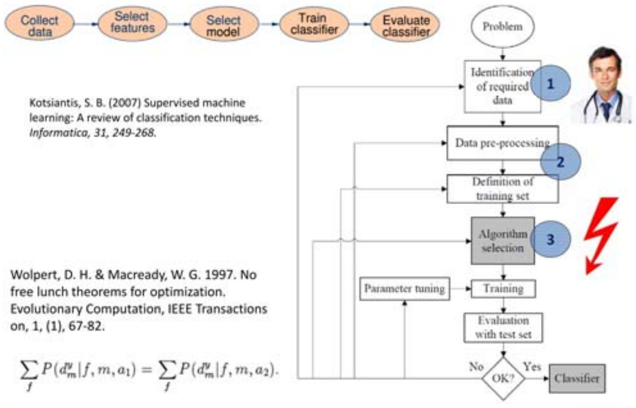


Why do we need Classification?



- Typical questions include:
  - Is this protein functioning as an enzyme?
  - Does this gene sequence contain a splice site?
  - Is this melanoma malign?
- Given object x – predict the class label y
  - If $y \in \{0,1\} \rightarrow$ binary classification problem
  - If $y \in \{1, \ldots, n\}$ and is $n \in \mathbb{N} \rightarrow$ multiclass problem
  - If $y \in \mathbb{R} \rightarrow$ regression problem

Learning Process: Algorithm selection is crucial



Kotsiantis, S. B. (2007) Supervised machine learning: A review of classification techniques. *Informatica, 31, 249-268.*

Wolpert, D. H. & Macready, W. G. 1997. No free lunch theorems for optimization. Evolutionary Computation, IEEE Transactions on, 1, (1), 67-82.

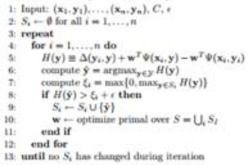
$$\sum_f P(d_m^y|f,m,a_1) = \sum_f P(d_m^y|f,m,a_2).$$

Classifiers Examples

- Naïve Bayes (NB) – see Bayes' theorem with independent assumptions (hence "naïve")
- Decision Trees (e.g. C4.5)
- NN – if x_1 is most similar to $x_2 \Rightarrow\ y_1 = y_2$

$$x_j = argmin_{x \in D}||x - x_i||^2 \Rightarrow y_i = y_j$$

- SVM – a plane/hyperplane separates two classes of data – very versatile for classification and clustering – also via the Kernel trick in high-dimensions

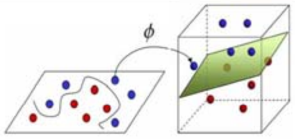


```
Input: $(\mathbf{x}_1,\mathbf{y}_1),\ldots,(\mathbf{x}_n,\mathbf{y}_n)$, $C$, $\epsilon$
$S_i \leftarrow \emptyset$ for all $i = 1,\ldots,n$
repeat
  for $i = 1,\ldots,n$ do
    $H(\mathbf{y}) \equiv \Delta(\mathbf{y}_i,\mathbf{y}) + \mathbf{w}^T\Psi(\mathbf{x}_i,\mathbf{y}) - \mathbf{w}^T\Psi(\mathbf{x}_i,\mathbf{y}_i)$
    compute $\hat{\mathbf{y}} = \mathrm{argmax}_{\mathbf{y}\in\mathcal{Y}} H(\mathbf{y})$
    compute $\xi_i = \max\{0, \max_{\mathbf{y}\in S_i} H(\mathbf{y})\}$
    if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ then
      $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
      $\mathbf{w} \leftarrow$ optimize primal over $S = \bigcup_i S_i$
    end if
  end for
until no $S_i$ has changed during iteration
```

Finley, T. & Joachims, T. Supervised clustering with support vector machines. Proceedings of the 22nd international conference on Machine learning, 2005. ACM, 217-224.

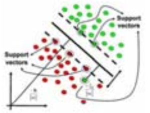
SVM – Vapnik, 1992

- Uses a nonlinear mapping to transform the original data (input space) into a higher dimension (feature space)

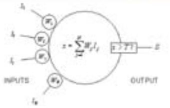


- = classification method for both linear and nonlinear data;
- Within the new dimension, it searches for the linear optimal separating **hyperplane** (i.e., "decision boundary");
- By nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated with a hyperplane;
- The SVM finds this hyperplane by using **support vectors** (these are the "essential" training tuples) and **margins** (defined by the support vectors);

SVM vs. ANN

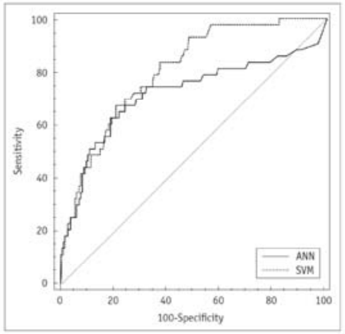


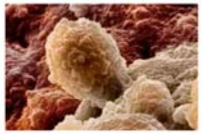
- SVM
  - Deterministic algorithm
  - Nice generalization properties
  - Hard to learn – learned in batch mode using quadratic programming techniques
  - Using kernels can learn very complex functions



- ANN
  - Nondeterministic algorithm
  - Generalizes well but doesn't have strong mathematical foundation
  - Can easily be learned in incremental fashion
  - To learn complex functions—use multilayer perceptron (nontrivial)

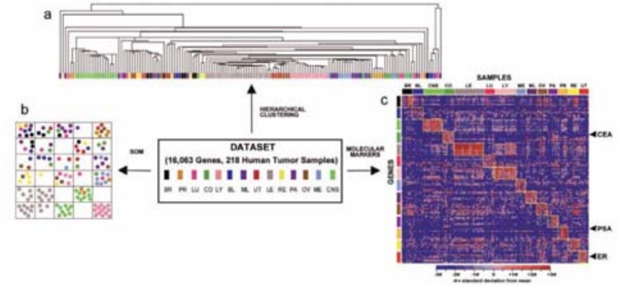
Clinical use: SVM are more accurate than ANN





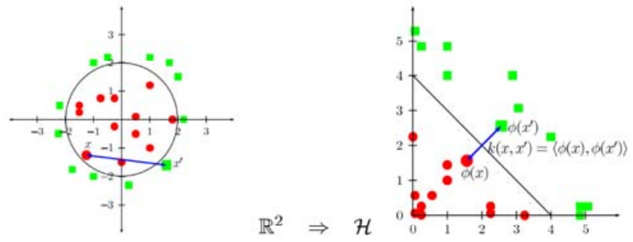
Kim, S. Y., Moon, S. K., Jung, D. C., Hwang, S. I., Sung, C. K., Cho, J. Y., Kim, S. H., Lee, J. & Lee, H. J. (2011) Pre-Operative Prediction of Advanced Prostatic Cancer Using Clinical Decision Support Systems: Accuracy Comparison between Support Vector Machine and Artificial Neural Network. *Korean J Radiol, 12,* ***5, 588-594.***

Example: Multiclass cancer diagnosis (for Exercise)



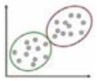
Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C.-H., Angelo, M., Ladd, C., Reich, M., Latulippe, E. & Mesirov, J. P. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences, 98, (26), 15149-15154, doi:10.1073/pnas.211566398.

Counterexample: Move problem to a feature space $\mathcal{H}$

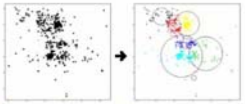


Borgwardt, K., Gretton, A., Rasch, J., Kriegel, H.-P., Schölkopf, B. & Smola, A. 2006. Integrating structured biological data by kernel max. mean discrepancy. Bioinformatics, 22, 14, e49-e57.

Why do we need Clustering in Health Informatics?



Why do we need Clustering?



- Group similar objects into clusters together, e.g.
  - For image segmentation
  - Grouping genes similarly affected by a disease
  - Clustering patients with similar diseases
  - Cluster biological samples for category discovery
  - Finding subtypes of diseases
  - Visualizing protein families
- Inference: given x_i, predict y_i by learning f
- No training data set – learn model and apply it

Example K-means

- Partite a set of n observations into k clusters so, that the intra-cluster variance is $argmin$
  - v ... variance (objective function)
  - S_i ... cluster
  - c_j ... mean ("centroid" for cluster j)
  - D ... set of all data points x_j
  - k ... number of clusters

Distance "medoid"

$$v(D) = argmin \sum_{j=1}^{k} \sum_{i=1}^{n} \|(x_i - c_j)\|^2$$

Jain, A. K. 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31, (8), 651-666, doi:http://dx.doi.org/10.1016/j.patrec.2009.09.011.

Example

Algorithm 1: Example for a classical weight balanced k-means algorithm

```
Input: $d, k, n \in \mathbb{N}$, $X := \{x_1, \dots, x_n\} \subset \mathbb{R}^d$, $S := \{s_1, \dots, s_k\} \subset \mathbb{R}^d$
Output: Clustering $C = (C_1, \dots, C_k)$ of $X$ and the arithmetic means $c_1, \dots, c_k$ as sites
1. Partition $X$ into a clustering $C = (C_1, \dots, C_k)$ by assigning $x_j \in X$ to a cluster
   $C_i$ that is closest to site $s_i \in S$.
2. Update each site $s_i$ as the center of gravity of cluster $C_i$; if $|C_i| = 0$, choose
   $s_i = x_l$ for a random $l \leq n$ with $x_l \neq s_j$ for all $j \leq k$. If the sites change, go to
   (1.).
```

Merely an increase in awareness of physicians on risk factors for ARA in children can be sufficient to change their attitudes towards antibiotics prescription.
Our results can also be useful when preparing recommendations for antibiotics prescription and to guide the standardized health data record.

Yildirim, P., Majnarić, L., Ekmekci, O. I. & Holzinger, A. 2013. On the Prediction of Clusters for Adverse Reactions and Allergies on Antibiotics for Children to Improve Biomedical Decision Making. In: Lecture Notes in Computer Science LNCS 8127. 431-445

Summary: The 10 top algorithms everybody should know

Wu et al. (2008) Top 10 algorithms in data mining. *Knowledge & Information Systems, 14, 1, 1-37.*

- **C4.5**
  - for generation of decision trees used for **classification,** (statistical classifier, Quinlan (1993));
- **k-means**
  - simple iterative method for partition of a dataset in a user-specified n of **clusters**, k (Lloyd (1957));
- **A-priori**
  - for finding frequent item sets using candidate generation and **clustering** (Agrawal & Srikant (1994));
- **EM**
  - Expectation–Maximization algorithm for finding maximum likelihood estimates of parameters in models (Dempster et al. (1977));
- **PageRank**
  - a search ranking algorithm using hyperlinks on the Web (Brin & Page (1998));
- **Adaptive Boost**
  - one of the most important ensemble methods (Freund & Shapire (1995));
- **k-Nearest Neighbor**
  - a method for **classifying** objects based on closest training sets in the feature space (Fix & Hodges (1951));
- **Naive Bayes**
  - can be trained efficiently in a supervised learning setting for classification (Domingos & Pazzani (1997));
- **CART**
  - **Classification** And Regression Trees as predictive model mapping observations about items to conclusions about the goal (Breiman et al 1984);
- **SVM** *support vector machines offer one of the most robust and accurate methods among all well-known algorithms (Vapnik (1995));*

02
Feature Engineering

"Applied Machine Learning is basically feature engineering".

Andrew Y. Ng, VP & Chief Scientist of Baidu;
Co-Chair/Founder of Coursera; Professor at Stanford University

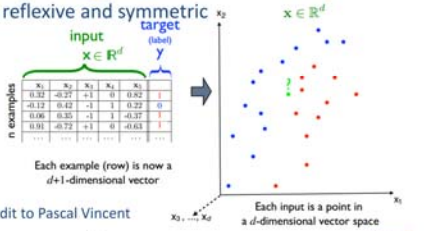
http://www.andrewng.org

Advance Organizer

- Feature:= specific measurable property of a phenomenon being observed.
- Feature engineering:= using domain knowledge to create features useful for ML. **("Applied ML is basically feature engineering. *Andrew Ng*").**
- Feature learning:= transformation of raw data input to a representation, which can be effectively exploited in ML.

Feature Space Basic Definitions

- Intuitively: a domain with a distance function
- Formally: Feature Space $\mathcal{F} = (\mathcal{D}, d)$
  - $\mathcal{D}$ = ordered set of features
  - $d\colon D \times D \rightarrow \mathbb{R}_0^+$ ... a total distance function; true for
    - $\forall p, q \in \mathcal{D}, p \neq q\colon d(p,q) > 0$ (strict)
    - and must be reflexive and symmetric



x_1	x_2	x_3	x_4	x_5	
0.32	-0.27	+1	0	0.82	1
-0.12	0.42	-1	1	0.22	0
0.06	0.35	-1	1	-0.37	1
0.91	-0.72	+1	0	-0.63	1
...	...	...	...	...	...

Each example (row) is now a d+1-dimensional vector

Each input is a point in a d-dimensional vector space

Image credit to Pascal Vincent

Metric Space (e.g. Euclidean Vector Space)

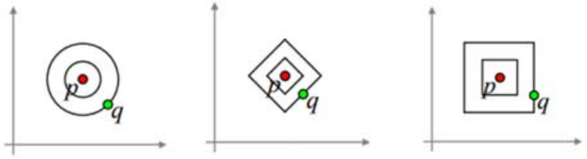
A **Metric Space** is a pair (X, d) where X is a set and $d : X \times X \rightarrow \mathbb{R}^+$, called the metric, s.t.

1. For all $x, y, z \in X$, $d(x,y) \leq d(x,z) + d(z,y)$.
2. For all $x, y \in X$, $d(x,y) = d(y,x)$.
3. $d(x,y) = 0$ if and only if $x = y$.

Remark 1. *One example is $\mathbb{R}^d$ with the Euclidean metric. Spheres S^n endowed with the spherical metric provide another example.*

$$d : \mathcal{X} \rightarrow \mathbb{R}$$
$$d(x,x) = 0$$
$$d(x^1, x^2) = d(x^2, x^1) \quad \text{symmetry}$$
$$d(x^1, x^2) \leq d(x^1, x^3) + d(x^3, x^2) \quad \text{triangle inequality}$$

Similarities of feature vectors

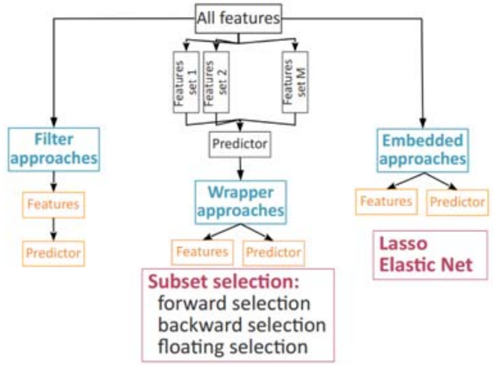


Euclidian norm

Manhattan norm

Maximums norm

Feature Selection: Overview



All features

Features set 1

Features set 2

Features set M

Filter approaches

Features

Predictor

Predictor

Wrapper approaches

Features

Predictor

Subset selection:
forward selection
backward selection
floating selection

Embedded approaches

Features

Predictor

Lasso
Elastic Net

Image credit to Chloe Azencott

Feature Selection vs. Feature Extraction

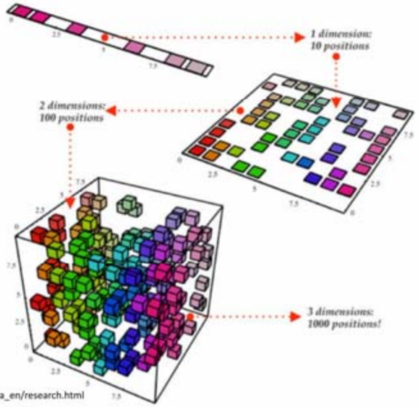
- Feature selection is just selecting a subset of the existing features without any transformation
- Feature extraction is *transforming* existing features into a lower dimensional space

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature selection}} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \xrightarrow{\text{feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}\right)$$

Blum, A. L. & Langley, P. 1997. Selection of relevant features and examples in machine learning. Artificial intelligence, 97, (1), 245-271.

03 Curse of Dimensionality

Remember: The curse of dimensionality

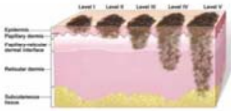

1 dimension: 10 positions
2 dimensions: 100 positions
3 dimensions: 1000 positions!

Bengio, S. & Bengio, Y. 2000. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Transactions on Neural Networks, 11, (3), 550-557.

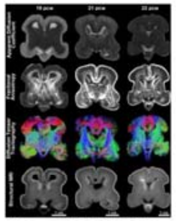
http://www.iro.umontreal.ca/~bengioy/yoshua_en/research.html

Examples for High-Dimensional Biomedical Data (1/6)

- Medical Image Data (16 - 1000+ features)



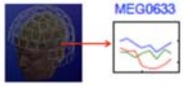
http://qsota.com/melanoma/



Nature 508, 199–206 doi:10.1038/nature13185

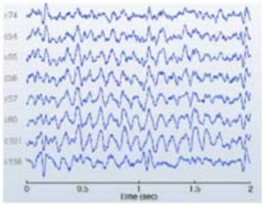
MEG Brain Imaging

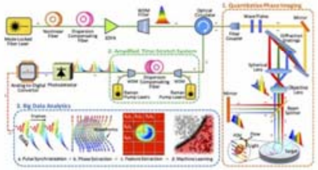
120 locations x 500 time points x 20 objects



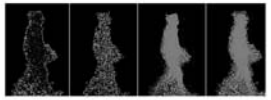
Examples for High-Dimensional Biomedical Data (2/6)

- Biomedical Signal Data (10 - 1000+ features)

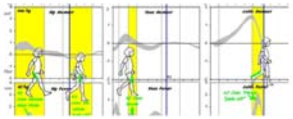




http://www.nature.com/articles/srep21471#f1



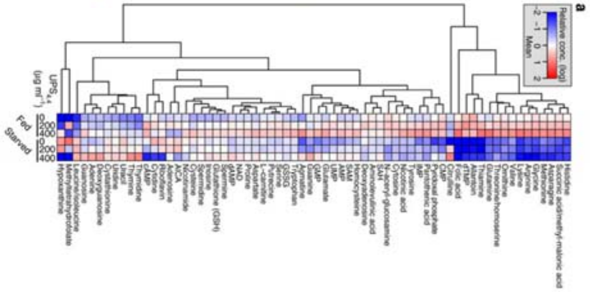
http://www.mdpi.com/1424-8220/14/4/6124/htm



http://www.clinicalgaitanalysis.com/data/

Examples for High-Dimensional Biomedical Data (3/6)

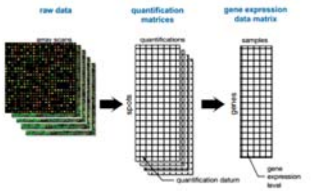
- Metabolome data (feature is the concentration of a specific metabolite; 50 – 2000+ features)

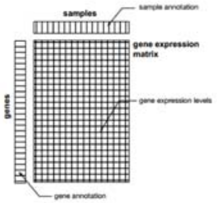


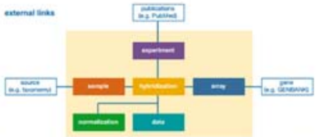
http://www.nature.com/ncomms/2015/151005/ncomms9524/fig_tab/ncomms9524_F5.html

Examples for High-Dimensional Biomedical Data (4/6)

Microarray Data (features correspond to genes, up to 30k features)







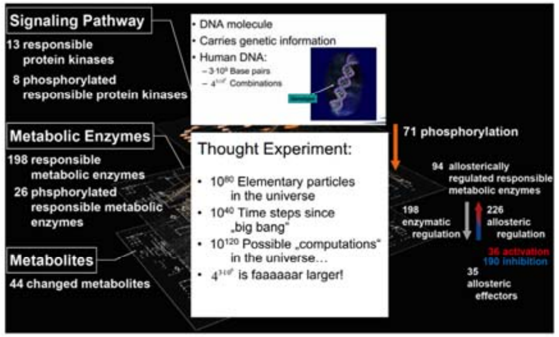
Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A. & Causton, H. C. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature genetics, 29, (4), 365-371.

Examples for High-Dimensional Biomedical Data (5/6)

- Text > 10^9 documents × 10^6 words/n-grams features correspond to words or terms, between 5k to 20k features
- Text (Natural Language) is definitely very important for health:
  - Handwritten Notes, Drawings
  - Patient consent forms
  - Patient reports
  - Radiology reports
  - Voice dictations, annotations
  - Literature !!!

https://www.researchgate.net/publication/255723699_An_Answer_to_Who_Needs_a_Stylus_on_Handwriting_Recognition_on_Mobile_Devices

Example for high-dimensional biomedical data (6/6)



Signaling Pathway

13 responsible protein kinases

8 phosphorylated responsible protein kinases

- DNA molecule
- Carries genetic information
- Human DNA:
  - $3 \cdot 10^9$ Base pairs
  - $4^{3 \cdot 10^9}$ Combinations

Metabolic Enzymes

198 responsible metabolic enzymes

26 phshorylated responsible metabolic enzymes

Thought Experiment:

- 10^{80} Elementary particles in the universe
- 10^{40} Time steps since „big bang"
- 10^{120} Possible „computations" in the universe…
- $4^{3 \cdot 10^9}$ is faaaaaar larger!

Metabolites

44 changed metabolites

71 phosphorylation

94 allosterically regulated responsible metabolic enzymes

198 enzymatic regulation

226 allosteric regulation

36 activation

190 inhibition

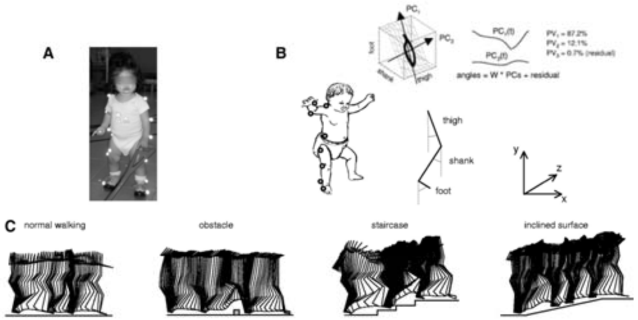
35 allosteric effectors

Yugi, K. et al. 2014. Reconstruction of Insulin Signal Flow from Phosphoproteome and Metabolome Data. Cell Reports, 8, (4), 1171-1183, doi:10.1016/j.celrep.2014.07.021.

Why are many features problematic?

- Hyperspace is large – all points are far apart
- Computationally challenging (both in time & space)
- Complexity grows with n of features
- Complex models less robust – more variance
- Statistically challenging – hard to learn
- Hard to interpret and **hard to visualize (humans are bound to R3/R2!)**
- Problem with redundant features and noise
- Question: Which algorithms will provide worse results with increasing irrelevant features?
- Answer: Distance-based algorithms generally trust all features of equal importance

Space and Time: Simple example on gait analysis



Dominici, N., Ivanenko, Y. P., Cappellini, G., Zampagni, M. L. & Lacquaniti, F. 2010. Kinematic Strategies in Newly Walking Toddlers Stepping Over Different Support Surfaces. Journal of Neurophysiology, 103, (3), 1673-1684, doi:10.1152/jn.00945.2009.

04 Dimensionality Reduction

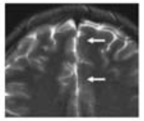
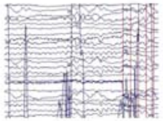
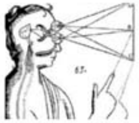
Why should we reduce the dimensionality?

- Data visualization only possible in $\mathbb{R}2$ (R3 cave)
- Human interpretability only in R2/R3 (visualization can help sometimes with parallel coordinates)
- Simpler (=less variance) models are more robust
- Computational complexity (time and space)
- Eliminate non-relevant attributes that can make it more difficult for algorithms to learn
- Bad results through (many) irrelevant attributes?
- *Note again: Distance-based algorithms generally trust that all features are equally important.*

Problem

- Given n data points in d dimensions
- Conversion to m data points in $r \ll d$ dimension
- Challenge: **minimal loss of information *)**

- *) this is always a grand challenge, e.g. in k-Anonymization – see later
- Very dangerous is the "modeling-of-artifacts"

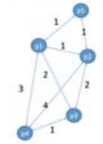
Approaches

- Linear methods (unsupervised):
  - PCA (Principal Component Analysis)
  - FA (Factor Analysis)
  - **MDS (Multi-dimensional Scaling)**
- Non-linear methods (unsupervised):
  - Isomap (Isometric feature mapping)
  - LLE (locally linear embedding)
  - Autoencoders
- Supervised methods:
  - LDA (Linear Discriminant Analysis)
- **Subspace Clustering with a human-in-the-loop**

Example 4: MDS: Find projection that best preserves *d*

- Given n x n matrix of pairwise distances between data points
- Compute n x k matrix X with coordinates of distances with some linear algebra magic
- Perform PCA on this matrix X

	p1	p2	p3	p4	p5
p1	0	1	2	3	1
p2	1	0	2	4	1
p3	2	2	0	1	3
p4	3	4	1	0	1
p5	1	1	3	1	0



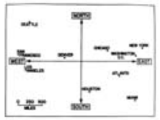
x_i Point in d dimensions

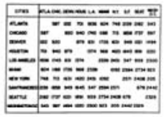
y_i Corresponding point in $r < d$ dimensio

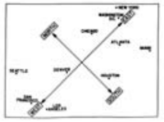
δ_{ij} Distance between x_i and x_j

d_{ij} Distance between y_i and y_j

- Define (e.g.) $E(\mathbf{y}) = \sum_{i,j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right)$
- Find y_i's that minimize E by gradient descent
- Invariant to translations, rotations and scalings







Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, (1), 1-27.

05 Subspace Clustering* and Analysis

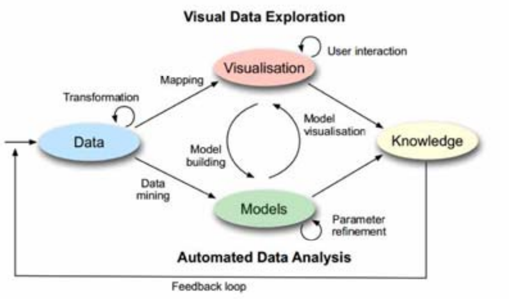
* Two major issues
(1) the algorithmic approach to clustering and
(2) the definition and assessment of **similarity versus dissimilarity**.

What is subspace clustering?

- Definitions:
- K clusters
- N data points
- D dimensions (original space)
- d dimensions (latent subspace)
- Subspace Clustering is the process of clustering data whilst reducing the d of each cluster to a cluster-dependent subspace

Agrawal, R., Gehrke, J., Gunopulos, D. & Raghavan, P. 1998. Automatic subspace clustering of high dimensional data for data mining applications. SIGMOD Rec., 27, (2), 94-105, doi:10.1145/276305.276314.

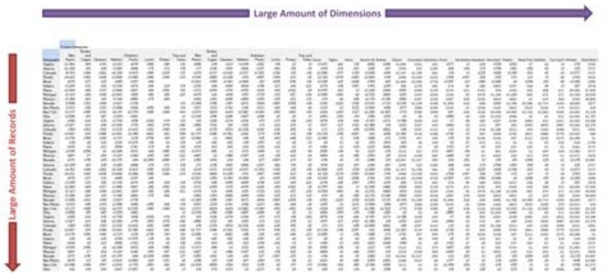
Visual Analytics Pipeline



Keim, D., Kohlhammer, J., Ellis, G. & Mansmann, F. (eds.) 2010. Mastering the Information Age: Solving Problems with Visual Analytics, Goslar: Eurographics.

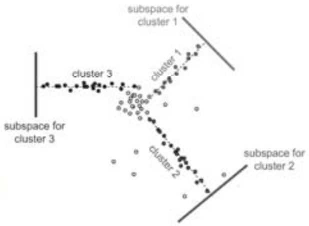
http://www.vismaster.eu/wp-content/uploads/2010/11/VisMaster-book-lowres.pdf

High-Dimensional Data e.g. from patient records


Large Amount of Dimensions
Large Amount of Records

High-Dimensional Data – The Curse of Dimensionality

- Many irrelevant dimensions
- Correlated and redundant dimensions
- Conflicting dimensions
- Wrong Interpretation of global analysis results

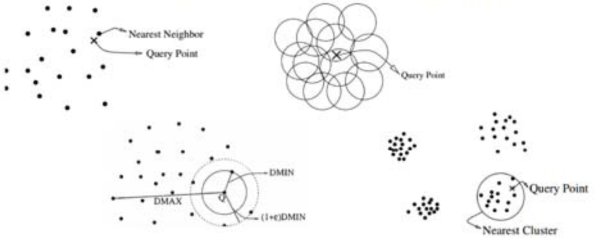


Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. 1999. When is "nearest neighbor" meaningful? *In:* Beeri, C. & Buneman, P. (eds.) *Database Theory ICDT 99, LNCS 1540.* Berlin: Springer, pp. 217-235.

Kriegel, H. P., Kroger, P. & Zimek, A. 2009. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. ACM Transactions on Knowledge Discovery from Data (TKDD), 3, (1), 1-58, doi:10.1145/1497577.1497578.

High-Dimensional Data – The Curse of Dimensionality

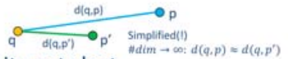
- NN problem: Given n data points and a query point in an m −dimensional metric space
- find the data point closest to the query point.



Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. 1999. When is "nearest neighbor" meaningful? *In:* Beeri, C. & Buneman, P. (eds.) *Database Theory ICDT 99, LNCS 1540.* Berlin: Springer, pp. 217-235.

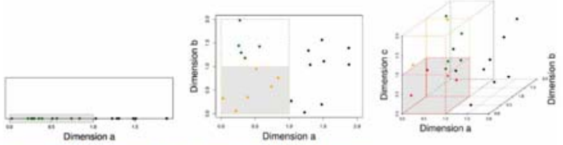
Challenges in High-Dim Data – Curse of Dimensionality

- Concentration Effect
  - Discriminability of similarity gets lost
  - Impact on usefulness of a similarity measure



Simplified(!)
$\#dim \rightarrow \infty: d(q,p) \approx d(q,p')$

- **High-Dimensional Data is Sparse**

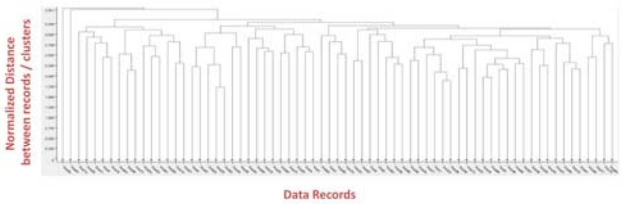


Optimization Problem and Combinatorial Issues

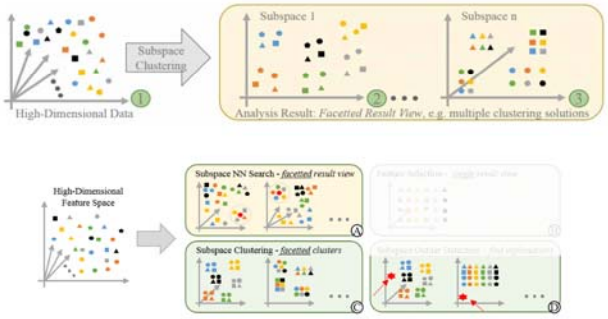
Feature selection and dimension reduction

2^d−1 possible subsets of dimensions (-> subspaces)

Example: Full Space Clustering of High-Dimensional Data

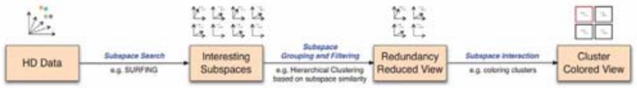


Overview of (major?) Subspace Analysis Techniques

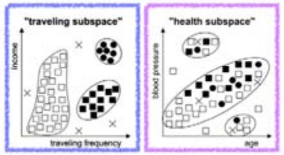

High-Dimensional Data
1
Subspace Clustering
Subspace 1
2
Subspace n
3
Analysis Result: Facetted Result View, e.g. multiple clustering solutions
High-Dimensional Feature Space
Subspace NN Search - facetted result view
A
B
Subspace Clustering - facetted clusters
C
D

- **Patterns may be found in subspaces (dimension combinations)**
- **Patterns may be complementary or redundant to each other**

Subspace Concept

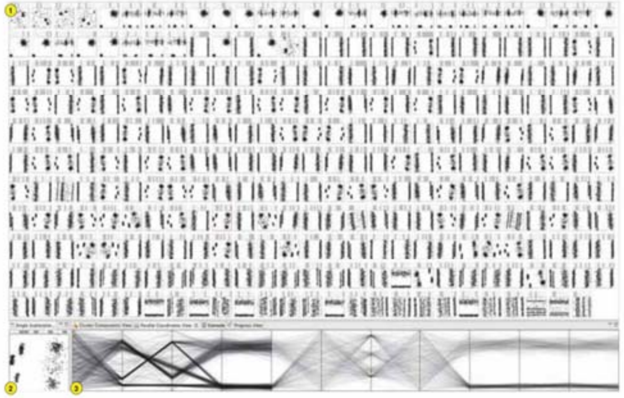


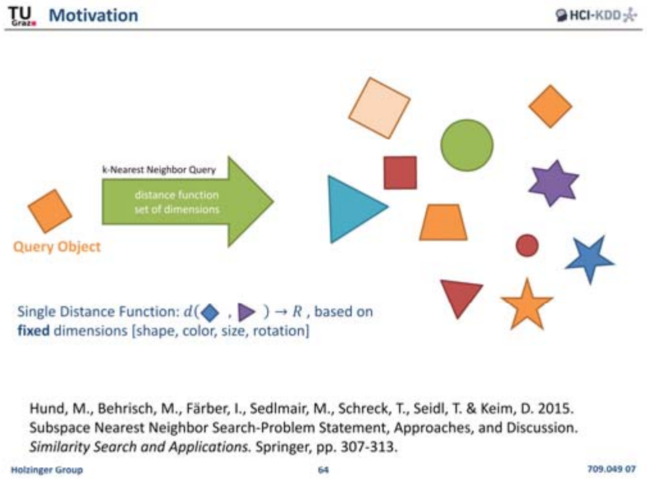
objectID	age	blood pres.	sportactiv	income	trav. freq.
1	ABC	ABC	ABC	ABC	ABC
2	ABC	ABC	ABC	ABC	ABC
3	ABC	ABC	ABC	ABC	ABC
4	ABC	ABC	ABC	ABC	ABC
5	ABC	ABC	ABC	ABC	ABC
6	ABC	ABC	ABC	ABC	ABC
7	ABC	ABC	ABC	ABC	ABC
8	ABC	ABC	ABC	ABC	ABC
9	ABC	ABC	ABC	ABC	ABC



Tatu, A., Maass, F., Faerber, I., Bertini, E., Schreck, T., Seidl, T. & Keim, D. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. IEEE Symposium onVisual Analytics Science and Technology (VAST), 2012 Seattle. IEEE, 63-72, doi:10.1109/VAST.2012.6400488.

Example of 12D Data -> 4095 subspaces (296 interesting)





Motivation

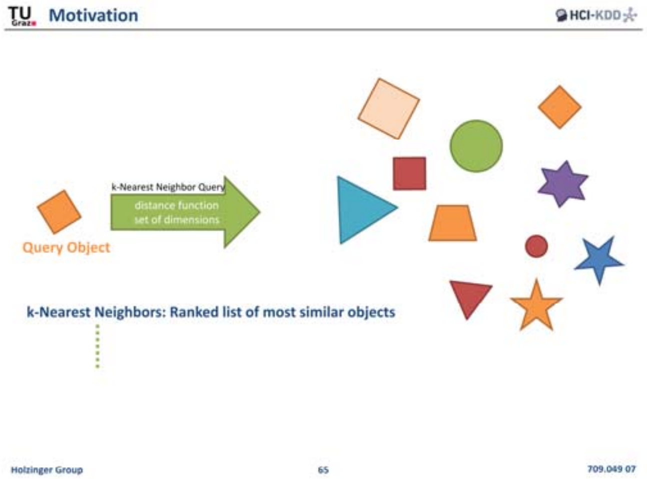
k-Nearest Neighbor Query

distance function
set of dimensions

Query Object

Single Distance Function: $d($◆ , ▶ $) \to R$, based on **fixed** dimensions [shape, color, size, rotation]

Hund, M., Behrisch, M., Färber, I., Sedlmair, M., Schreck, T., Seidl, T. & Keim, D. 2015. Subspace Nearest Neighbor Search-Problem Statement, Approaches, and Discussion. *Similarity Search and Applications.* Springer, pp. 307-313.



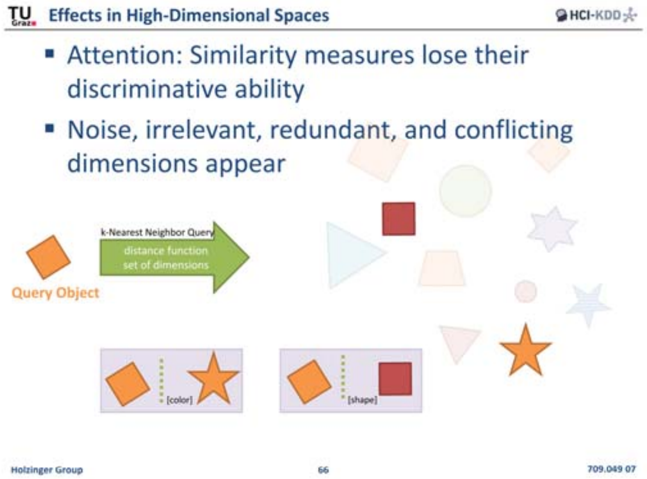
Motivation

k-Nearest Neighbor Query

distance function
set of dimensions

Query Object

k-Nearest Neighbors: Ranked list of most similar objects



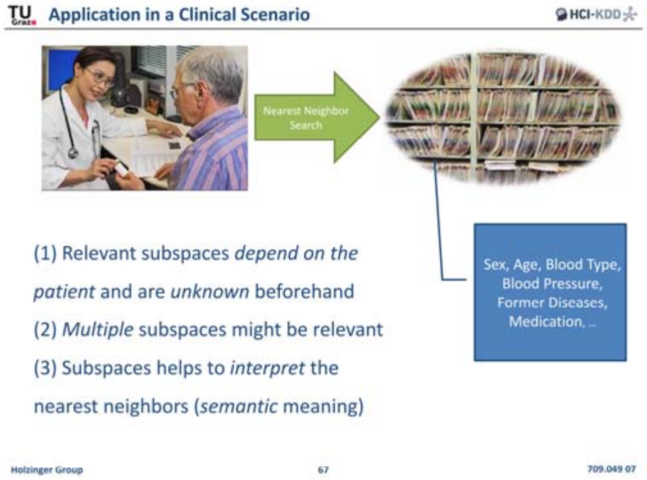
Effects in High-Dimensional Spaces

- Attention: Similarity measures lose their discriminative ability
- Noise, irrelevant, redundant, and conflicting dimensions appear

k-Nearest Neighbor Query

distance function
set of dimensions

Query Object



Application in a Clinical Scenario

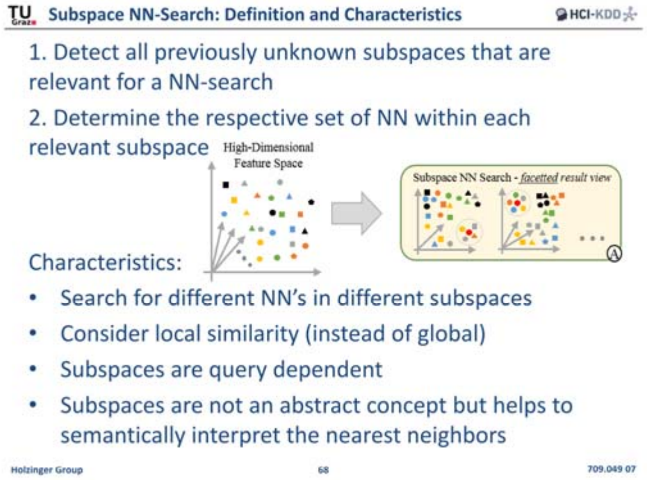
Nearest Neighbor Search

Sex, Age, Blood Type, Blood Pressure, Former Diseases, Medication, …

(1) Relevant subspaces *depend on the patient* and are *unknown* beforehand

(2) *Multiple* subspaces might be relevant

(3) Subspaces helps to *interpret* the nearest neighbors (*semantic* meaning)

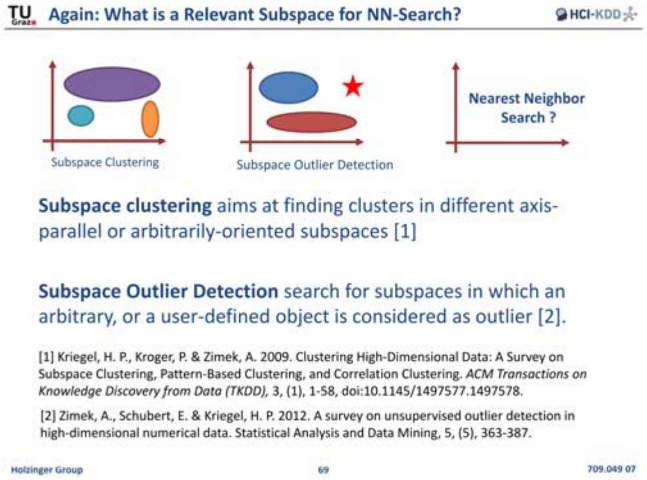


Subspace NN-Search: Definition and Characteristics

1. Detect all previously unknown subspaces that are relevant for a NN-search
2. Determine the respective set of NN within each relevant subspace

Characteristics:

- Search for different NN's in different subspaces
- Consider local similarity (instead of global)
- Subspaces are query dependent
- Subspaces are not an abstract concept but helps to semantically interpret the nearest neighbors



Again: What is a Relevant Subspace for NN-Search?

Subspace Clustering

Subspace Outlier Detection

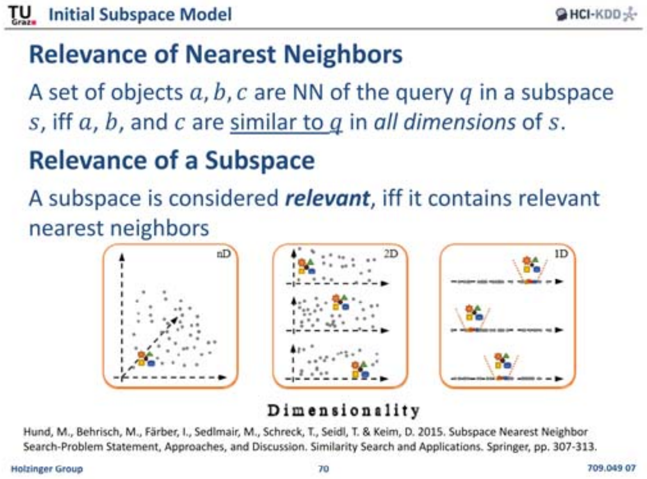
Nearest Neighbor Search ?

Subspace clustering aims at finding clusters in different axis-parallel or arbitrarily-oriented subspaces [1]

Subspace Outlier Detection search for subspaces in which an arbitrary, or a user-defined object is considered as outlier [2].

[1] Kriegel, H. P., Kroger, P. & Zimek, A. 2009. Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering. *ACM Transactions on Knowledge Discovery from Data (TKDD),* 3, (1), 1-58, doi:10.1145/1497577.1497578.

[2] Zimek, A., Schubert, E. & Kriegel, H. P. 2012. A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining, 5, (5), 363-387.



Initial Subspace Model

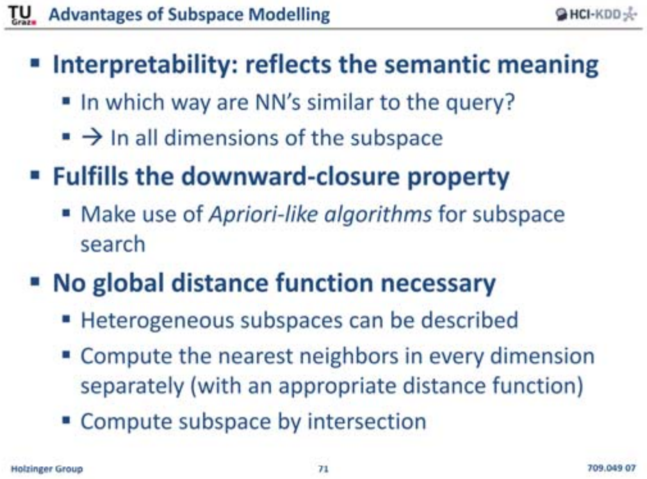
Relevance of Nearest Neighbors

A set of objects a, b, c are NN of the query q in a subspace s, iff a, b, and c are similar to q in *all dimensions* of s.

Relevance of a Subspace

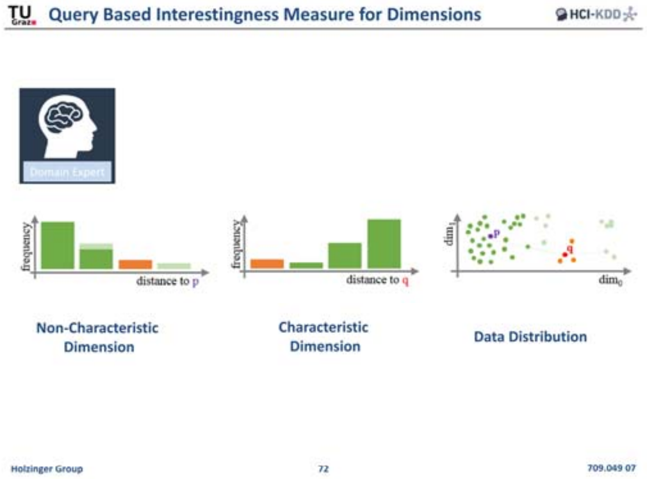
A subspace is considered ***relevant***, iff it contains relevant nearest neighbors

Hund, M., Behrisch, M., Färber, I., Sedlmair, M., Schreck, T., Seidl, T. & Keim, D. 2015. Subspace Nearest Neighbor Search-Problem Statement, Approaches, and Discussion. Similarity Search and Applications. Springer, pp. 307-313.



Advantages of Subspace Modelling

- **Interpretability: reflects the semantic meaning**
  - In which way are NN's similar to the query?
  - → In all dimensions of the subspace
- **Fulfills the downward-closure property**
  - Make use of *Apriori-like algorithms* for subspace search
- **No global distance function necessary**
  - Heterogeneous subspaces can be described
  - Compute the nearest neighbors in every dimension separately (with an appropriate distance function)
  - Compute subspace by intersection



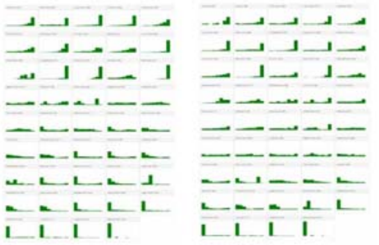
Query Based Interestingness Measure for Dimensions

Non-Characteristic Dimension

Characteristic Dimension

Data Distribution

Query Based Interestingness Measure for Dimensions



query A

query B

Discussion and Open Research Questions

(1)Determine Nearest Neighbors per Dimension

(2) Efficient Search Strategy

(3) Query-Based Interestingness for Dimensions

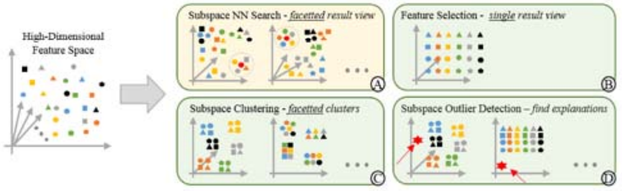
(4) Subspace Quality Criterion (Depends on Analysis Task)

(5) Evaluation Methods and Development of Benchmark Datasets

(6) Multi-input Subspace Nearest Neighbor Search

(7) Visualization and User Interaction

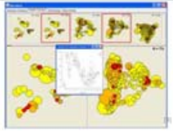
Summary: Subspace Clustering in medical data

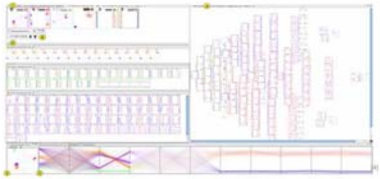

High-Dimensional Feature Space
Subspace NN Search - facetted result view
A
Feature Selection - single result view
B
Subspace Clustering - facetted clusters
C
Subspace Outlier Detection – find explanations
D

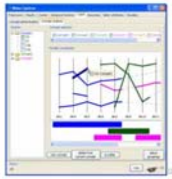
Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L. & Holzinger, A. 2015. Analysis of Patient Groups and Immunization Results Based on Subspace Clustering. In: Guo, Y., Friston, K., Aldo, F., Hill, S. & Peng, H. (eds.) Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250. Cham: Springer International Publishing, pp. 358-368, doi:10.1007/978-3-319-23344-4_35.

Further Subspace Cluster Visualization Techniques

- VISA by Assent et al. (2007)
- CoDa by Günnemann et al (2010)
- Morpheus by Müller et al. (2008)
- Visual Analytics Framework by Tatu et al. (2012)







Visual Analytics for Subspace Steering

- Existing techniques: **exploration** of subspace clusters
- Visualizations to **make sense** of clusters and its subspaces

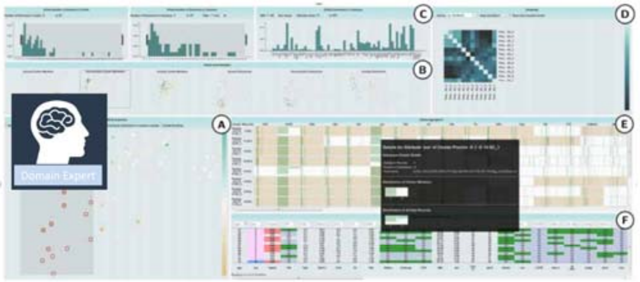
Is the parameter setting appropriate for the data?

What happens if algorithms cannot scale with the #dimensions?

- We need methods to **steer algorithms** while computing relevant subspaces
  - Pruning of intermediate results
  - Adjust parameters to domain knowledge
  - ...

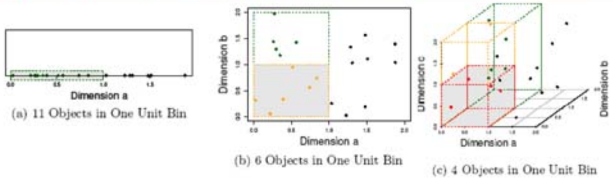
Domain Expert

The doctor-in-the-loop



Hund, M., Boehm, D., Sturm, W., Sedlmair, M., Schreck, T., Ullrich, T., Keim, D. A., Majnaric, L. & Holzinger, A. 2016. Visual analytics for concept exploration in subspaces of patient groups: Making sense of complex datasets with the Doctor-in-the-loop. Brain Informatics, 3, (4), 233-247, doi:10.1007/s40708-016-0043-5.

Always Remember: The curse of dimensionality



(a) 11 Objects in One Unit Bin

(b) 6 Objects in One Unit Bin

(c) 4 Objects in One Unit Bin

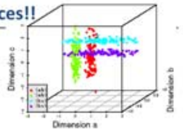
- Data in only one dimension is relatively packed
- Adding a dimension "stretch" the points across that dimension, making them further apart
- Adding more dimensions will make the points further apart—high dimensional data is extremely sparse
- Distance measure becomes meaningless—due to equidistance

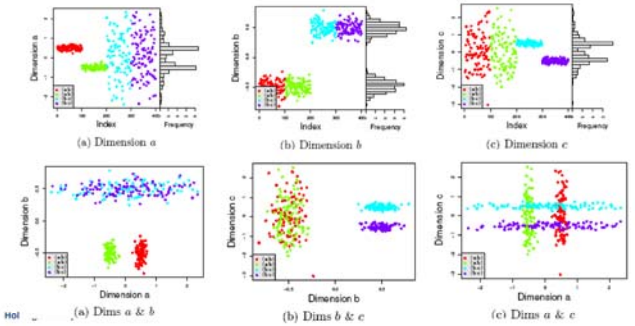
Please remember some definitions

- **Data set** - consists of a matrix of data values, rows represent individual instances and columns represent dimensions.
- **Instance** - refers to a vector of *d* measurements.
- **Cluster** - group of instances in a dataset that are more similar to each other than to other instances. Often, similarity is measured using a distance metric over some or all of the dimensions in the dataset.
- **Subspace** - is a subset of the *d* dimensions of a given dataset.
- **Subspace Clustering** – seek to find clusters in a dataset by selecting the most *relevant* dimensions for each cluster separately .
- **Feature Selection** - process of determining and selecting the dimensions (features) that are most relevant to the data mining task.

Note: Interesting Clusters may ONLY exist in subspaces!!

Parsons, L., Haque, E. & Liu, H. 2004. Subspace clustering for high dimensional data: a review. SIGKDD Explorations 6, (1), 90-105.





(a) Dimension *a*

(b) Dimension *b*

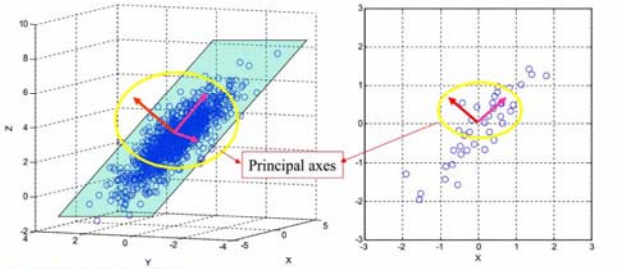
(c) Dimension *c*

(a) Dims *a* & *b*

(b) Dims *b* & *c*

(c) Dims *a* & *c*

Similar concept : Principal Component Analysis (PCA)



- We assume that
- 1) data sets concentrate to a low d-dim. linear subspace
- 2) axes of the subspaces are representations of the data
- 3) identifying the axes can be done by PCA

06 “What is interesting?” Projection Pursuit

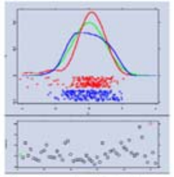
Huber (1985): “What is interesting?”

- **Projection pursuit** : Find a subset of coordinates of the data which display “interesting” features. Often the selection of the subset of coordinates is manual, but there are automated algorithms which can find these subsets automatically also. Finally one has to inspect each projection and decide if its “interesting”.

Huber P.J.: Projection pursuit. *Ann. Statist. 13*, 2 (1985), 435-525.

Interesting = the least Gaussian

- Remember: Gaussian distribution maximizes the entropy!
- Now the objective is to minimize the entropy:
- min $H(t)$ for $t = \omega^T x$
- (i.e. t is normalized)



http://fedc.wiwi.hu-berlin.de/xplore/tutorials/mvahtmlnode115.html

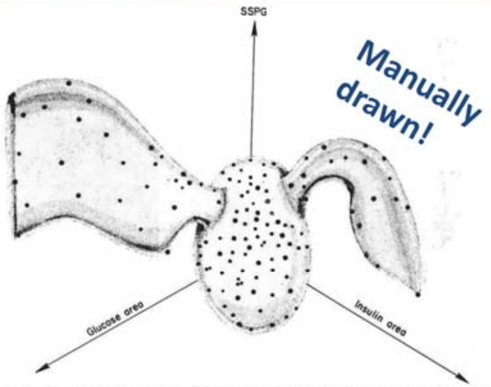
Friedman, J. H. & Tukey, J. W. 1974. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers, 100, (9), 881-890.

Example: Miller-Reavens Diabetes Study from 1979 (1/2)

- 145 diabetes patients
- 6 dimensional data set:
  - 1) age,
  - 2) relative weight,
  - 3) fasting plasma glucose,
  - 4) area under the plasma glucose curve for the three hour glucose tolerance test (OGTT),
  - 5) area under the plasma insulin curve for the OGTT,
  - 6) steady state plasma glucose response.
- Method: Projection Pursuit (PP)
- Result: $\mathbb{R}^6 \longrightarrow \mathbb{R}^3$

Reaven, G. & Miller, R. (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia, 16,* ***1, 17-24.***

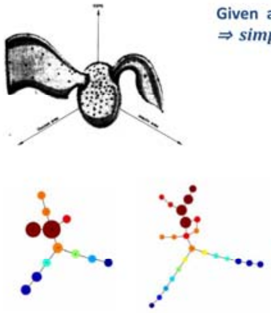
Miller-Reaves Diabetes Study (2/2)



Manually drawn!

Reaven, G. & Miller, R. (1979) An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia, 16,* ***1, 17-24.***

Mapping the data from $\mathbb{R}^6$ to $\mathbb{R}^2$



Given a point cloud data set X and a covering *U*
⇒ *simplicial complex*

$$f: \boldsymbol{X} \to \mathbb{R}$$

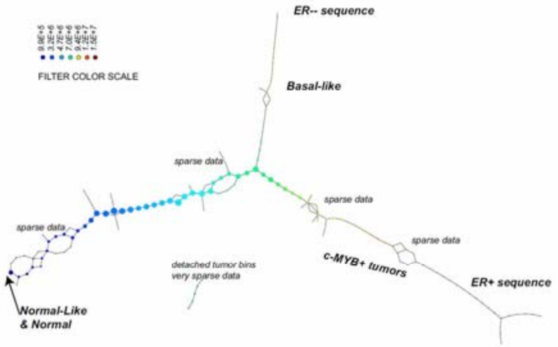
$$f: \boldsymbol{X} \to \boldsymbol{Z}$$

$$\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$$

$$f_\varepsilon(x) = C_\varepsilon \sum_y \exp\left(\frac{-d(x,y)^2}{\varepsilon}\right)$$

Singh, G., Mémoli, F. & Carlsson, G. (2007). *Topological methods for the analysis of high dimensional data sets and 3D object recognition. Eurographics Symposium on Point-Based Graphics, Euro Graphics Society, 91-100.*

Future topic: Topology based data analysis



Nicolau, M., Levine, A. J. & Carlsson, G. (2011) Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences, 108,* ***17, 7265-7270.***

Conclusion and Future Challenges

Slide 8-39 Future Outlook

- Sometimes we have
  - A small number of data sets
  - Rare events – “little data”
  - NP-hard problems (e.g. k-Anonymization, Protein-Folding, Graph Coloring, Subspace Clustering, …)
- Then we still need the “human-in-the-loop”

HCI-KDD

Future Outlook

- Time (e.g. entropy) and Space (e.g. topology)
- Knowledge Discovery from “unstructured” ;-) (Forrester: >80%) data and applications of structured components as methods to index and organize data -> Content Analytics
- Open data, Big data, sometimes: “little data”
- Integration in “real-world” (e.g. Hospital context)
- How can we measure the benefits of visual analysis as compared to traditional methods?
- Can (and how can) we develop powerful visual analytics tools for the non-expert end user?

Thank you!

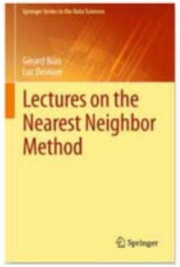
Questions

Sample Questions

- Why would we wish at all to reduce the dimensionality of a data set?
- Why is feature selection so important? What is the difference between feature selection and feature extraction?
- What types of feature selection do you know?
- Can Neural Networks also be used to select features?
- Why do we need a human expert in the loop in subspace clustering?
- What is the advantage of the Projection Pursuit method?
- Why is algorithm selection so critical?

Appendix

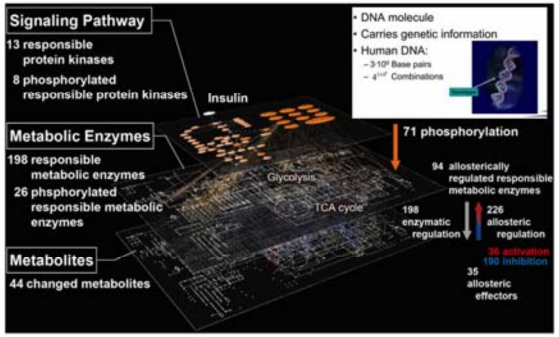
Appendix: Nearest Neighbors


Springer Series in the Data Sciences
Gérard Biau
Luc Devroye
Lectures on the Nearest Neighbor Method
Springer

“Children learn effortlessly by example and exhibit a remarkable capacity of generalization. The field of machine learning, on the other hand, stumbles along clumsily in search of algorithms and methods, but nothing available today comes even close to an average two-year-old toddler … “

Biau, G. & Devroye, L. 2016. Lectures on the nearest neighbor method, Springer, doi:10.1007/978-3-319-25388-6.

Complexity and really BIG DATA

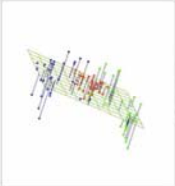

Signaling Pathway
13 responsible protein kinases
8 phosphorylated responsible protein kinases
Insulin
Metabolic Enzymes
198 responsible metabolic enzymes
26 phshorylated responsible metabolic enzymes
Metabolites
44 changed metabolites
Glycolysis
TCA cycle
71 phosphorylation
94 allosterically regulated responsible metabolic enzymes
198 enzymatic regulation
226 allosteric regulation
36 activation
190 inhibition
35 allosteric effectors
• DNA molecule
• Carries genetic information
• Human DNA:
– 3·10⁹ Base pairs
– 4^10⁹ Combinations

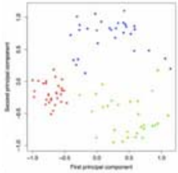
Yugi, K. et al. 2014. Reconstruction of Insulin Signal Flow from Phosphoproteome and Metabolome Data. Cell Reports, 8, (4), 1171-1183, doi:10.1016/j.celrep.2014.07.021.

Dimensionality reduction methods (selected)

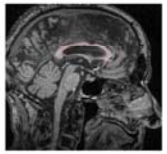
- Linear methods (unsupervised):
  - PCA (Principal Component Analysis)
  - FA (Factor Analysis)
  - MDS (Multi-dimensional Scaling)
- Non-linear methods (unsupervised):
  - Isomap (Isometric feature mapping)
  - LLE (locally linear embedding)
  - Autoencoders
- Supervised methods:
  - LDA (Linear Discriminant Analysis)
- Subspace Clustering with a human-in-the-loop

Example 1: PCA





- Subtract mean from data (center X)
- (Typically) scale each dimension by its variance
  - Helps to pay less attention to magnitude of dimensions
- Compute covariance matrix S $\quad \mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$
- Compute k largest eigenvectors of S
- These eigenvectors are the k principal components

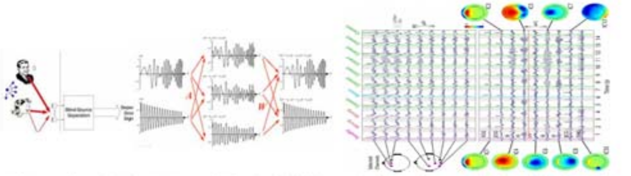


Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, New York, Springer, doi:10.1007/978-0-387-84858-7

Example 2 ICA (Motivation: Blind Source Separation)

- Suppose that there are k unknown independent sources
  $$s(t) = [s_1(t), \dots, s_k(t)]^T \text{ with } Es(t) = 0$$
- A data vector x(*t*) is observed at each time point *t*, such that $\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t)$

where $\mathbf{A}$ is a $n \times k$ full rank scalar matrix



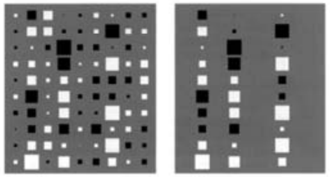
Holzinger, A., Scherer, R., Seeber, M., Wagner, J. & Müller-Putz, G. 2012. Computational Sensemaking on Examples of Knowledge Discovery from Neuroscience Data: Towards Enhancing Stroke Rehabilitation. In: Böhm, C., Khuri, S., Lhotská, L. & Renda, M. (eds.) Information Technology in Bio- and Medical Informatics, Lecture Notes in Computer Science, LNCS 7451. Heidelberg, New York: Springer, pp. 166-168

Example 3: Factor Analysis (FA)

- FA describes variability of observations given unobserved **latent variables = factors.**
- Factors explain correlation between variables
- Similar to PCA, the difference is the conditional probability of the data (ψ = diagonal matrix):

$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{Wz} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$

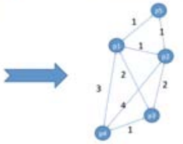
Bishop, C. M. 2006. Pattern Recognition and Machine Learning, Heidelberg, Springer, Chapter 12.2.4



Example 4: MDS: Find projection that best preserves *d*

- Given n x n matrix of pairwise distances between data points
- Compute n x k matrix X with coordinates of distances with some linear algebra magic
- Perform PCA on this matrix X

	p1	p2	p3	p4	p5
p1	0	1	2	3	1
p2	1	0	2	4	1
p3	2	2	0	1	3
p4	3	4	1	0	1
p5	1	1	3	1	0



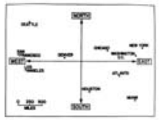
x_i Point in d dimensions

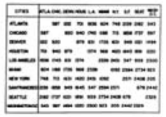
y_i Corresponding point in $r < d$ dimensio

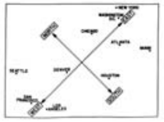
δ_{ij} Distance between x_i and x_j

d_{ij} Distance between y_i and y_j

- Define (e.g.) $\quad E(\mathbf{y}) = \sum_{i,j}\left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}}\right)$
- Find y_i's that minimize E by gradient descent
- Invariant to translations, rotations and scalings







Kruskal, J. B. 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, (1), 1-27.

Example 5: Latent Dirichlet Allocation (LDA)

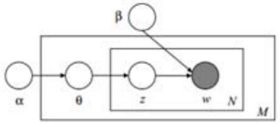
"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

```
initialize $\phi^0_{ni} := 1/k$ for all $i$ and $n$
initialize $\gamma_i := \alpha_i + N/k$ for all $i$
repeat
    for $n = 1$ to $N$
        for $i = 1$ to $k$
            $\phi^{t+1}_{ni} := \beta_{iw_n}\exp(\Psi(\gamma^t_i))$
        normalize $\phi^{t+1}_n$ to sum to 1.
    $\gamma^{t+1} := \alpha + \sum_{n=1}^{N}\phi^{t+1}_n$
until convergence
```

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d$$

Blei, D. M., Ng, A. Y. & Jordan, M. I. 2003. Latent Dirichlet allocation. Journal of Machine Learning Research, 3, (4-5), 993-1022.



Example 6: Isomap

A Global Geometric Framework for Nonlinear Dimensionality Reduction

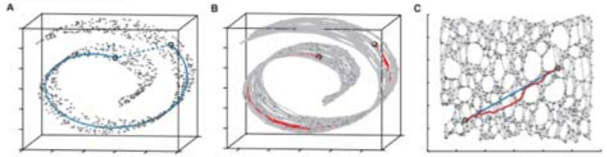
Joshua B. Tenenbaum,[1*] Vin de Silva,[2] John C. Langford[3]

Scientists working with large volumes of high-dimensional data, such as global climate patterns, stellar spectra, or human gene distributions, regularly confront the problem of dimensionality reduction: finding meaningful low-dimensional structures hidden in their high-dimensional observations. The human brain confronts the same problem in everyday perception, extracting from its high-dimensional sensory inputs—30,000 auditory nerve fibers or 10^6 optic nerve fibers—a manageably small number of perceptually relevant features. Here we describe an approach to solving dimensionality reduction problems that uses easily measured local metric information to learn the underlying global geometry of a data set. Unlike classical techniques such as principal component analysis (PCA) and multidimensional scaling (MDS), our approach is capable of discovering the nonlinear degrees of freedom that underlie complex natural observations, such as human handwriting or images of a face under different viewing conditions. In contrast to previous algorithms for nonlinear dimensionality reduction, ours efficiently computes a globally optimal solution, and, for an important class of data manifolds, is guaranteed to converge asymptotically to the true structure.

Goal: Find projection onto *nonlinear* manifold

1. Construct neighborhood graph G:
   For all x_i, x_j
   If distance$(x_i, x_j) < \epsilon$
   Then add edge (x_i, x_j) to G
2. Compute shortest distances along graph $\delta_G(x_i, x_j)$ (e.g., by Floyd's algorithm)
3. Apply multidimensional scaling to $\delta_G(x_i, x_j)$

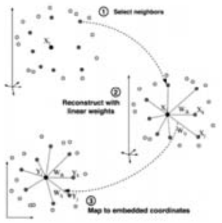
http://isomap.stanford.edu/

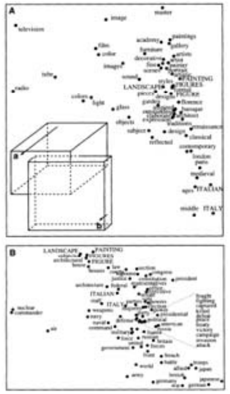


Tenenbaum, J. B., De Silva, V. & Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. Science, 290, (5500), 2319-2323, doi:10.1126/science.290.5500.2319.

Example 7: Locally Linear Embedding (LLE)

Roweis, S. T. & Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. Science, 290, (5500), 2323-2326, doi:10.1126/science.290.5500.2323.

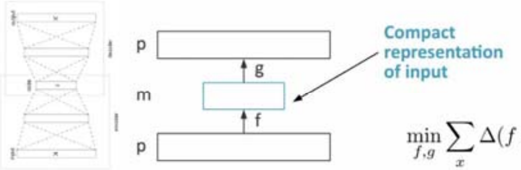

① Select neighbors
② Reconstruct with linear weights
③ Map to embedded coordinates



$$\varepsilon(W) = \sum_i \left|\vec{X}_i - \Sigma_j W_{ij}\vec{X}_j\right|^2$$

$$\Phi(Y) = \sum_i \left|\vec{Y}_i - \Sigma_j W_{ij}\vec{Y}_j\right|^2$$

Example 8: Autoencoders



$$\min_{f,g}\sum_x \Delta(f \circ g, x)$$

- History: Dim-reduction with NN: Learning representations by back-propagating errors
- Goal: output matches input

Rumelhart, D. A., Hinton, G. E. & Williams, R. J. 1986. Learning representations by back-propagating errors. Nature, 323, 533-536.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. The Journal of Machine Learning Research, 11, 3371-3408.

Autoencoders are "old"

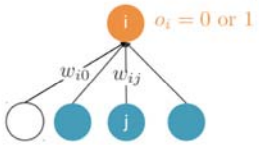
- **Sigmoidal neurons and backpropagation:** Rumelhart*), D. A., Hinton, G. E. & Williams, R. J. 1986. Learning representations by back-propagating errors. Nature, 323, 533-536.
  $$\Delta(y, x) = ||y - x||_2^2$$
- **Linear autoencoders:** Baldi, P. & Hornik, K. 1989. Neural networks and principal component analysis: Learning from examples without local minima. Neural networks, 2, (1), 53-58.
  $$\min_{A,B}\sum_x ||ABx - x||_2^2$$

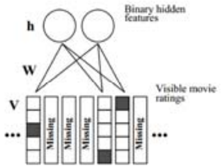
*) David Rumelhart (1942-2011) was Cognitive Scientist working on math. Psychology

Autoencoders -> Restricted Boltzmann Machines

- Based on Information processing in dynamical systems: Foundations of harmony theory by Smolensky (1986): Stochastic neural networks where the unit activation i = probabilistic

$$Pr(o_i = 1) = \frac{1}{1 + e^{-w_{i0} + \sum_j o_j w_{ij}}}$$





Right: A restricted Boltzmann machine with binaryhidden units and softmax visible units

Salakhutdinov, R., Mnih, A. & Hinton, G. (2007) Restricted Boltzmann machines for collaborative filtering. ICML, 791-798.

Summary

- Goal: Having m < p features
- Feature selection via
  - A) Filter approaches
  - B) Wrapper approaches
  - C) Embedded approaches (Lasso, Electric net, see Tibshirani, Hastie …)
- Feature extraction
  - A) Linear: e.g. PCA
  - B) Non-linear: Autoencoders (map the input to the output via a smaller layer)