Representations

Measurements and Data

Patterns

Data Types, Transformations, Scale

September 26, 2019

Recap: The dilemma of learning

What should we do about over-fitting?

- Minimize **expected classification error**!
- Maximize **generalization** (whatever it is).

What can we do about over-fitting?

- Minimize empirical classification error!
- Maximize estimated empirical generalization performance by cross validation.

Estimation of Dependences Based on Empirical Data

Vladimir Vapnik, Springer Verlag (1982)

What is the learning problem?

We search a function $f(x) \in \mathcal{C}$ out of the hypothesis class / solution space $\mathcal{C}$ so that

$$\begin{aligned} f &: \mathcal{X} \to \mathcal{Y} \\ & x \mapsto y = f(x) \end{aligned}$$

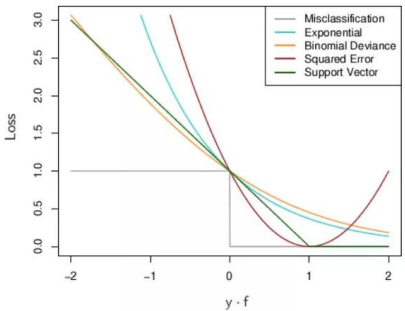
Often, we index the function $f(x) = f_\theta(x)$ by a parameter θ.

Estimation of Dependences Based on Empirical Data

Quality of the estimate for given X

The loss function Q measures the deviation between dependent variables y and prediction $f(x)$

$$Q(y, f(x)) = \begin{cases} \big(y - f(x)\big)^2 & \text{quadratic loss (regression)} \\ \mathbb{I}_{\{y \neq f(x)\}} & \text{0-1 loss (classification)} \\ \exp(-\beta y f(x)) & \text{exponential loss (classification)} \end{cases}$$



Expected risk

Conditional expected risk

Given the random variable X the conditional expected risk is defined as

$$R(f,X) = \int_{\mathcal{Y}} Q(Y,f(X))P(Y|X)dY$$

Total expected risk ...

... for the random variables X, Y is defined as

$$\begin{aligned} R(f) &= \mathbb{E}_X[R(f,X)] = \int_{\mathcal{X}} R(f,X)P(X)dX \\ &= \int_{\mathcal{Y}} \int_{\mathcal{X}} Q(Y,f(X))P(X,Y)\,dX\,dY \end{aligned}$$

Empirical risk

Training data $\mathcal{Z}^{\text{train}}$ and test data $\mathcal{Z}^{\text{test}}$

Usually, the samples are split into training data and test data. Additional validation data is used to guide estimator selection.

Test data cannot be used before the final estimator has been selected!

$$\mathcal{Z}^{\text{train}} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$$
$$\mathcal{Z}^{\text{test}} = \{(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})\}$$

Training error $\hat{R}(\hat{f}, \mathcal{Z}^{\text{train}})$ for Empirical Risk Minimizer $\hat{f}$

select ERM s.t. $\hat{f} \in \arg\min_{f \in \mathcal{C}} \hat{R}(f, \mathcal{Z}^{\text{train}})$

training error: $\hat{R}(\hat{f}, \mathcal{Z}^{\text{train}}) = \frac{1}{n} \sum_{i=1}^{n} Q(Y_i, \hat{f}(X_i))$

Empirical test error and expected risk

Test error $\hat{R}(\hat{f}, \mathcal{Z}^{\text{test}})$

Test data cannot be used before the final estimator has been selected. The test error amounts to

$$\hat{R}(\hat{f}, \mathcal{Z}^{\text{test}}) = \frac{1}{m} \sum_{i=n+1}^{m+n} Q(Y_i, \hat{f}(X_i))$$

When we use test data for validation, then estimator adaptation introduces statistical dependencies between outcome of the learning process (estimator) and test data. This design flaw yields a too optimistic estimate of the test error.

Distinguish between test error and expected risk!

$$\hat{R}(\hat{f}, \mathcal{Z}^{\text{test}}) \neq \mathbb{E}_X[R(\hat{f}, X)] \qquad \Rightarrow$$

What is the probability $\mathbb{P}\Big(\big|\hat{R}(\hat{f}, \mathcal{Z}^{\text{test}}) - \mathbb{E}_X[R(\hat{f}, X)]\big| > \epsilon\Big)$?

The test error empirically estimates the expected risk. To assess the quality of this estimate, we should report mean and variance or another measure of deviation.

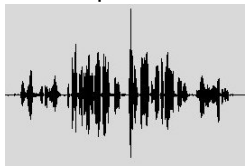
Intelligent Hearing Aids and Classification

Phonak designs the next generation of intelligent hearing instruments (HI). The HI autonomously classifies the acoustic environment and adapts its sound processing strategy to improve speech comprehensibility.

Waveforms of four acoustic environments:

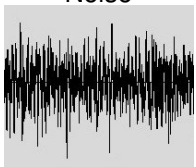
Music

Speech



Speech in noise

Noise

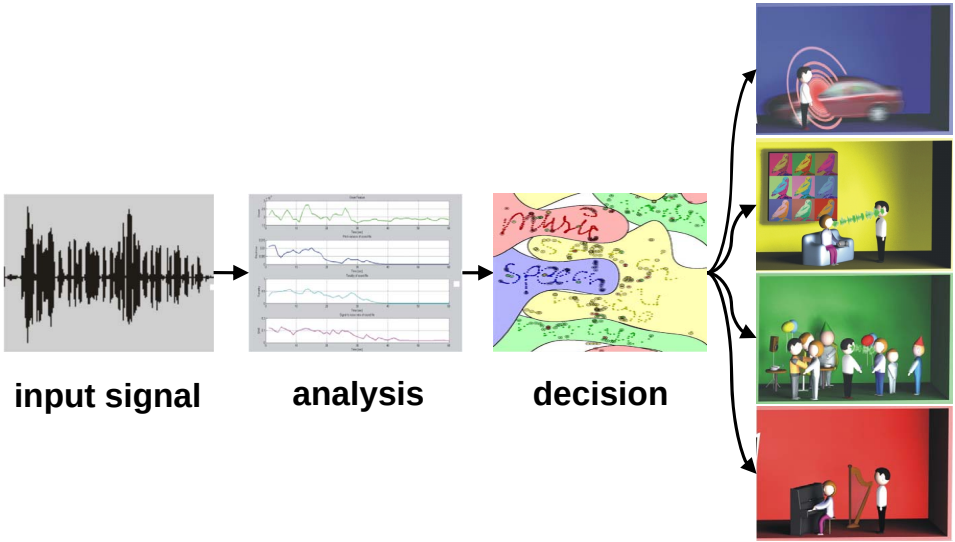


Acoustic Analysis as a Machine Learning Problem

1. **Feature Extraction**: Raw data often have to be transformed to reduce information. (Many) potentially useful features are extracted and a subset is then selected for processing! $\Rightarrow$ feature extraction
2. **Classifier Design**
3. **Postprocessing of classification results**: This process adapts the classifier output to enable stable scoring and to suppress temporal fluctuations.
4. **Scoring** prepares subsequent decision making like hearing aid control (e.g. different parameter settings).

The Processing Pipeline of Hearing Instruments

Data Flow of acoustic signal and processing steps of a hearing instrument.



Sound Classification: an Example

HI: Phonak Savia

Features:

BBSNR :	Broadband Signal to Noise Ratio
CG Means :	Center of Gravity (Means)
CGVar :	Center of Gravity (Variance)
OnsetComMean :	Common Onset Means
Tonality	
PitchVar :	Pitch Variance
LowVsHigh :	Low Freq vs High Freq
WWPtc :	Wind protection
Telefon	
MeanRMSLevel :	(Power : average root-mean-square Level)
LowFreqPow :	Low Frequency Power.

Classification in 4 classes:

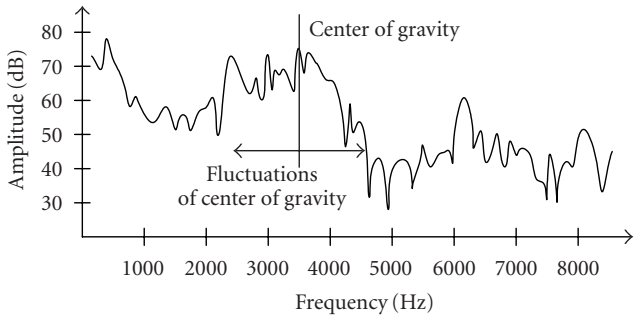
1. speech
2. music
3. speech in noise
4. noise

Spectral Center of Gravity

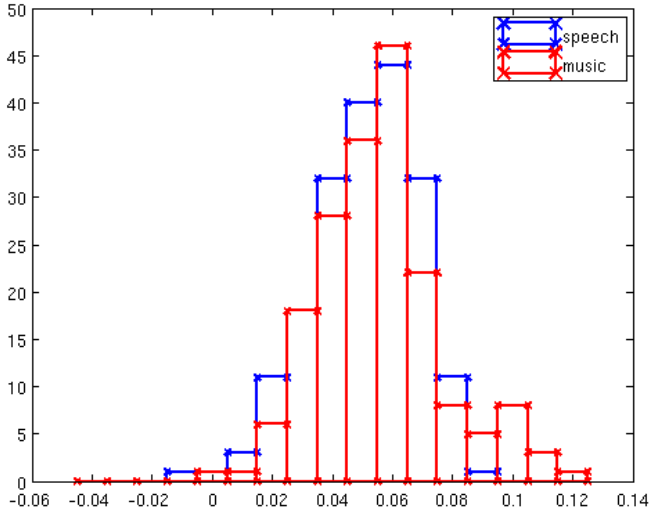
The spectral profile of a sound can contribute to the classification in that its form may differ for different sound classes, such as for music or noise. Moreover, the shape of the spectrum of most sound sources remains constant as the overall level of the sound is changed.

M. Büchler et al., EURASIP Journal on Applied Signal Processing 2005:18, 2991-3002

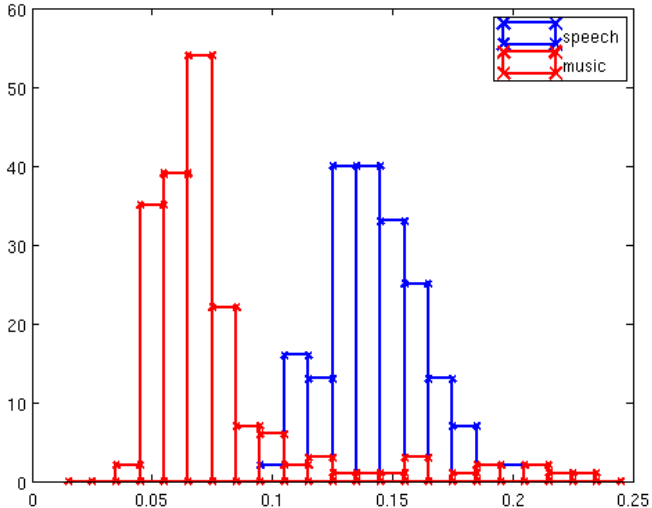
One of the psychoacoustically interesting features are the spectral center of gravity and the fluctuations of the spectral center of gravity.



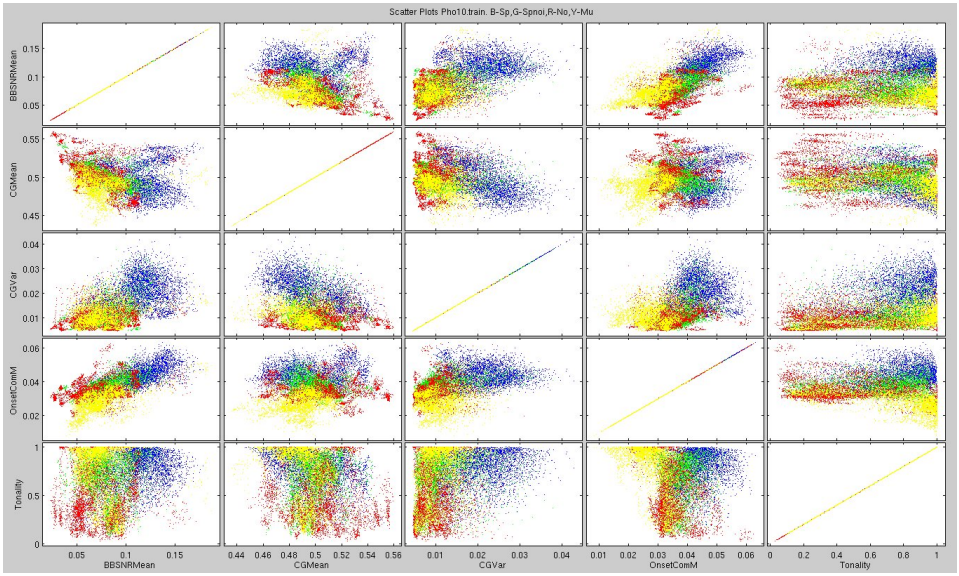
Histograms of CGMeans



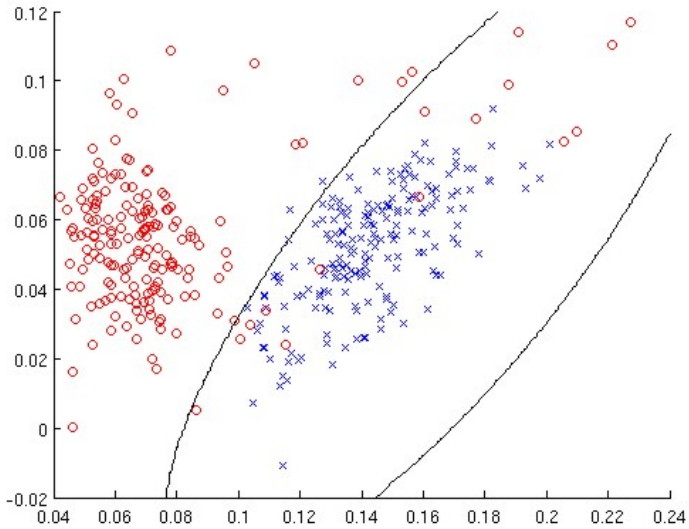
Histograms of BBSNR



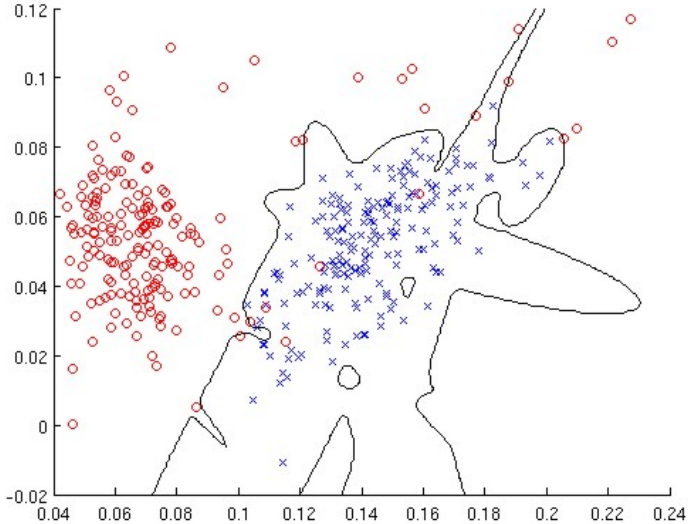
Scatter Plots of Phonak Features



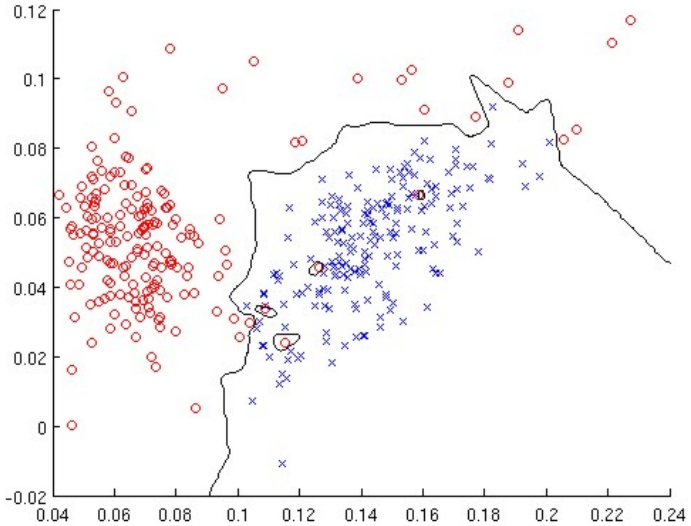
Classification with One Gaussian Component



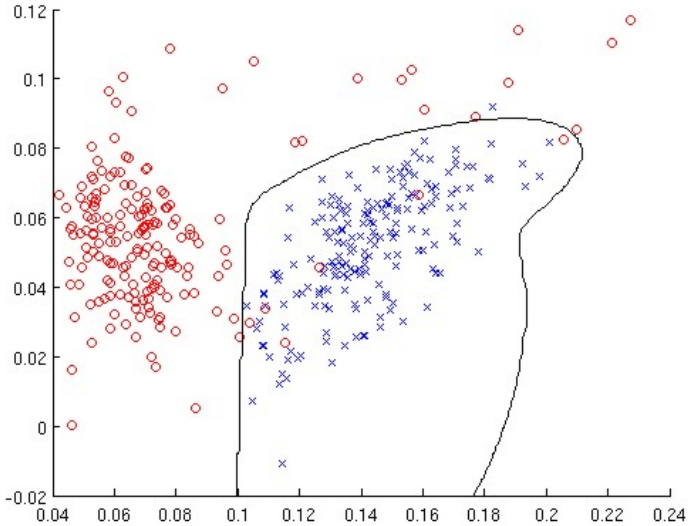
Overfitting (10 Gaussians)



Overfitting (Nearest Neighbor Classifier)



Optimized Non-Linear Classification



Taxonomy of Data

Objects, measurements, patterns and more!

Pattern analysis requires to find structure in sets of object representations.

Object space: We are given a design / configuration / object space $\mathcal{O}$!

Measurement: Given an object set, a measurement X maps an object set into a domain $\mathbb{K}$.

$$X : \mathcal{O}^{(1)} \times \cdots \times \mathcal{O}^{(R)} \to \mathbb{K}$$
$$(o_1, \ldots, o_R) \mapsto X_{o_1,\ldots,o_R}$$

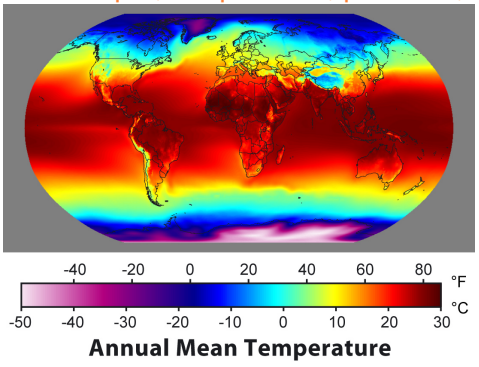
Examples:

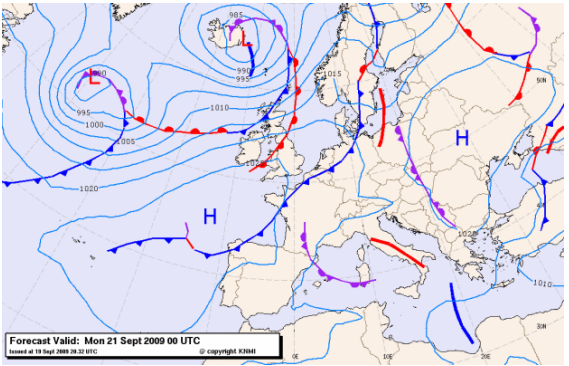
Feature vectors:	$X : \mathcal{O} \to \mathbb{R}^d$,	$o \mapsto X_o$
Classification data:	$X : \mathcal{O} \to \mathbb{R}^d \times \{1, \ldots, k\}$,	$o \mapsto (X_o, Y_o)$
Regression data:	$X : \mathcal{O} \to \mathbb{R}^d \times \mathbb{R}$,	$o \mapsto (X_o, Y_o)$
Proximity data:	$X : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$,	$(o_1, o_2) \mapsto X_{o_1,o_2}$

Examples of Data

a) monadic data: $X : \mathcal{O} \to \mathbb{R}^d, \quad o \mapsto X_o$

water depth, temperature, pressure, intensity, ...



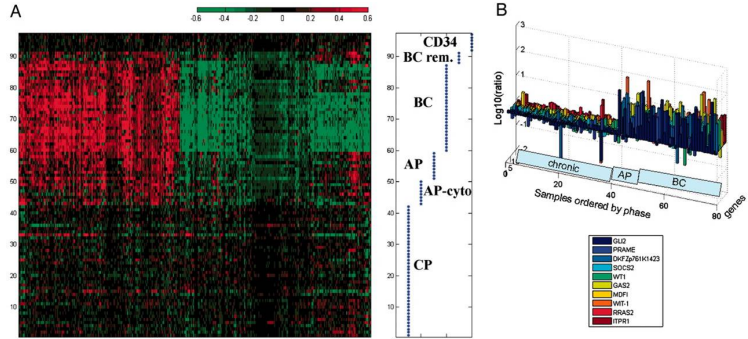


http://www.globalwarmingart.com/images/a/aa/Annual_Average_Temperature_Map.jpg

Monadic data characterize configurations or objects without reference to other configurations. The temperature and the pressure are measured for each location in absolute terms.

b) dyadic data: $X : \mathcal{O}^{(1)} \times \mathcal{O}^{(2)} \to \mathbb{R}, \quad (o_1, o_2) \mapsto X_{o_1,o_2}$

{users} × {websites}
{word counts} × {documents}
{users} × {permissions} (role based access control)
{diseases} × {gene expression levels}

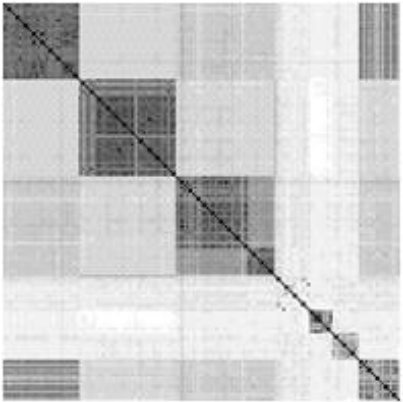


Radich J P et al. PNAS 2006;103:2794-2799

pairwise data: $\mathcal{O} \times \mathcal{O} \to \mathbb{R}, \quad (o_1, o_2) \mapsto X_{o_1,o_2}$

{image patches} $\times$ {image patches}

{proteins} $\times$ {proteins}



145 globins have been compared to each other by sequence alignment. The globins have been selected from 4 classes of α-globins, β-globins, myoglobins and globins of insects and plants. The high similarity is depicted by dark entries in the matrix.

c) polyadic data:

$$\mathcal{O}^{(1)} \times \mathcal{O}^{(2)} \times \mathcal{O}^{(3)} \to \mathbb{R}, \quad (o_1, o_2, o_3) \mapsto X_{o_1,o_2,o_3}$$

$\{\text{test persons}\} \times \{\text{behaviors}\} \times \{\text{traits}\}$

Preferential choice data in market analysis of consumers have a three way characteristics

$\{\text{test persons}\} \times \{\text{item 1}\} \times \{\text{item 2}\}$

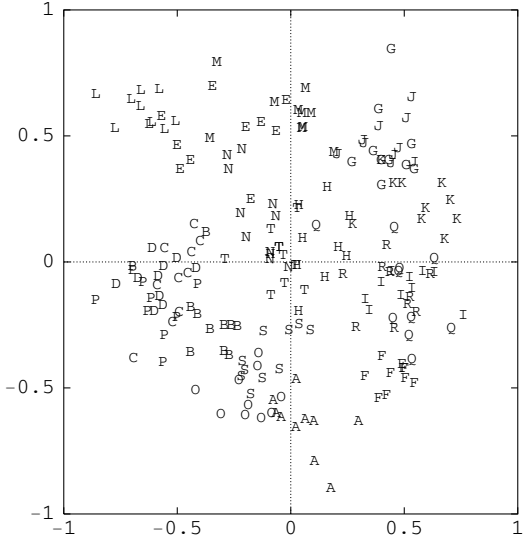
Three way data or an even more complex tuples of data are known in the literature but it is increasingly difficult to extract patterns from these data sets. However, the nature of the datum has to be considered when we propose a structure definition to extract patterns.

William G Jacoby, *Data Theory and Dimensional Analysis* 1991, SAGE Publications, Inc.

Table of Data Types

Data type	design space $\mathcal{O}$	measurement space $\mathcal{X}$
vectorial data	monadic, $\mathcal{O}^{(1)}$	multi scale, $\mathbb{R}^D$
contingency	monadic, $\mathcal{O}^{(1)}$	categorial,
tables	monadic, $\mathcal{O}^{(1)}$	$\bigotimes_{d=1}^{D}\{1,\ldots,n_d\}$
histogram data	monadic, $\mathcal{O}^{(1)}$	d-dim. probability simplex,
similarity data	pairwise, $\mathcal{O}^{(1)} \times \mathcal{O}^{(1)}$	quant. or ordinal, univariate $\mathbb{R}$
multiway similarity data	3–adic, $\mathcal{O}^{(1)} \times \mathcal{O}^{(2)} \times \mathcal{O}^{(2)}$	quant. or ordinal, univariate $\mathbb{R}$
co-occurrence	dyadic, $\mathcal{O}^{(1)} \times \mathcal{O}^{(2)}$	absolute, $\mathbb{N}$
n-th $\times$ occurrence	polyadic, $\mathcal{O}^{R}$	absolute, $\mathbb{N}$
single stimulus	dyadic, $\mathcal{O}^{(1)} \times \mathcal{O}^{(2)}$	quantitative, $\mathbb{R}$
rank preference	dyadic, $\mathcal{O}^{(1)} \times \mathcal{O}^{(2)}$	ordinal, $\mathbb{N}$
preferential choice	3–adic, $\mathcal{O}^{(1)} \times \mathcal{O}^{(2)} \times \mathcal{O}^{(2)}$	categorial, $\{-1, 0, 1\}$

Example for Vector Data



Data of 20 Gaussian sources in $\mathbb{R}^{20}$, projected onto 2 dimensions with PCA

Scales

Nominal or **categorial scale**

qualitative, but without quantitative measurements,
e.g. binary scale $\mathcal{X} = \{0, 1\}$ *(presence or absence of properties) or taste categories "sweet, sour, salty, bitter, umami".*

Ordinal scale

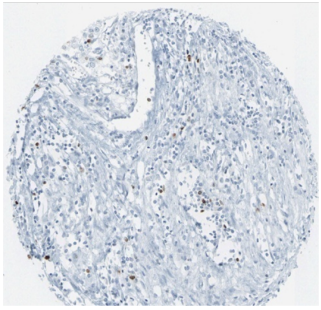
measurement values are meaningful only with respect to other measurements, i.e., the rank order of measurements carries the information, not the numerical differences *(e.g. information on the ranking of different marathon races!?)*

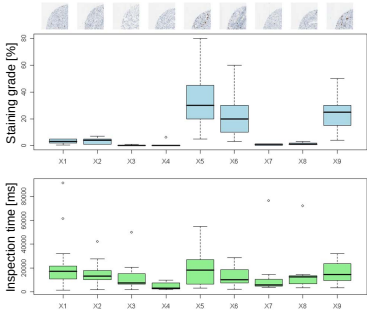
Quantitative scale

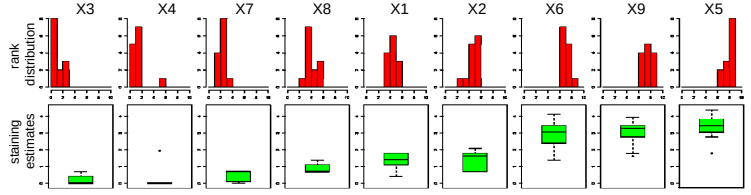
- **interval scale**: the relation of numerical differences carries the information. Invariance w.r.t. translation and scaling (Fahrenheit scale of temperature).
- **ratio scale**: zero value of the scale carries information but not the measurement unit. (Kelvin scale).
- **Absolute scale**: Absolute values are meaningful. (grades of final exams)

Data whitening: Normalize the values of a feature vector by the standard deviation (or another scale quantity) in this component. Thereby, differences in dynamic range (e.g., by different measurement units) are eliminated.

Staining estimation by pathologists







Transformation Invariances

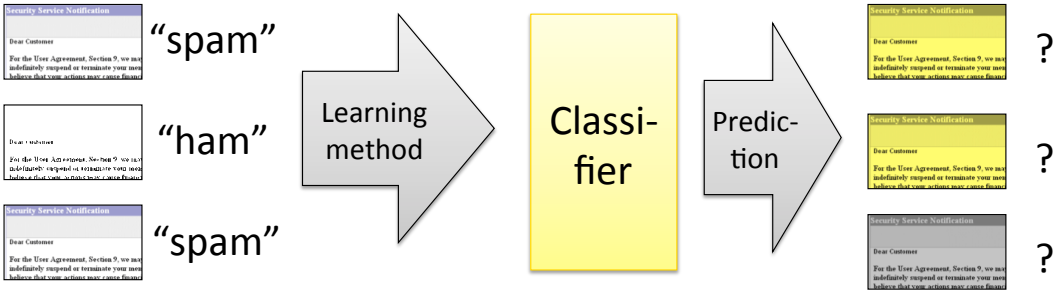
scale type	transformation invariances
nominal	$\mathcal{T} = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ bijective}\}$
ordinal	$\mathcal{T} = \{f : \mathbb{R} \to \mathbb{R} \mid f(x_1) < f(x_2),\ \forall x_1 < x_2\}$
interval	$\mathcal{T} = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax + c,\ a \in \mathbb{R}^+,\ c \in \mathbb{R}\}$
ratio	$\mathcal{T} = \{f : \mathbb{R} \to \mathbb{R} \mid f(x) = ax,\ a \in \mathbb{R}^+\}$
absolute	$\mathcal{T} = \{f : \mathbb{R} \to \mathbb{R} \mid f \text{ is identity map}\}$

Formal characterisation of different scale types and their invariance properties.

Importance of invariances: if the measurements are invariant under a set of transformation then the mathematical definition of structure should obey the same invariances.
Otherwise, our structure search procedure breaks the symmetry in an a priori (not data-dependent) way.

Learning Pipeline

Learning and prediction by algorithm $\mathcal{A}$ and hypothesis class $\mathcal{C}$



$\mathcal{X}$ data $\quad \mathcal{Y}$ label $\quad \mathcal{A} : \mathcal{X} \times \mathcal{Y} \to \mathcal{C}$ algorithm $\quad f : \mathcal{X} \to \mathcal{Y}$ classifier $f \in \mathcal{C}$

Mathematical Spaces

Topological spaces

Let $\mathcal{X}$ be a non-empty set and $\mathfrak{J}$ a collection of subsets of $\mathcal{X}$ such that:

1. $\mathcal{X} \in \mathfrak{J}$;
2. $\emptyset \in \mathfrak{J}$;
3. If $\mathcal{O}_1, \ldots, \mathcal{O}_n \in \mathfrak{J}$, then $\mathcal{O}_1 \cap \cdots \cap \mathcal{O}_n \in \mathfrak{J}$;
4. If for each $\alpha \in \mathcal{I}, \mathcal{O}_\alpha \in \mathfrak{J}$, then $\cup_{\alpha \in \mathcal{I}} \mathcal{O}_\alpha \in \mathfrak{J}$;

The pair of objects $(\mathcal{X}, \mathfrak{J})$ is called a topological space.

Topological spaces only describe the closeness/neighborhood of objects but they do not models any quantitative differences (distances) between the "degrees of closeness". The concept of a topological space is one of the most fruitful concepts of modern mathematics. It is the proper setting for discussions based on considerations of continuity.

Topological spaces allow us to introduce the concept of a neighborhood and to define neighborhood spaces in a very natural way.

Mathematical Spaces

Metric space

A pair of objects $(\mathcal{X}, d)$ consisting of a non-empty set $\mathcal{X}$ and a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a metric space provided that:

(1) Positivity: $d(x,y) \geq 0, \quad x, y \in \mathcal{X}$

(2) Uniqueness: $d(x,y) = 0$ if and only if $x = y$, $\forall\, x, y \in \mathcal{X}$

(3) Symmetry: $d(x,y) = d(y,x)$, $x, y \in \mathcal{X}$

(4) $\triangle$ inequality: $d(x,z) \leq d(x,y) + d(y,z)$, $x, y, z \in \mathcal{X}$

The function d is called a distance function or metric on $\mathcal{X}$ and the set $\mathcal{X}$ is called the *underlying set*.

Example: $(\mathbb{R}, d)$ is a metric space, where d is the function defined by $d(a,b) = |a - b|$, for all $a, b \in \mathbb{R}$.

Euclidean vector spaces

Let $\mathcal{X}$ be a non-empty set and $\mathcal{V} = (\mathcal{X}, +, \cdot)$ a vector space. A function $\phi : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which assigns two vextors $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ a real number is called scalar product on $\mathcal{V}$ if the following properties hold:

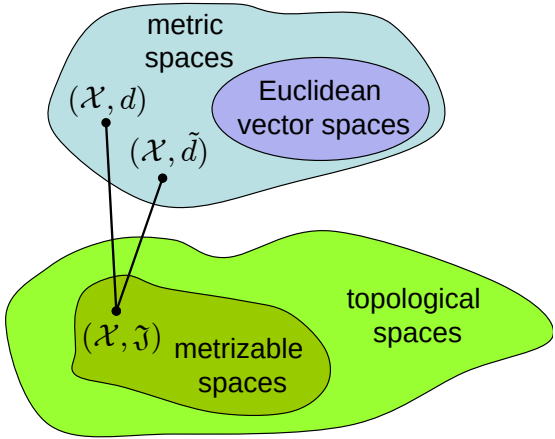
(1) distributivity: $\phi(\mathbf{x}_1 + \mathbf{x}_2, \mathbf{y}) = \phi(\mathbf{x}_1, \mathbf{y}) + \phi(\mathbf{x}_2, \mathbf{y})$

(2) commutativity: $\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{y}, \mathbf{x})$

(3) homogeneity: $\phi(\alpha\mathbf{x}, \mathbf{y}) = \alpha\phi(\mathbf{x}, \mathbf{y}), \forall\, \alpha \in \mathbb{R}$

(4) positive definiteness: $\phi(\mathbf{x}, \mathbf{x}) > 0$ for all $\mathbf{x} \neq 0$

A vector space with such a scalar product is called Euclidean vector space. The scalar product defines the norm $\|\mathbf{x}\| \stackrel{\triangle}{=} \sqrt{\phi(\mathbf{x}, \mathbf{x})}$.



Every element of a metric space corresponds to an element in a metrizable space.

From a machine learning point of view we have to answer the question how precisely we can actually gather metric information rather than topological information in an application scenario. Such an analysis then suggests the appropriate space to model the structures in the data.

Probability Spaces

Elementary event

$\omega_1, \ldots, \omega_N$ are samples points

Sample space

$\Omega = \{\omega_1, \ldots, \omega_N\}$
example:
n coin flips: $\Omega = \{\omega | \omega = (a_1, \ldots, a_n), a_i \in \{head, tail\}\}, N(\Omega) = 2^n$

Family of sets

An event A of an experiment is a set of elementary events with the following condition:

- $A \subset \Omega$
- result of an experiment: $\omega \in A$ or $\omega \notin A$

Probability Spaces

Algebra of events

Let A, B be events. $\mathcal{A}$ is an algebra of events, i.e., a set of subsets $A \subset \Omega$, for which holds:

- $\Omega \in \mathcal{A}$
- if $A \in \mathcal{A} \ \wedge \ B \in \mathcal{A}$,
  then $A \cup B \in \mathcal{A} \ \wedge \ A \cap B \in \mathcal{A} \ \wedge \ A \setminus B \in \mathcal{A}$

Example of an event with n coin flips: all sequences of n coin flips with more than $\frac{n}{2}(1+\epsilon)$ heads.

Probability of events

Assign weights $p(\omega_i)$ to elementary events $\omega_i \in \Omega$ with the following properties:

- $0 \leq p(\omega_i) \leq 1$ (non-negativity)
- $\sum_i p(\omega_i) = 1$ (normalization)

Probability of an event $A \in \mathcal{A}$ with $\mathbf{P}(A) = \sum_{\{i:\omega_i \in A\}} p(\omega_i)$

Probability model

A probability model or a probability space is a triple

$$(\Omega, \mathcal{A}, \mathbf{P})$$

with the sample set $\Omega = \{\omega_1, \ldots, \omega_n\}$, the event algebra $\mathcal{A}$ and the probabilities $\mathcal{P} = \{\mathbf{P}(A) | A \in \mathcal{A}\}$