Introduction

Central Problem of Pattern Recognition:
Supervised and Unsupervised Learning
Data Types, Transformations, Scale

Joachim M. Buhmann

Information Science & Engineering Group
Institute for Machine Learning
D-INFK, ETH Zurich

September 19, 2019

The Learning Problem of Pattern Recognition

- **Representation of objects.** $\Rightarrow$ Data representation
  Choosing the wrong data representation can induce inappropriate similarity measures !
- What is a pattern? **Definition/modeling of structure.**
  A statistical definition of good and poor structures is mandatory for rational pattern recognition!
- **Optimization**: Search for preferred structures
  Multiscale optimization yields efficient algorithms to detect good structures in data.
- **Validation**: are the structures indeed in the data or are they explained by fluctuations?

  Without validation, any pattern recognition strategy is doomed to fail.

What are Data ?

Measurements: (`Encyclopedia Britannica`)

Association of numbers with physical quantities and natural phenomena by comparing an unknown quantity with a known quantity of the same kind.

Merriam-Webster Online Dictionary – a definition of **data**:

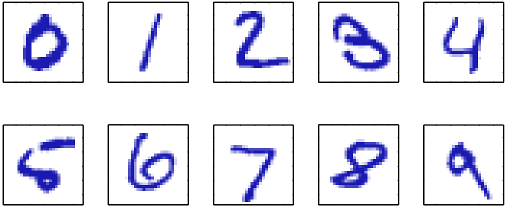
```
Etymology: Latin, plural of datum
1 : factual information (as measurements or statistics)
    used as a basis for reasoning, discussion, or
    calculation
2 : information output by a sensing device or organ that
    includes both useful and irrelevant or redundant
    information and must be processed to be meaningful
3 : information in numerical form that can be digitally
    transmitted or processed
```

Objects and Measurements

Goal: We like to represent objects of interest and characterize them according to their typical patterns for ...

- ... detection,
- ... classification,
- ... abstraction (compression), ...

Measurements represent objects in a data space, e.g., digits as objects and pixel intensity as measurements.



Feature Space

Measurement space $\mathcal{X}$: the mathematical space in which the data are represented, e.g., numerical ($\mathcal{X} \subset \mathbb{R}^d$), boolean ($\mathcal{X} = \mathbb{B}$) or categorial ($\mathcal{X} = \{1, \ldots, k\}$) features.

Features are derived quantities or indirect observations which often significantly compress the information content of measurements.
Examples are edges, corners, motion vectors in images or video, mel-cepstral features in acoustics, wavelet amplitudes in seismology, ...

Remark: The selection of a specific feature space predetermines the metric to compare data; this choice is the first significant design decision in a machine learning system.

The Goal of Learning

Estimation of Dependencies Based on Empirical Data

(V. Vapnik 1983, Springer Verlag)

Typical learning problems:

- classification: learning an indicator function
- regression: learning a real valued function
- density estimation: learning a probability density of the data source

- dimension reduction: learning a linear or nonlinear projection
- data compression: learning a coding efficient representation

Learning requires to infer a **functional** or **statistical** relationship between variables when we only observe noisy samples. Approximation and interpolation in function estimation are such procedures.
The problem without additional assumptions is mathematically ill-defined since many different functions might be compatible with noisy observations.
We, therefore, require that our inference has to “work” on future data. Mathematically, the **expected quality** of inference should be high and not necessarily the **empirically observed quality**.

Supervised Learning: Classification

Learning with a teacher: The conceptually simplest form of learning is "supervised learning". A teacher (oracle) provides the correct answer during training.

Data are pairs of features and response variables $\{(x_1, y_1), \ldots (x_n, y_n) \ : \ x_i \in \mathcal{X} \subset \mathbb{R}^d, y_i \in \mathbb{K}\}$ with

- $\mathbb{K} = \{1, \ldots, k\}$ for classification where $\mathbb{K}$ is an index set for the classes;
- $\mathbb{K} = [0, 1]^k$ is the space of assignments for probabilistic classification.
- $\mathbb{K} \subset \mathbb{R}$ for regression.

Problem: The data are noise contaminated, e.g., the response variable $y = f(x, w) + \eta$ depends on the function with parameters w and (Gaussian white) noise η.

Question: How can we infer a functional relationship $f(x, w)$ from data which are described by the statistical relationship $\mathbf{P}(X = x, Y = y)$?

Answer: statistical learning theory! Define a function class

$$\mathcal{C} = \{f(x, w) \ : \ w \in \mathcal{W}, x \in \mathbb{R}^d\}$$

where w indexes the function (hypothesis) class $\mathcal{C}$.

It turns out that the "complexity" of the function class $\mathcal{C}$ is the essential concept to describe the difficulty of learning. If we have too few data and we work with a too complex function class then learning algorithms have a strong tendency to overfit, i.e., to confuse/interpret noise as signal.

What you know already!

Lecture: **Introduction to Machine Learning (Spring)**

- (Semi-)supervised, unsupervised learning
- Key Machine Learning concepts
- Overview of most important algorithms
- Discriminative vs. generative modeling

Here:

- Formal statistical learning theory perspective
- More depth for selected topics: validation, structured SVMs, ...
- Advanced topics: PAC learning, sequences, time series, ...

Key concepts of machine learning

- Trade-off: training error vs. model complexity
  - Regularizers prevent overfitting
  - Hyper parameter/model selection via cross-validation
- Correspondence prob. modeling: Loss $\widehat{=}$ likelihood, regularizer $\widehat{=}$ prior
- Kernel trick: replace inner products by kernel function
- Neural nets for feature learning
- Discriminative vs. generative models
  - Discrim.: Learn function $f : \mathcal{X} \to \mathcal{Y}$
  - Gen.: Learn joint distribution $\mathbf{P}(X = x, Y = y)$
    - $\mathbf{P}(Y)$, $\mathbf{P}(X|Y)$
    - $\mathbf{P}(X)$, $\mathbf{P}(Y|X)$ $(\sim f)$
- Unsupervised learning as latent variable modeling (clustering $\widehat{=}$ classification, dim. Reduction $\widehat{=}$ regression)
  - Training by EM algorithm

What you should know?

Representation/ features	Linear hypotheses; nonlinear hypotheses with nonlinear feature transforms, kernels, learn nonlinear features via neural nets		
Paradigm:	Discriminative vs. generative		
Probabilistic / Optimization Model:	Likelihood Loss-function Squared loss = Gaussian lik., 0/1, Perceptron, Hinge, cost sensitive, multi-class hinge, reconstruction error, logistic loss=Bernoulli lik., cross-entropy loss=Categorical lik.	* +	Prior Regularization L^2 norm (=Gaussian prior), L^1 norm (=Laplace prior), early stopping, dropout Categorical;. Beta/Dirichlet priors
Method:	Exact solution, Gradient Descent, (mini-batch) SGD, Reductions, EM, Bayesian model averaging		
Evaluation metric:	Mean squared error, Accuracy, F1 score, AUC, Confusion matrices, compression performance, log-likelihood on validation set		
Model selection:	K-fold Cross-Validation, Monte Carlo CV, Bayesian model selection		

Source: Lecture Slides, Introduction to Machine Learning 2018, Andreas Krause, ETHZ

Recall: Introduction to Machine Learning (Spring)

Supervised learning

Linear regression, ridge regression, Perceptron, Support Vector Machines, kernelized SVM, kernel ridge regression, k-Nearest Neighbor, l1-SVM, Lasso, logistic regression, (deep) neural nets, convolutional neural networks, Gaussian and categorical (Naive) Bayes Classifiers, Gaussian mixture Bayes classifiers, ...

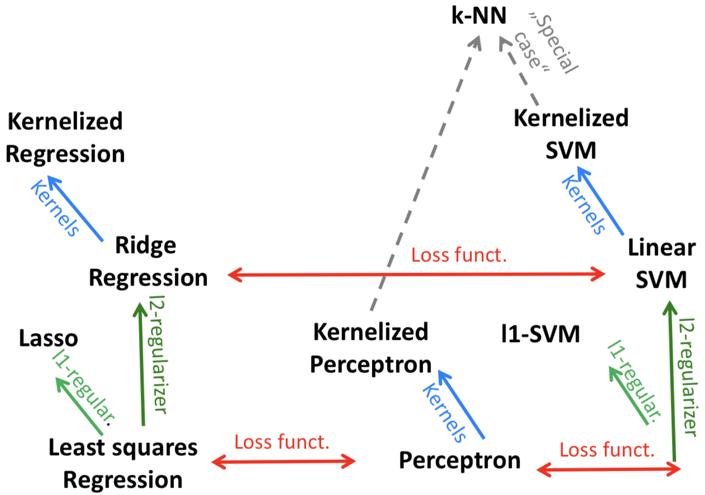
Unsupervised learning

k-Means, Gaussian mixtures, semi-supervised GMMs, anomaly detection, Principal Component Analysis, Kernel-PCA, neural net autoencoders, GANs

Optimization algorithms

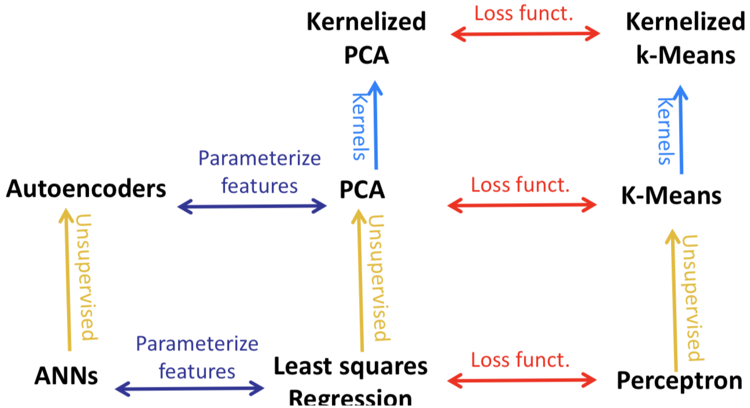
(Stochastic) Gradient Descent, EM Algorithm

Supervised learning via risk minimization



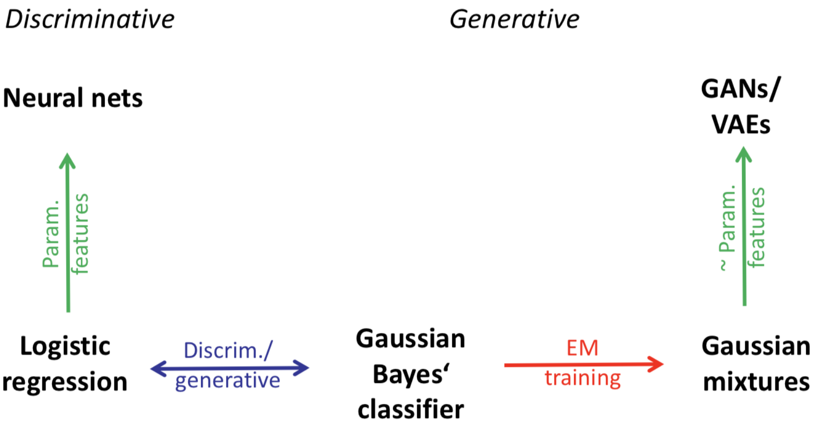
Source: Lecture Slides, Introduction to Machine Learning 2018, Andreas Krause, ETHZ

Supervised vs. unsupervised learning



Source: Lecture Slides, Introduction to Machine Learning 2018, Andreas Krause, ETHZ

Generative vs. discriminative modeling



Source: Lecture Slides, Introduction to Machine Learning 2018, Andreas Krause, ETHZ