Andreas Holzinger

185.A83 Machine Learning for Health Informatics

2017S, VU, 2.0 h, 3.0 ECTS

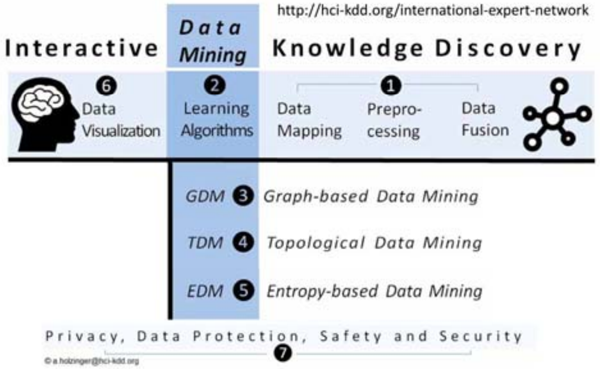
Lecture 07 - Module 05 – Week 18 - 02.05.2017

Dimensionality Reduction and Subspace Clustering: Example for the Doctor-in-the-Loop

a.holzinger@hci-kdd.org

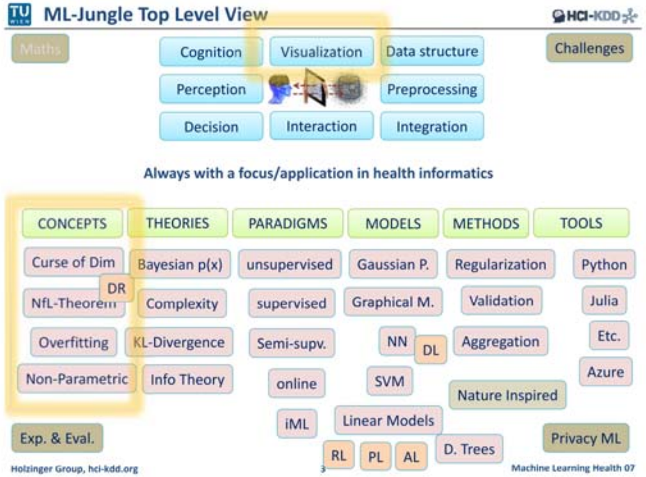
http://hci-kdd.org/machine-learning-for-health-informatics-course

ML needs a concerted effort fostering integrated research



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

ML-Jungle Top Level View

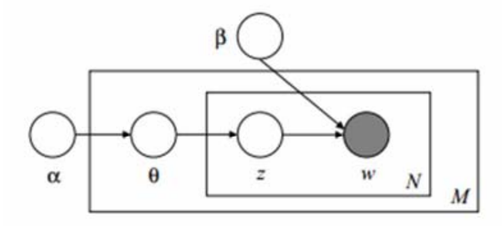


Red thread through this lecture

- 01 Classification vs Clustering
- 02 Feature spaces, feature engineering
  - Feature selection, feature extraction
- 03 The curse of dimensionality
- 04 Dimensionality reduction
  - PCA, ICA, FA, MDS, LDA – Isomap, LLE, Autoencoder
- 05 Subspace clustering and analysis
- 06 Projection Pursuit: "What is interesting?"

LDA = ?

- Latent Dirichlet Allocation
- LDA = linear discriminant analysis (Attention!)

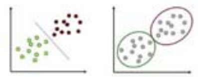


01 Classification vs. Clustering

Key Challenges

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

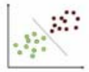
Classification (A) vs. Clustering (C) – Intro Quiz



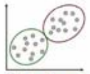
1) The data is not labeled (A/C)?
2) Identify structure/patterns (A/C)?
3) Predicting an item set, identifying to which set of categories a new observation belongs (A/C)?
4) Assigning a set of objects into groups (A/C)?
5) Having many labelled data points (A/C)
6) Using the concept of supervised learning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

Classification vs. Clustering on one slide

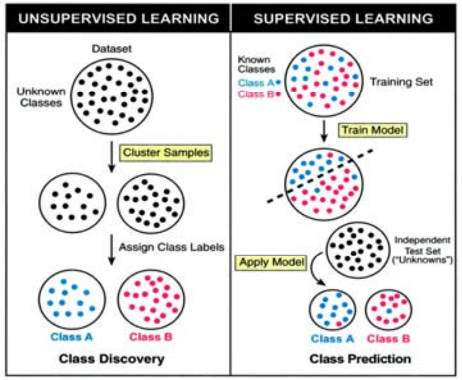
- **Classification (Supervised learning, Pattern Recogn., Prediction)**
  - Supervision = the training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations;
  - New data is **classified based on the training set**
  - Important for clinical decision 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;

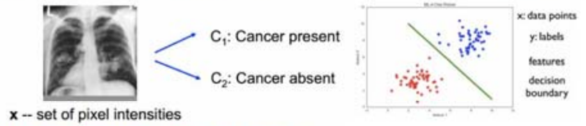


Discovery vs. Prediction



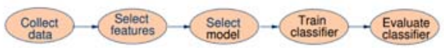
Ramaswamy, S. & Golub, T. R. (2002) DNA Microarrays in Clinical Oncology. *Journal of Clinical Oncology, 20, 7, 1932-1941.*

Why do we need Classification?



- 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

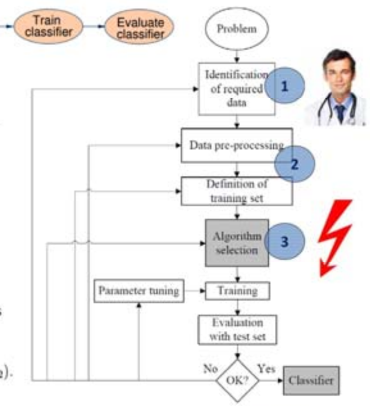
Learning Process: Algorithm selection is crucial



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 $x_2 \Rightarrow y_1 = y_2$

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

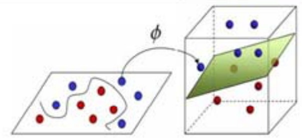
- SVM – a plane/hyperplane separates two classes of data – very versatile for classification and clustering – also via the Kernel trick in high-dimensions

```
Input: (x_1, y_1), ..., (x_n, y_n), C, ε
S_i ← ∅ for all i = 1, ..., n
repeat
  for i = 1, ..., n do
    H(y) ≡ Δ(y_i, y) + w^T Ψ(x_i, y) − w^T Ψ(x_i, y_i)
    compute ŷ = argmax_{y∈Y} H(y)
    compute ξ_i = max{0, max_{y∈S_i} H(y)}
    if H(ŷ) > ξ_i + ε then
      S_i ← S_i ∪ {ŷ}
      w ← optimize primal over S = ∪_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.

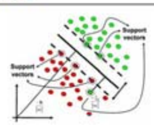
SVM – Vapnik, 1992

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

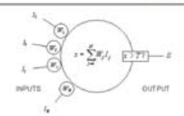


- = classification method for both linear and nonlinear data;
- Within the new dimension, it searches for the linear optimal separating **hyperplane** (i.e., "decision boundary");
- By nonlinear mapping to a sufficiently high dimension, data from two classes can always be separated with a hyperplane;
- 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

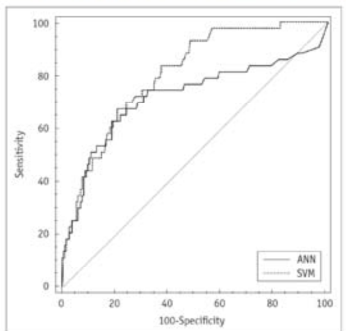


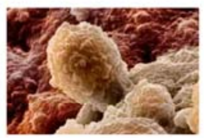
- **SVM**
  - Deterministic algorithm
  - Nice generalization properties
  - Hard to learn – learned in batch mode using quadratic programming techniques
  - Using kernels can learn very complex functions



- **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)

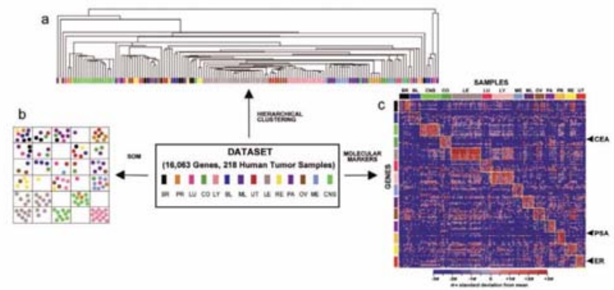
Clinical use: SVM are more accurate than ANN





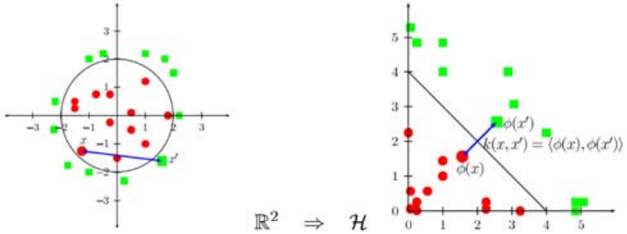
Kim, S. Y., Moon, S. K., Jung, D. C., Hwang, S. I., Sung, C. K., Cho, J. Y., Kim, S. H., Lee, J. & Lee, H. J. (2011) Pre-Operative Prediction of Advanced Prostatic Cancer Using Clinical Decision Support Systems: Accuracy Comparison between Support Vector Machine and Artificial Neural Network. *Korean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

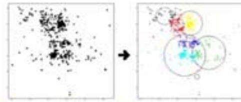
Summary: The 10 top algorithms - Quiz

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));
- **Apriori**
  - 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));*

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 data set into k clusters so that intra-cluster variance is a minimum
  - V ... variance (objective function)
  - S_i ... cluster
  - Y_i ... mean
  - D ... set of all points xj
  - k ... number of clusters

$$V(D) = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

Example

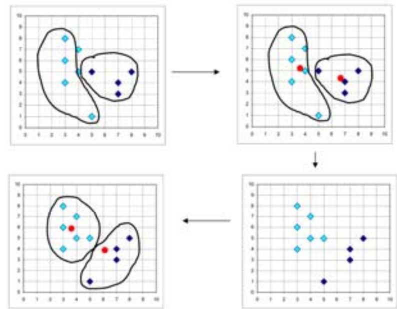
Algorithm 1: Example for a classical weight balanced k-means algorithm

Input: $d, k, n \in \mathbb{N}$, $X := \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $S := \{s_1, \ldots, s_k\} \subset \mathbb{R}^d$
Output: Clustering $C = (C_1, \ldots, C_k)$ of X and the arithmetic means $c_1, \ldots, c_k$ as sites

1. Partition X into a clustering $C = (C_1, \ldots, C_k)$ 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



Advantages of k-Means?

- What is the computational time of k-means?
- NP-hard in Euclidean space, however, if k and d can be fixed than it can be solved within:

$$\mathcal{O}(npkt)$$

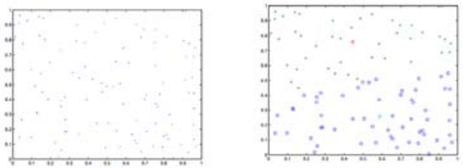
compute kn distances in p dimensions

number of iterations

Can be small if there's indeed a cluster structure in the data

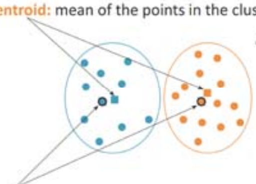
Jain, A. K. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters,* 31, (8), 651-666.

The Basics: Centroid and Medoid



- **Centroid:** mean of the points in the cluster.

$$\mu = \frac{1}{|C|} \sum_{x \in C} x$$



- **Medoid:** point in the cluster that is closest to the centroid.

$$m = \arg\min_{x \in C} d(x, \mu)$$

02 Feature Engineering

"Applied ML is basically feature engineering.
Andrew Yan-Tak Ng"

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
  - $d\colon D \times D \to \mathbb{R}_0^+$ ... a total distance function; true for
    - $\forall p,q \in \mathcal{D}, p \neq q: d(p,q) > 0$ (strict)
    - and must be reflexive and symmetric

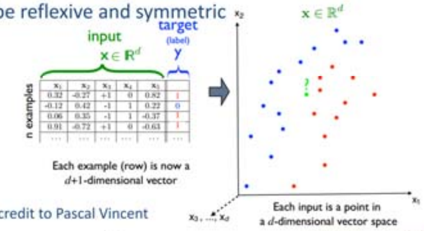


Image credit to Pascal Vincent

Digression: Metric Space (e.g. Euclidean Vector Space)

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

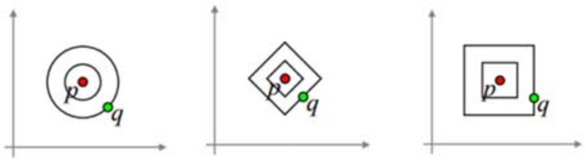
1. For all $x, y, z \in X$, $d(x, y) \leq d(x, z) + d(z, y)$.
2. For all $x, y \in X$, $d(x, y) = d(y, x)$.
3. $d(x, y) = 0$ if and only if $x = y$.

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

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

Let do a Quiz again: Similarities of feature vectors

Look at the examples below, which distance measures would you select?



Euclidian norm Manhattan norm Maximums norm

Feature Selection: Overview

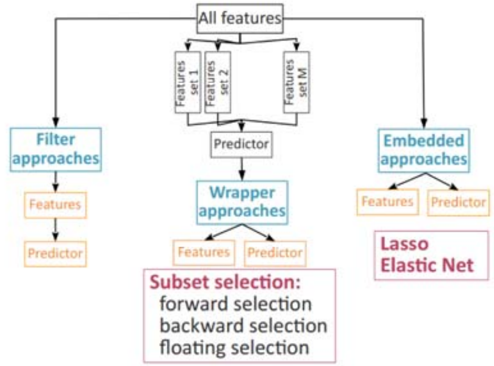
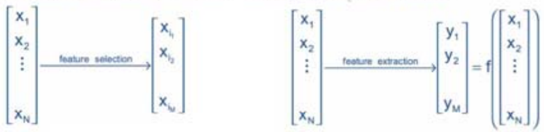


Image credit to Chloe Azencott

Feature Selection vs. Feature Extraction

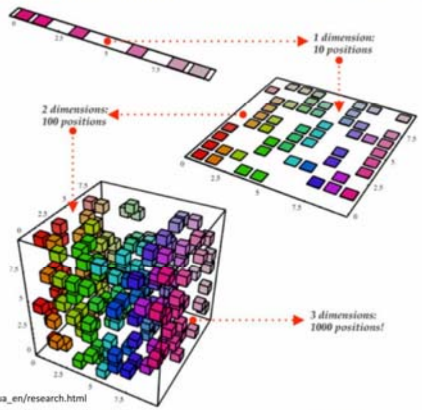
- Feature selection is just selecting a subset of the existing features without any transformation
- Feature extraction is *transforming* existing features into a lower dimensional space



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

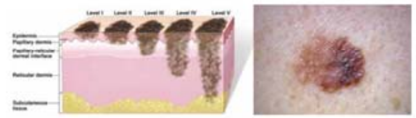


Bengio, S. & Bengio, Y. 2000. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Transactions on Neural Networks, 11, (3), 550-557.

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

Examples for High-Dimensional Biomedical Data

- Medical Image Data (16 - 1000+ features)

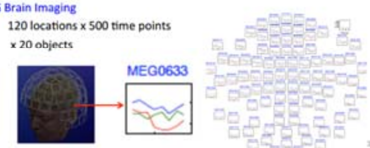


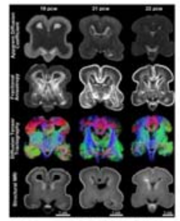
http://qsota.com/melanoma/

MEG Brain Imaging

120 locations x 500 time points

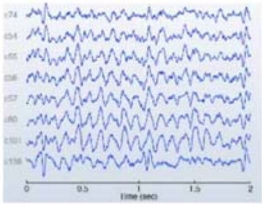
x 20 objects

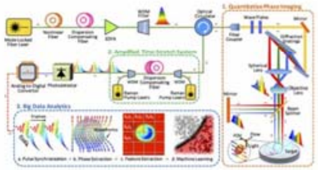




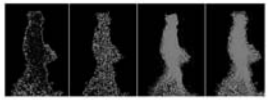
Nature 508, 199–206 doi:10.1038/nature13185

- Biomedical Signal Data (10 - 1000+ features)

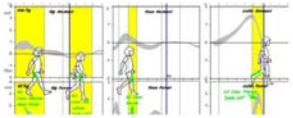




http://www.nature.com/articles/srep21471#f1

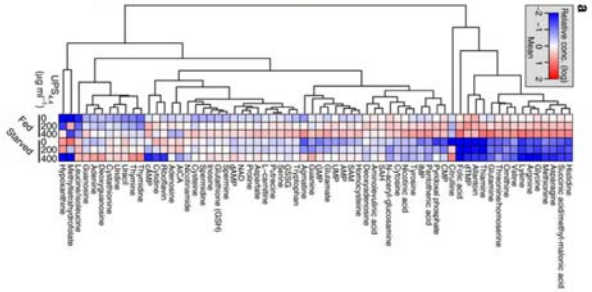


http://www.mdpi.com/1424-8220/14/4/6124/htm



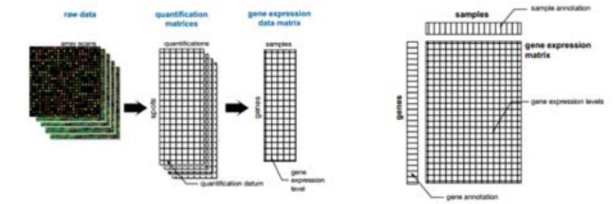
http://www.clinicalgaitanalysis.com/data/

- 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

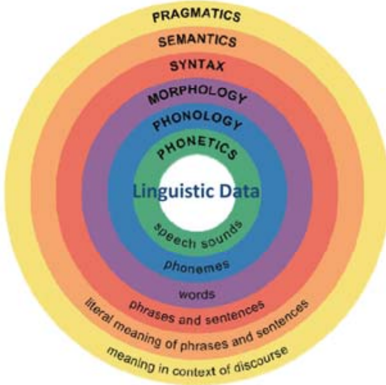
Microarray Data (features correspond to genes, up to 30k features)



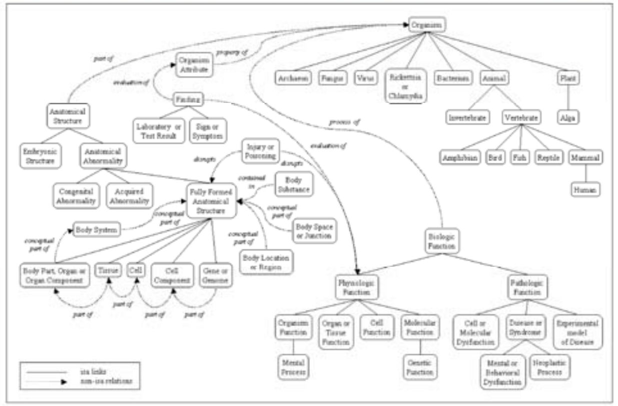
Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A. & Causton, H. C. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nature genetics, 29, (4), 365-371.

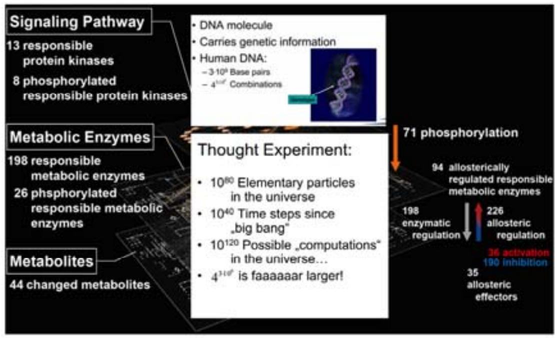
- Text > 10^9 documents × 10^6 words/n-grams features correspond to words or terms, between 5k to 20k features
- Text (Natural Language) is definitely very important for health:
  - Handwritten Notes, Drawings
  - Patient consent forms
  - Patient reports
  - Radiology reports
  - Voice dictations, annotations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

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



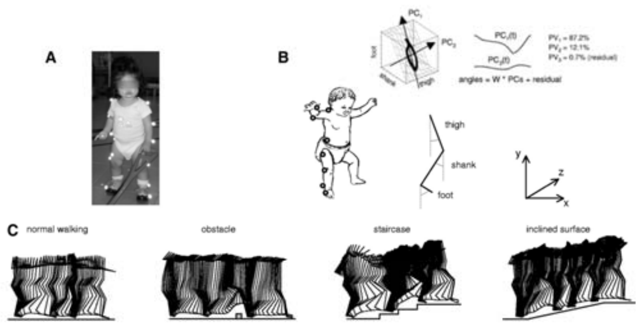
Thomas, J. J. & Cook, K. A. 2005. *Illuminating the path: The research and development agenda for visual analytics, New York, IEEE Computer Society Press.*





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.

- Hyperspace is large – all points are far apart
- Computationally challenging (in time and 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
- 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



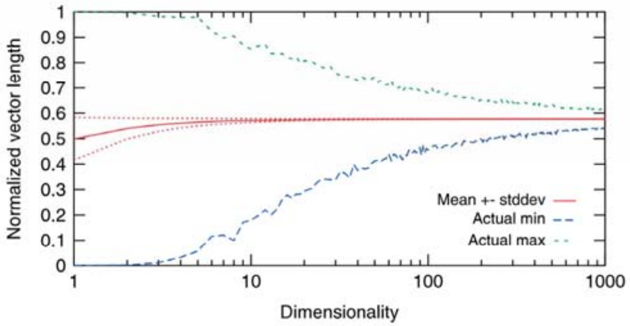
Dominici, N., Ivanenko, Y. P., Cappellini, G., Zampagni, M. L. & Lacquaniti, F. 2010. Kinematic Strategies in Newly Walking Toddlers Stepping Over Different Support Surfaces. Journal of Neurophysiology, 103, (3), 1673-1684, doi:10.1152/jn.00945.2009.

Problems of the CoD in relation to subspace clustering

- Aspect 1: Optimization Problem
- Aspect 2: Concentration Effect
- Aspect 3: Irrelevant Attributes
- Aspect 4: Correlated Attributes

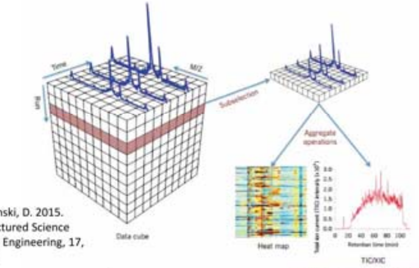
Kriegel, H. P., Kröger, P. & Zimek, A. 2012. Subspace clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2, (4), 351-364, doi:10.1002/widm.1057.

Curse of Dimensionality



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, doi:10.1002/sam.11161.

Example: Neonatal Screening (2/3)



Yao, Y., Bowen, B. P., Baron, D. & Poznanski, D. 2015. SciDB for High-Performance Array-Structured Science Data at NERSC. Computing in Science & Engineering, 17, (3), 44-52, doi:10.1109/MCSE.2015.43.

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.*

Challenge

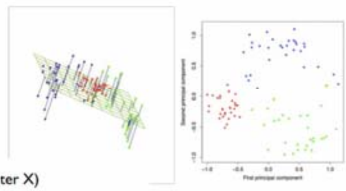
- Given n data points in d dimensions
- Conversion to m data points in $r < d$ dimensions
- Challenge: **minimal loss of information *)**

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

Approaches

- Linear methods (unsupervised):
  - PCA
  - FA
  - MDS
- Supervised methods:
  - LDA
- Non-linear methods (unsupervised):
  - Isomap (Isometric feature mapping)
  - LLE (locally linear embedding)
  - Autoencoders

Example 1: PCA

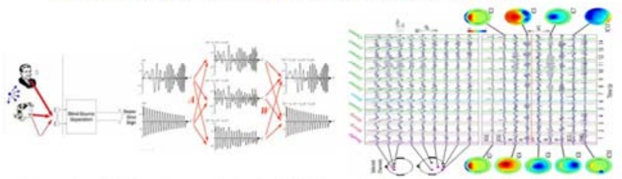


- Subtract mean from data (center X)
- (Typically) scale each dimension by its variance
  - Helps to pay less attention to magnitude of dimensions
- Compute covariance matrix S $\quad \mathbf{S} = \frac{1}{N}\mathbf{X}^\mathsf{T}\mathbf{X}$
- Compute k largest eigenvectors of S
- These eigenvectors are 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
  $\mathrm{s}(t) = [s_1(t), \ldots, s_k(t)]^T$ with $E\mathrm{s}(t) = 0$
- A data vector x(*t*) is observed at each time point *t*, such that $\mathrm{x}(t) = \mathbf{A}\,\mathrm{s}(t)$
  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: FA

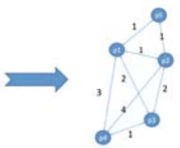
- Factor Analysis describes the variability of observations in terms of unobserved latent variables (these are called "factors") and noise
  - The factors explain the correlation between the var
  - Variance can be explained by Gaussian noise (and can be calculated)
  - Advantage: generative approach and models BOTH the noise of the observations and their correlation!
  - You can make assumptions on the distributions of noise and factors

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

–Find a set of points whose pairwise distances match a given distance matrix

- 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

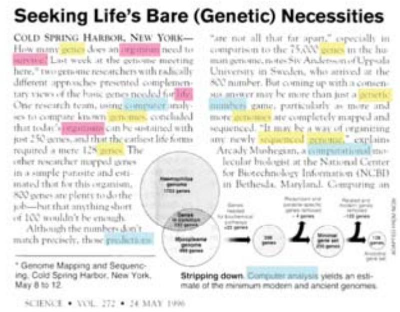
y_i Corresponding point in $r < d$ dimensions

δ_{ij} Distance between x_i and x_j

d_{ij} Distance between y_i and y_j

- Define (e.g.) $E(\mathbf{y}) = \sum_{i,j} \left(\frac{d_{ij} - \delta_{ij}}{\delta_{ij}} \right)^2$
- Find y_i's that minimize E by gradient descent
- Invariant to translations, rotations and scalings

Example 5: LDA


Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Example 6: Isomap

A Global Geometric Framework for Nonlinear Dimensionality Reduction

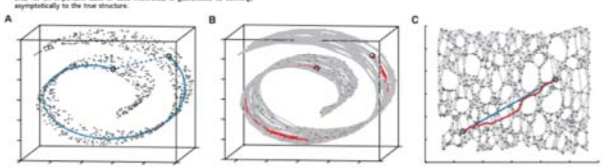
Joshua B. Tenenbaum,[1*] Vin de Silva,[2] John C. Langford[3]

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

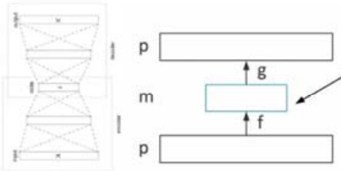
1. Construct neighborhood graph G:
   For all x_i, x_j
   If distance$(x_i, x_j) < \epsilon$
   Then add edge (x_i, x_j) to G
2. Compute shortest distances along graph $\delta_G(x_i, x_j)$ (e.g., by Floyd's algorithm)
3. Apply multidimensional scaling to $\delta_G(x_i, x_j)$

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 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.

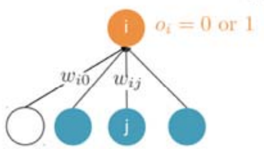
$$\min_{A,B} \sum_x ||ABx - x||_2^2$$

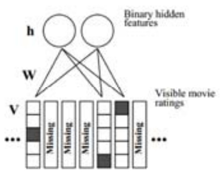
*) David Rumelhart (1942-2011) was Cognitive Scientist working on math. Psychology

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)

05 Subspace Clustering* & Subspace Analysis

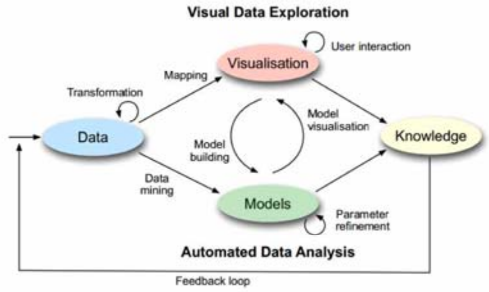
* Two major issues
(1) the algorithmic approach to clustering and
(2) the definition and assessment of **similarity versus dissimilarity**.

What is subspace clustering?

- K clusters
- N data points
- *D* dimensions (original space)
- *d* dimensions (latent subspace)
- SC = 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.

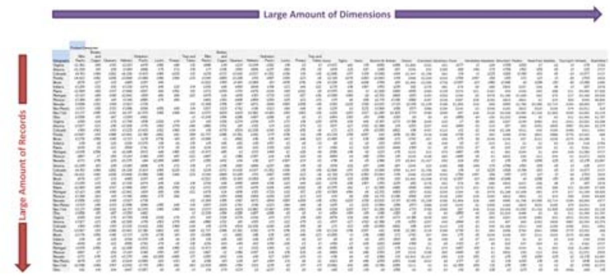
Visual Analytics Pipeline



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



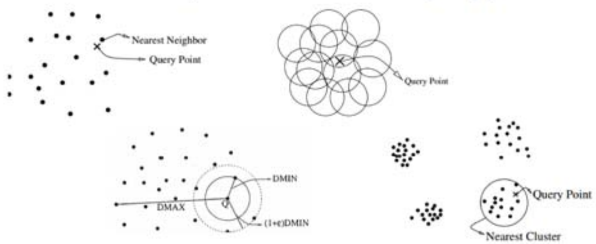
High-Dimensional Data – The Curse of Dimensionality

- Irrelevant Dimensions
- Correlated and Redundant Dimensions
- Conflicting Dimensions
- Challenging Interpretation of data and 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.

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.

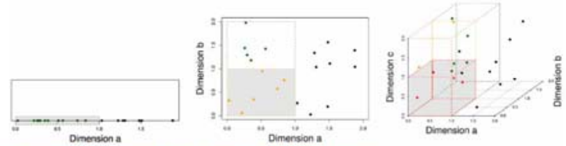


Beyer, K., Goldstein, J., Ramakrishnan, R. & Shaft, U. 1999. When is "nearest neighbor" meaningful? *In:* Beeri, C. & Buneman, P. (eds.) *Database Theory ICDT 99, LNCS 1540.* Berlin: Springer, pp. 217-235.

Challenges in High-Dim Data – Curse of Dimensionality

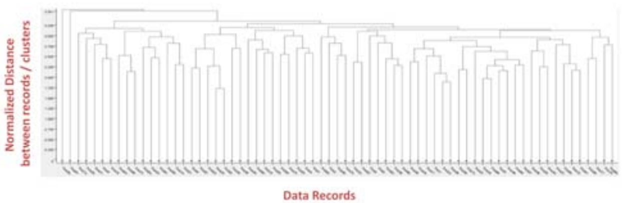
- Concentration Effect
  - Discriminability of similarity gets lost
  - Impact on usefulness of a similarity measure

- **High-Dimensional Data is Sparse**

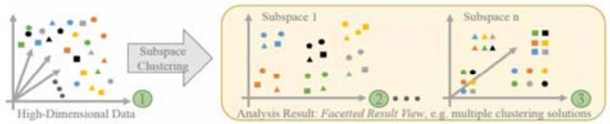


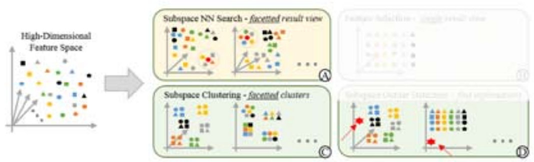
Optimization Problem and Combinatorial Issues
Feature selection and dimension reduction
2^d-1 possible subsets of dimensions (-> subspaces)

Example: Full Space Clustering of High-Dimensional Data



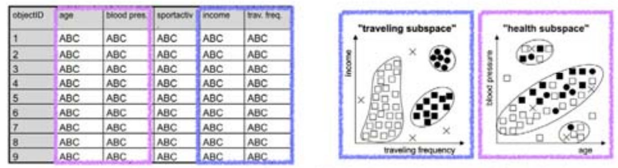
Overview of (major?) Subspace Analysis Techniques





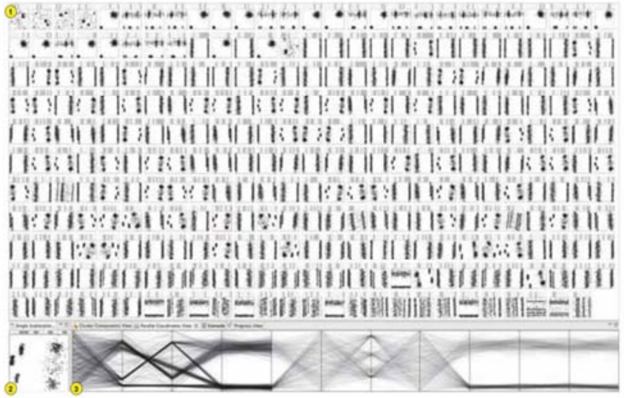
- **Patterns may be found in subspaces (dimension combinations)**
- **Patterns may be complementary or redundant to each other**

Subspace Concept

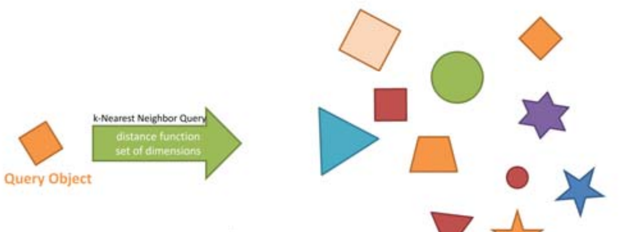


Tatu, A., Maass, F., Faerber, I., Bertini, E., Schreck, T., Seidl, T. & Keim, D. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. IEEE Symposium onVisual Analytics Science and Technology (VAST), 2012 Seattle. IEEE, 63-72, doi:10.1109/VAST.2012.6400488.

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



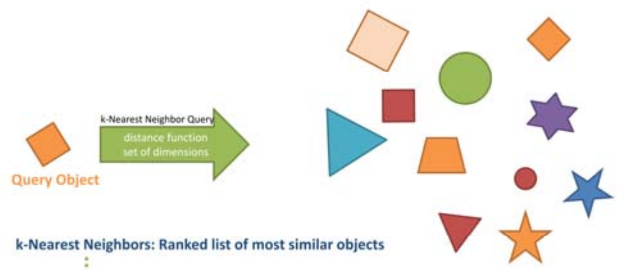
Motivation



Single Distance Function: $d($◆, ▶$) \rightarrow R$, 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.

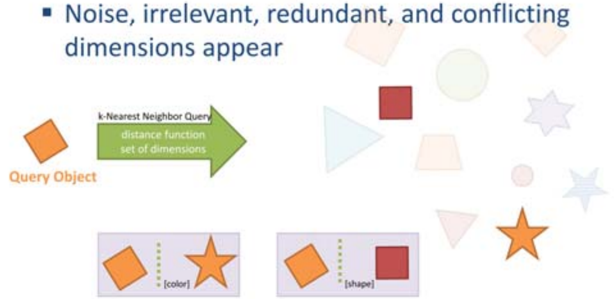
Motivation



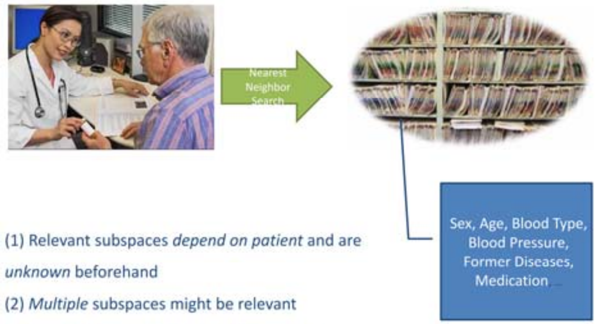
k-Nearest Neighbors: Ranked list of most similar objects

Effects in High-Dimensional Spaces

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



Application in a Clinical Scenario



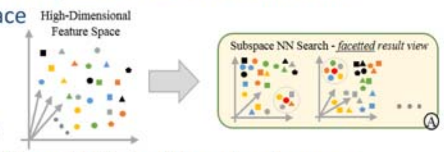
(1) Relevant subspaces *depend on patient* and are *unknown* beforehand

(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 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.

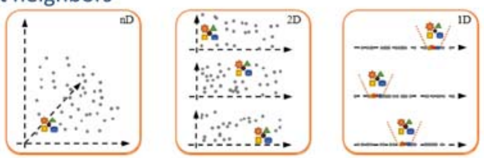
Initial Subspace Model

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



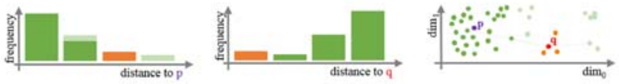
Dimensionality

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

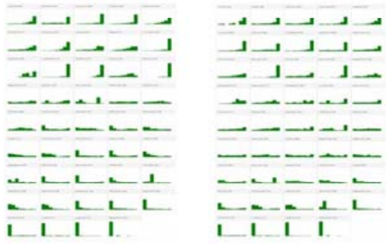


Non-Characteristic Dimension

Characteristic Dimension

Data Distribution

Query Based Interestingness Measure for Dimensions



query = butter

query = gauda cheese

Initial experiments

Supplementary Material

- http://files.dbvis.de/sisap2015

Dataset

- USDA National Nutrition Database
- http://ndb.nal.usda.gov/

Experiment

- Full Space (Eucl. distance, 50 dim.)
- Subspaces (our model)

Full Space	Subspace 1	Subspace 2
butter, whipped	butter, whipped	butter, whipped
butter, without salt	butter oil, anhydrous	butter, without salt
butter oil, anhydrous	butter, without salt	salad drsng, mayo
kellogg's, fruit bars	lard	margarine
margarine	salad drsng, mayo	chicken, broilers
pancakes	oil, soybn	pork, backfat
waffle	oil, cocnt	candies, butterscotch
cream	oil, olive	candies, hard
cheese, cream	oil, safflower	candies, jellybeans
pie crust	vegetable oil, palm kernel	candies, mars snackfood
cheese, mozzarella	oil, canola	chewing gum
kellogg's cereals	oil, sunflower	puddings, vanilla
soup	margarine	jellies
cheese, limburger	shortening	sweeteners, tabletop
peppers	chicken, broilers	syrups, corn
sauce tabasco	oil, corn, peanut, and olive	syrups, maple

Discussion and Open Research Questions

(1)Determine Nearest Neighbors per Dimension

(2) Efficient Search Strategy

(3) Query-Based Interestingness for Dimensions

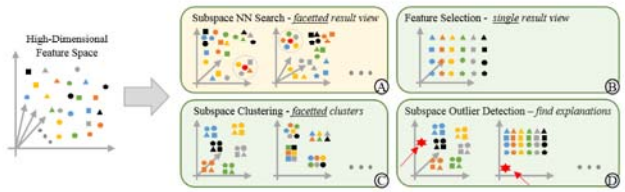
(4) Subspace Quality Criterion (Depends on Analysis Task)

(5) Evaluation Methods and Development of Benchmark Datasets

(6) Multi-input Subspace Nearest Neighbor Search

(7) Visualization and User Interaction

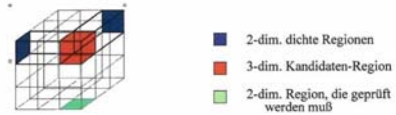
Subspace Clustering



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.

Projected Clustering / Subspace Clustering / Alternative

- Variety of different algorithms, e.g. PROCLUS [1], CLIQUE [2], RESCUE [3]
- Example CLIQUE:



- Challenges
- Exponential # of possible subspaces
- Result highly depend on parameters
- Highly redundant results (clusters + subspaces)

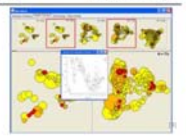
Example Clust Nails Tatu et al (2012)

Which dimensions occur more often in clusters?
Which occur often together?
Which values do records in a specific cluster have?

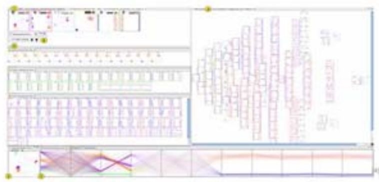
Tatu, A., Albuquerque, G., Eisemann, M., Schneidewind, J., Theisel, H., Magnor, M. & Keim, D. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on, 2009. IEEE, 59-66.

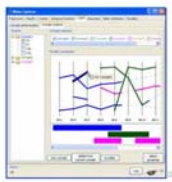
Tatu, A., Maass, F., Faerber, I., Bertini, E., Schreck, T., Seidl, T. & Keim, D. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. IEEE Symposium onVisual Analytics Science and Technology (VAST), 2012 Seattle. IEEE, 63-72, doi:10.1109/VAST.2012.6400488.

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), see before





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?

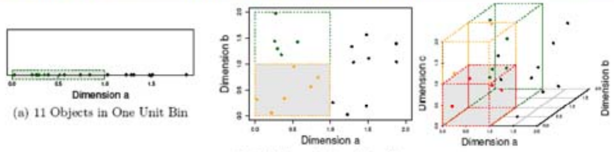
- We need methods to **steer algorithms** while computing relevant subspaces
  - Pruning of intermediate results
  - Adjust parameters to domain knowledge
  - ...

Domain Expert

Fig. 3 A screenshot of our visual analytics tool SubVIS. It enables the user to interactively explore a large number of subspace clusters. A general overview of the similarities between the subspaces is given by an MDS projection (A). Small multiples (B) allows to preview projections of different distance functions and a quick change of the MDS plot. On the very top (C) the user is provided with some distribution properties of the subspaces such as the #dimensions. A heatmap (D) provides more details of relationships between the pair-wise distances. An aggregation table (E) shows the values of the aggregated cluster members and the table lense (F) provides details on demand.

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.

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

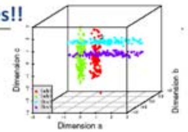
Data in only one dimension is relatively packed
Adding a dimension "stretch" the points across that dimension, making them further apart
Adding more dimensions will make the points further apart—high dimensional data is extremely sparse
Distance measure becomes meaningless—due to equidistance

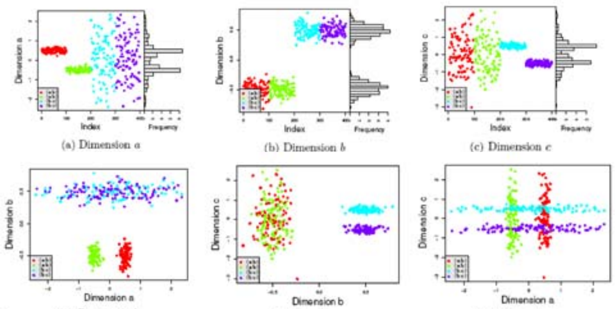
Repeat some definitions

- Dataset - 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.

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.

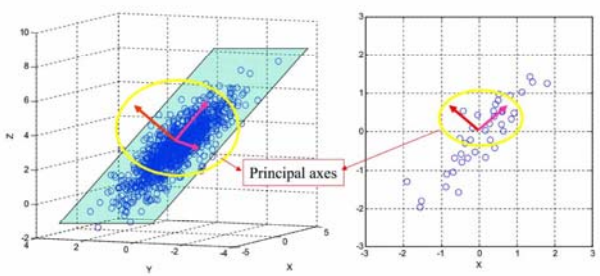




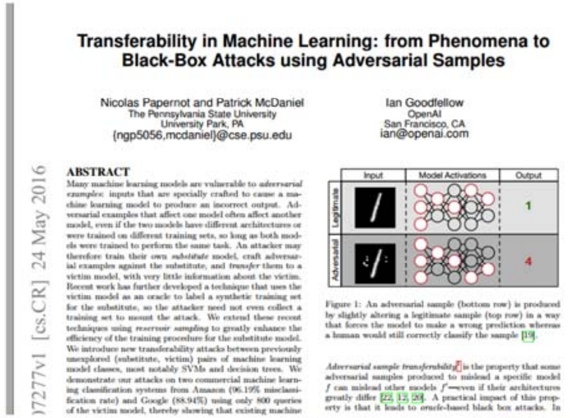
(a) Dimension *a* (b) Dimension *b* (c) Dimension *c*

(a) Dims *a* & *b* (b) Dims *b* & *c* (c) Dims *a* & *c*

Similar concept : Principal Component Analysis (PCA)



Black-Box approach



Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Nicolas Papernot and Patrick McDaniel
The Pennsylvania State University
University Park, PA
{ngp5056,mcdaniel}@cse.psu.edu

Ian Goodfellow
OpenAI
San Francisco, CA
ian@openai.com

[cs.CR] 24 May 2016

ABSTRACT

Many machine learning models are vulnerable to *adversarial examples*: inputs that are specially crafted to cause a machine learning model to produce an incorrect output. Adversarial examples that affect one model often affect another model, even if the two models have different architectures or were trained on different training sets, so long as both models were trained to perform the same task. An attacker may therefore train their own *substitute* model, craft adversarial examples against the substitute, and *transfer* them to a victim model, with very little information about the victim. Recent work has further developed a technique that uses the victim model as an oracle to label a synthetic training set for the substitute, so the attacker need not even collect a training set to mount the attack. We extend these recent techniques using *reservoir sampling* to greatly enhance the efficiency of the training procedure for the substitute model. We introduce new transferability attacks between previously unexplored (substitute, victim) pairs of machine learning model classes, most notably SVMs and decision trees. We demonstrate our attacks on two commercial machine learning classification systems from Amazon (96.19% misclassification rate) and Google (88.94%) using only 800 queries of the victim model, thereby showing that existing machine

Input | Model Activations | Output
Legitimate 1
Adversarial 4

Figure 1: An adversarial sample (bottom row) is produced by slightly altering a legitimate sample (top row) in a way that forces the model to make a wrong prediction whereas a human would still correctly classify the sample [19].

Adversarial sample transferability is the property that some adversarial samples produced to mislead a specific model f can mislead other models f'—even if their architectures greatly differ [22, 12, 20]. A practical impact of this property is that it leads to *oracle-based* black box attacks. In

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.

Least Gaussian projections of the data (interesting?)

how to define non-Gaussianity?

covariance and mean given: Gaussian distribution maximizes the entropy

Objective: minimize $H(t)$ for $t = \boldsymbol{w}^T \boldsymbol{x}$
t is normalized to zero mean and unit variance

This is difficult to optimize
→ finding unimodal super-Gaussians
→ finding multimodal distributions

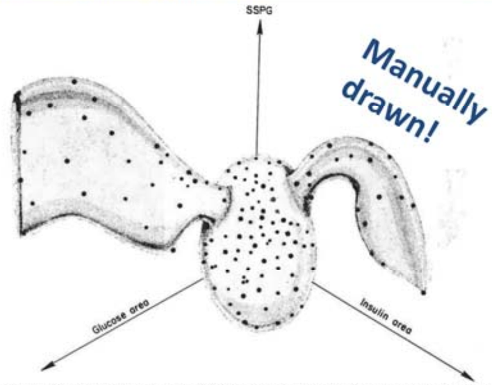
Other criteria are given for ICA: kurtosis and different contrast functions which measure non-Gaussianity

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.***

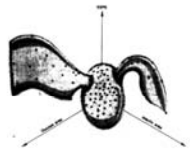
Miller-Reaves Diabetes Study (2/2)



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 $\mathbb{R}^2$

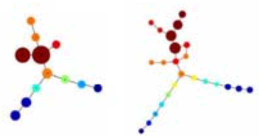


Given a point cloud data set X and a covering *U*
⇒ *simplicial complex*

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

$$f: X \to Z$$

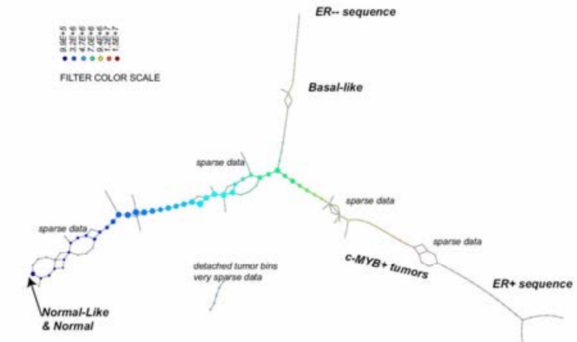
$$\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$$



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

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.*

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.***

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: small data
- Integration in "real-world" (e.g. Hospital), mobile
- 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!

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?

Sample Questions

- What are the problems in high-dimensional spaces?
- When is the human-in-the-loop beneficial?
- What is a Autoencoder and when would you use it?
- When would you use PCA?
- What did the authors of the Miller-Reavens study do?
- Why is the question "what is interesting?" a hard question?