Statistical Natural Language Processing

Part III: Basics of Natural Language Processing

Henning Wachsmuth

https://ai.uni-hannover.de

Learning Objectives

Concepts

- Basic concepts from linguistics
- Challenges of language understanding
- The notion of machine learning

Methods

- Fundamental techniques in natural language processing
- Standard techniques in machine learning
- Overfitting and underfitting of data

Notice

- Some concepts and methods are reviewed briefly here only.
- For more details, see for example my bachelor's lecture Introduction to Natural Language Processing

Outline of the Course

I. Overview

II. Basics of Data Science

III. Basics of Natural Language Processing

- Linguistics
- Fundamental NLP Techniques
- Machine Learning
- Data Mining
- Conclusion

IV. Representation Learning

V. NLP using Clustering

VI. NLP using Classification and Regression

VII. NLP using Sequence Labeling

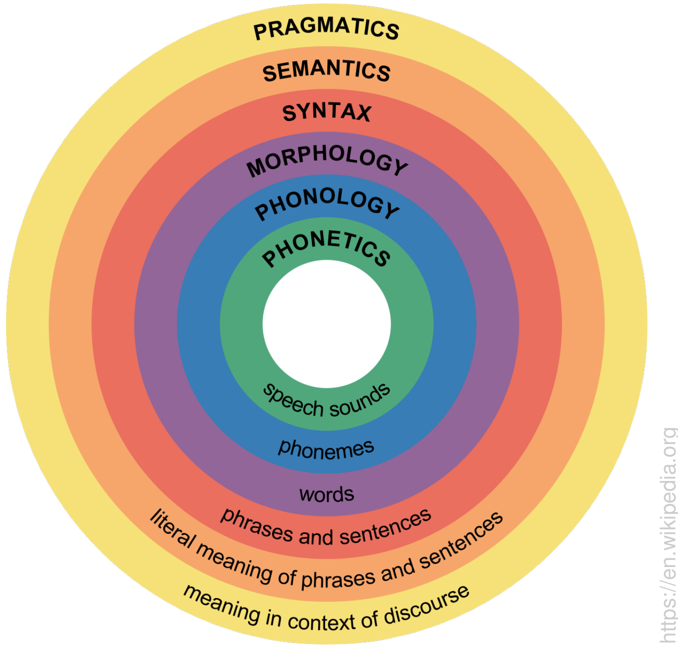
VIII. NLP using Neural Networks

IX. NLP using Transformers

X. Practical Issues

Linguistics

Linguistics



Linguistics

- The study of natural language(s) in terms of form, meaning, and context

Linguistic Levels

- Phonetics. Physical aspects of speech sounds
- Phonology. Linguistic sounds of a particular language
- Morphology. Senseful components of words
- Syntax. Structural relationships between words, usually in a sentence
- Semantics. Meaning of single words and compositions of words
- Discourse. Composition of linguistic units larger than a sentence
- Pragmatics. Use of language to accomplish certain goals

Linguistics

Main Morphological Concepts

Word

- The smallest unit of language that is to be uttered in isolation

"cats" and "ran" in "cats ran."

Lemma

- The dictionary form of a word

"cat" for "cats"   "run" for "ran"

Stem

- The part of a word that never changes

"cat" for "cats"   "ran" for "ran"

Token

- The smallest text unit in NLP: A word, number, symbol, or similar

  Whitespaces are usually not considered as tokens.

"cats", "ran", and "." in "cats ran."

Linguistics

Main Syntactic Concepts

Part-of-speech (POS)

- The lexical category (or word class) of a word
- Abstract classes. Nouns, verbs, adjectives, adverbs, prepositions, ...
- POS tags. NN (single nouns), NNS (plural n.'s), NNP (proper n.'s), ...

Phrases

- A contiguous sequence of words, functioning as a single meaning unit
- Phrases often contain nested phrases.
- Types. Noun phrase (NP), verb phrase (VP), prepositional phrase (PP)
  Sometimes also adjectival phrase (AP) and adverbial phrase (AdvP)

Clause

- The smallest grammatical unit that can express a complete proposition
- Types. Main clause, subordinate clause

Sentence

- A grammatically independent linguistic unit with one or more words

Linguistics

Main Semantic Concepts

Two types of semantics

- Lexical. The meaning of words and multi-word expressions
- Compositional. The meaning of the composition of words

Entities

- An object from the real world
- Named entities. Persons, locations, organizations, products, ...

  "Prof. Dr. Henning Wachsmuth" "Hannover" "Uni Hannover"

- Numeric entities. Values, quantities, ranges, periods, dates, ...

  "in this year" "2023-10-26" "$ 100 000" "762-12377"

Relations

- Semantic. Relations between entities, e.g., person *lives in* location
- Temporal. Relations describing courses of events, e.g., after <A>, <B>

Linguistics

Main Discourse and Pragmatics Concepts

Discourse (structure)

- Linguistic utterances larger than a sentence, e.g., paragraphs or articles
- Discourse segment. Linguistic building block within a discourse
- Coherence relation. Semantic or pragmatic relations between segments

Coreference

- Two or more expressions in a text that refer to the same thing
- Types. Pronouns, coreferring noun phrases, ...

> "Apple Inc. is based in the US. The company is called Apple. They make hardware.

Speech acts

- Linguistic utterances with a performative function

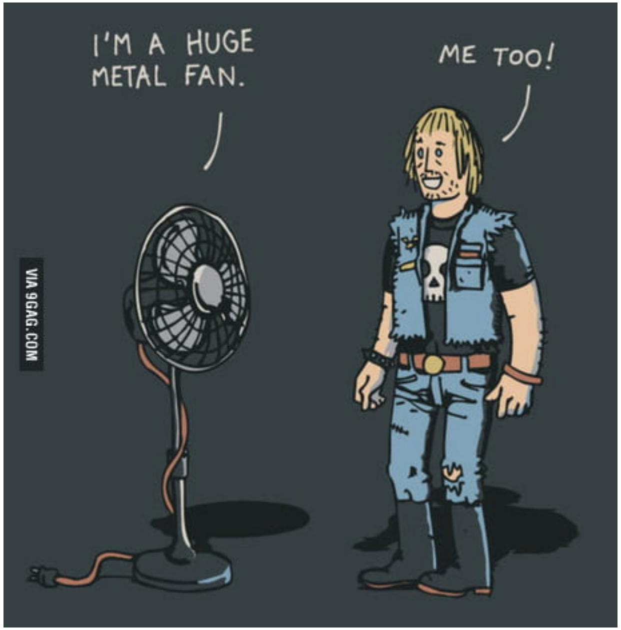
Presupposition and implicature

- Presupposition. Linguistic utterances assume things.
- Implicature. Linguistic utterances suggest things.

Challenges in Language Understanding

Ambiguity in natural language

- Phonetic. “wreck a nice beach”
- Word sense. “I went to the bank.”
- Part of speech. “I made her duck.”
- Attachment. “I saw a man with a telescope.”
- Coordination. “If you love money problems show up.”
- Quantifier scope. “I didn’t buy a car.”
- Speech act. “Have you emptied the dishwasher?”



https://9gag.com

Selected other challenges

- World knowledge. “Putin’s view of Ukraine is wrong”
- Domain dependency. “Read the book!”
- Language dependency. “Bad”

Fundamental NLP Techniques

Fundamental NLP Techniques

Recap: Rule-based NLP

- Based mainly on manually defined rules that encode expert knowledge
- Knowledge includes rules, lexicons, grammars, and similar.

Statistics in rule-based NLP

- Some techniques employ statistics to make decisions or to weigh rules.
- As such, they reflect the transition from rule-based to statistical NLP.

Selected techniques

- Hand-crafted decision trees. Apply nested if-then-else rules to a text.
- Finite-state transducers. Sequentially rewrite input to output sequence.
- Template-based generation. Create texts by filling predefined slots.
- Lexicon-based matching. Match text spans with terms from a lexicon.
- Regular expressions. Extract text spans that follow sequential patterns.
- Probabilistic parsing. Infer hierarchical structures of text spans.
- Language modeling. Generate sequences of words and other tokens.

Fundamental NLP Techniques

NLP Processes

From single tasks to processes

- Most NLP applications aim at combinations of different types of output.
- This means that several analysis or synthesis steps may be needed.

Ways to realize NLP processes

- Pipelines. Sequentially apply a set of methods to a text, such that the output of one method is the input to the next.
- Joint models. Perform multiple analysis/synthesis steps simultaneously.
- Neural models. Neural networks often operate on the raw input text.

  In any way, some kind of sequential pipeline is used for most NLP processes.

Example: Information extraction (IE)

- The mining of entities, their attributes, and their relations from text
- Usually, IE requires a pipeline of several analysis steps.
- The output is structured information, e.g., for the use in databases.

Fundamental NLP Techniques

Example: Information Extraction

Example: Extraction of founding dates

Time entity — Organization entity

" 2014 ad revenues of Google are going to reach

Reference — Time entity

$20B . The search company was founded in '98 .

Reference — Time entity — Founded relation

Its IPO followed in 2004 . [...] "

Output: Founded("Google", 1998)

Text analysis pipeline for this example

input text → Sentence splitting → Tokenization → Part-of-spech tagging → Phrase chunking → Time entity recognition → Named entity recognition → Reference resolution → Founded relation detection → **founded relations**

Machine Learning

Machine Learning

Example: Decision Making

?

Learning task

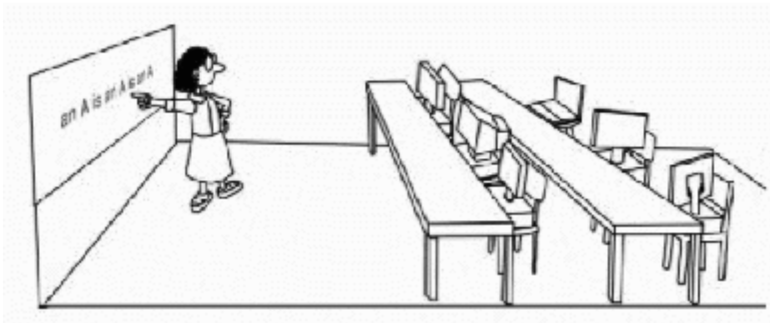
- What criteria form the basis of a decision?
- How is the decision made?

Machine Learning

Definitions

Machine learning (Samuel, 1959)

- The ability of an algorithm to learn without being explicitly programmed



An algorithm is said to learn... (Mitchell, 1997)

- ... from experience
- ... with respect to a given prediction task
- ... and some performance measure,
- ... if its performance on the task increases with the experience.

Machine Learning
Prediction

Prediction task

- A real-world problem that can be solved by a *target function* $\gamma : O \to C$
- Input. Objects $o_1, o_2, \ldots$ of some real-world concept O
- Output. Information $c_1, c_2, \ldots$ of some target variable C to solve the task

  The values of C are all of the same kind, for instance, all nominal labels.

(Ideal) Target function γ

- A function that interprets any object $o \in O$ to infer $\gamma(o) \in C$
- γ is operationalized by a human or some other real-world mechanism.
- Machine learning aims at prediction tasks where γ is unknown.

  This includes most NLP tasks, also those that can be tackled well with rules.

Prediction using machine learning

- Machine learning finds statistical patterns in examples of O that are relevant to infer C.

Machine Learning

Relation of NLP and Machine Learning

Machine learning in NLP

- Task. Predict output information $c \in C$ for a given text (or span of text).
- Experience. Texts, possibly annotated for C
- Performance. In terms of some effectiveness measure

Output information

- Text labels and scores. Topic, sentiment, grades, ...
- Span annotations. Tokens, entities, ...
- Span classifications. Entity types, POS tags, ...
- Span relations. Entity relations, coherence relations, ...
- Probabilities. For example, of next words to generate

Two-way relationship

- NLP often uses machine learning to produce output information.
- Its output may be the input to machine learning, e.g., to train a classifier.

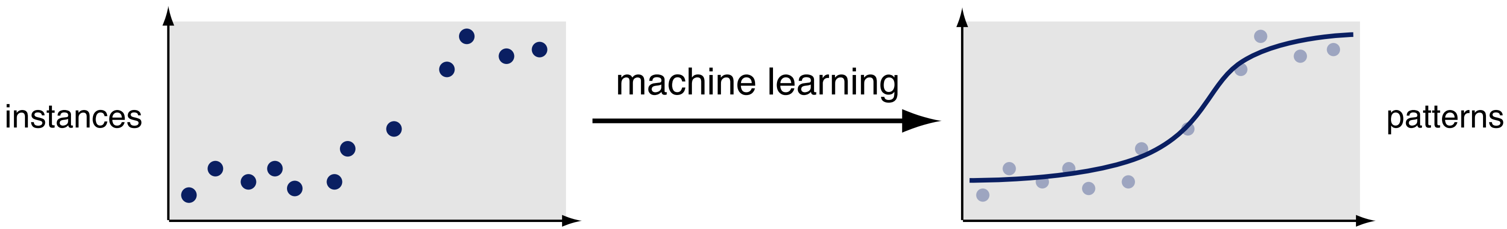
Machine Learning

Model

Machine learning models

- A model $y : X \to C$ is a mapping from formalized input instances X to an output target variable C
- y generalizes found patterns in X to approximate the target function γ.

  Machine learning seeks for the optimal y with respect to some performance measure.



Model vs. target function

- γ and y differ in the complexity and representation of their domain.
- Complexity. Objects $o \in O$ are abstracted into vectors $\mathbf{x} \in X$ using some mapping function α, $\mathbf{x} = \alpha(o)$.
- Representation. $y(\mathbf{x})$ is the formalized counterpart of $\gamma(o)$.

Machine Learning

From the Real World to the Model

Real-world domain

- O is a set of objects, C is a target variable.
- $\gamma : O \rightarrow C$ is the ideal target function for O.
- Task. Given some $o \in O$, determine the information $\gamma(o) \in C$.

Model domain

- X is a set of vectors (often called the *feature space*), C as before.
- $c : X \rightarrow C$ is the ideal predictor for X.
- Task. Given some $\mathbf{x} \in X$, determine its class or value $c(\mathbf{x}) \in C$.

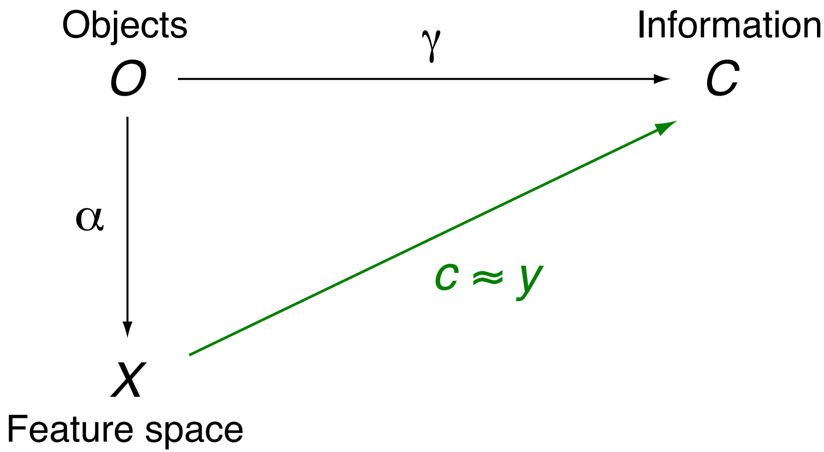
Example: Spam detection

- O is a set of emails, $C = \{$“spam”, “no spam”$\}$.
- X represents an email’s distribution of words.
- γ is a human expert on spam, c is unknown.
- Task. Given an email, is it spam or not?

https://datenschutz.org

Machine Learning

Overview of the Concepts



Notation

γ	Unknown ideal target function for real-world objects
α	(Feature) Mapping function
c	Unknown ideal predictor for vectors from the feature space
y	Machine learning model to be learned
$c \approx y$	c is approximated by y (based on a set of instances)

Machine Learning

How to Learn

Learning types

- Machine learning differs in terms of what kind of patterns are learned as well as to what kind of data it is applied to.
- Major types. *Supervised* and *unsupervised* learning
  They are most important for NLP and in the focus of this course.

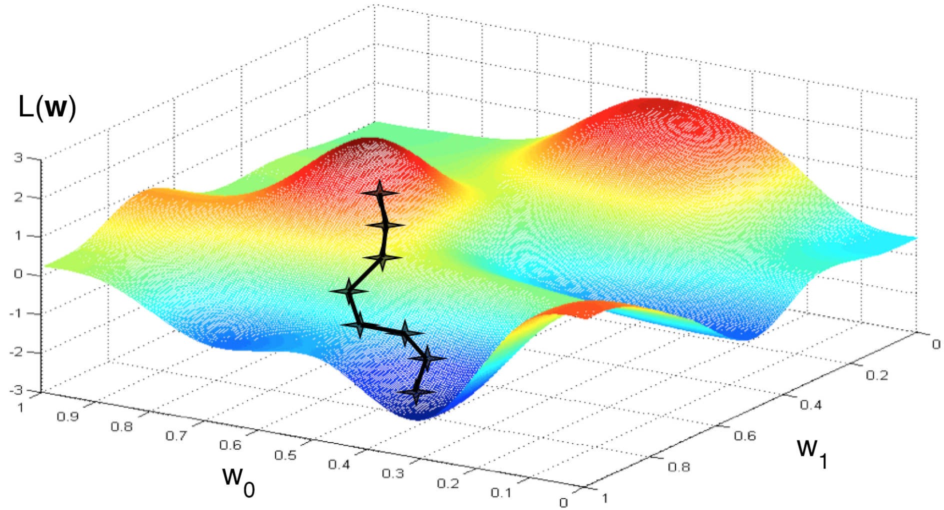
Major types in a nutshell

- Supervised. Derive a model from patterns in annotated training data (where the ground truth is known).
- Unsupervised. Derive model from unannotated data (no ground truth).

Learning algorithms

- Algorithms differ in terms of what patterns can be found, how they are represented, and how models are optimized.
- Some algorithms are discussed in detail in later lecture parts.

Machine Learning
Optimization



Training

- A learning algorithm incrementally creates candidate models y.
- y defines weights $\mathbf{w}$ for processing vectors $\mathbf{x}$.
  Not all learning algorithms assign weights explicitly.
- On a training set, y is tested against a *loss function* $\mathcal{L}(\mathbf{w})$.
  The loss function is a cost function that reflects the performance measure.
- Based on the loss, $\mathbf{w}$ is adapted to create the next model y'.
- For this, an *optimization procedure* is used, such as gradient descent.

Hyperparameter optimization

- Most learning algorithms have *hyperparameters* whose best values depend on how well the training set reflects the real distribution.
- Hyperparameters need to be optimized against a validation set.

Supervised Learning

Supervised (machine) learning

- A learning algorithm builds a model y on *known* training data, i.e., pairs of a vector $\mathbf{x}^{(i)}$ and the associated output information $c(\mathbf{x}^{(i)})$.
- y can then be used to predict output information for unknown data.

Why "supervised"?

- The learning process is guided by instances of correct predictions.

Supervised classification vs. regression

- Classification. Assign a nominal class to an instance.
- Regression. Predict a numeric value for an instance.

Manifold applications in NLP

- Classification. Standard technique for any text classification task, for extracting relations between entities, and similar
- Regression. Used to predict scores, ratings, probabilities, ...

Supervised Learning

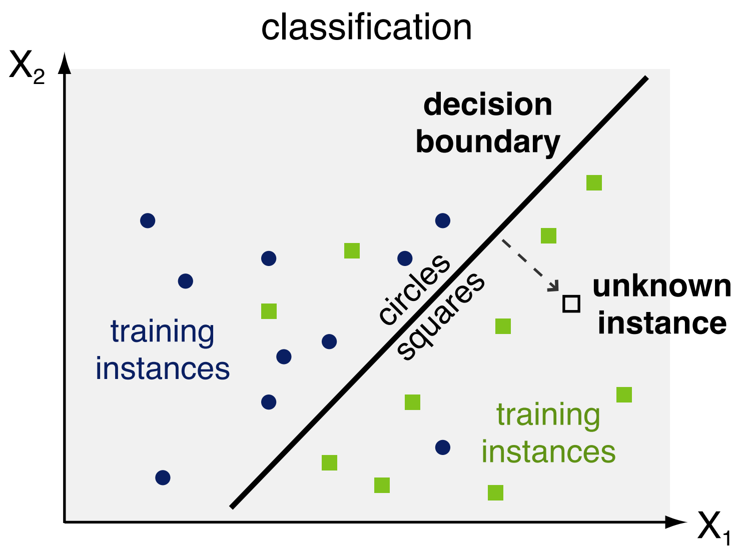
Classification

Classification

- The task to assign an object to the most likely of a set of two or more predefined discrete classes

Supervised classification

- An optimal decision boundary y is sought for on training vectors X with known classes C.
- The boundary decides the class of unknown instances.



Binary vs. multiple-class classification

- Binary classifiers separate the instances of two classes.
- Multiple classes are handled through multiple binary classifiers, e.g., using one-versus-all classification.

Supervised Learning

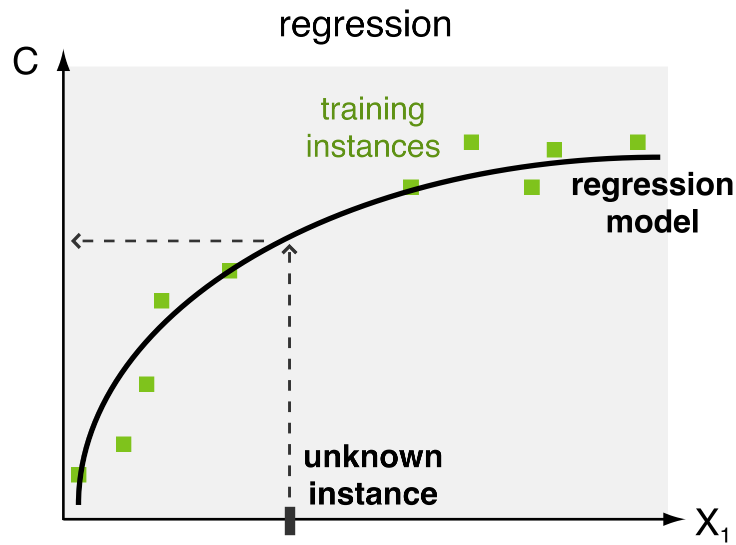
Regression

Regression

- The task is to assign a given object to the most likely value of a real-valued, continuous target variable

Supervised regression

- An optimal regression function y is sought for on training vectors X with known values C.
- The function decides the value of unknown instances.



Linear regression models

- Only constants and parameters multiplied by independent variables:

$$y(\mathbf{x}) = w_0 + w_1 \cdot x_1 + \ldots + w_m \cdot x_m \quad \text{with } x_i \in \mathbf{x} \text{ and } w_i \in \mathbf{w}$$

Unsupervised Learning

Unsupervised (machine) learning

- A model y is derived from vectors X only, without output information.
- y reveals the organization and association of input data.
- Techniques. Clustering, autoencoders, prinicipal component analysis, ...

  The focus is on clustering here.

Clustering

- The grouping of a set of instances into a possibly but not necessarily predefined number of classes (aka *clusters*).

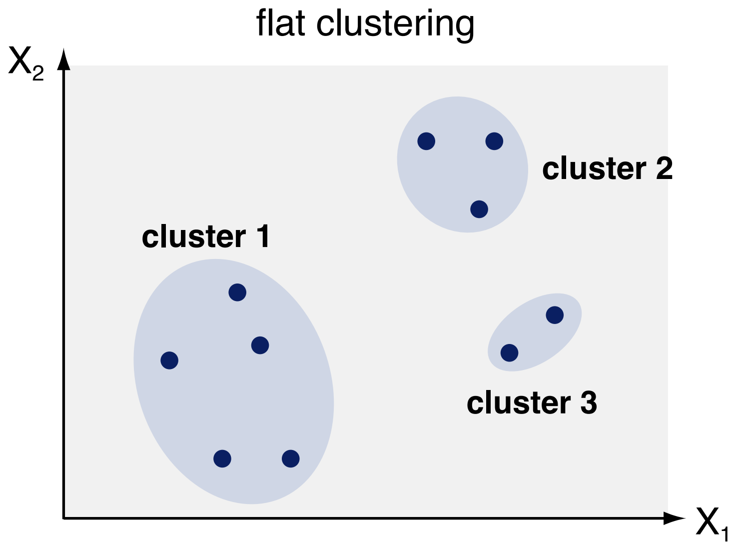
  The meaning of a class is usually unknown in advance.
- Hard clustering. Each instance belongs to a single cluster.
- Soft clustering. Instances belong to each cluster with some weight.

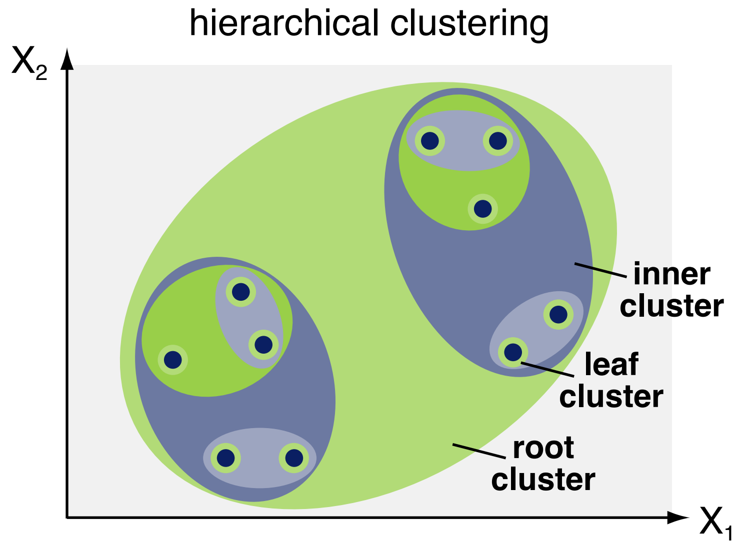
Applications in NLP

- Detection of texts with similar properties, mining of topics, ...

Unsupervised Learning

Flat vs. Hierarchical Clustering





Flat clustering

- Group instances into a (possibly predefined) number of clusters.
- No associations between the clusters are specified.

Hierarchical clustering

- Create a binary tree over all instances.
- Each tree node represents a cluster of a certain size.

Data Mining

Data Mining

Data mining

- The inference of new (or "hidden") output information of specified types from typically huge amounts of input data
- Data mining hence deals with prediction tasks.

Data mining in a nutshell

- Representation. Map data to instances of a defined representation.
- Machine learning. Find statistical patterns in the instances that are relevant to the prediction task (aka *training*).
- Generalization. Apply the found patterns to infer new information from unseen data (aka *prediction*).

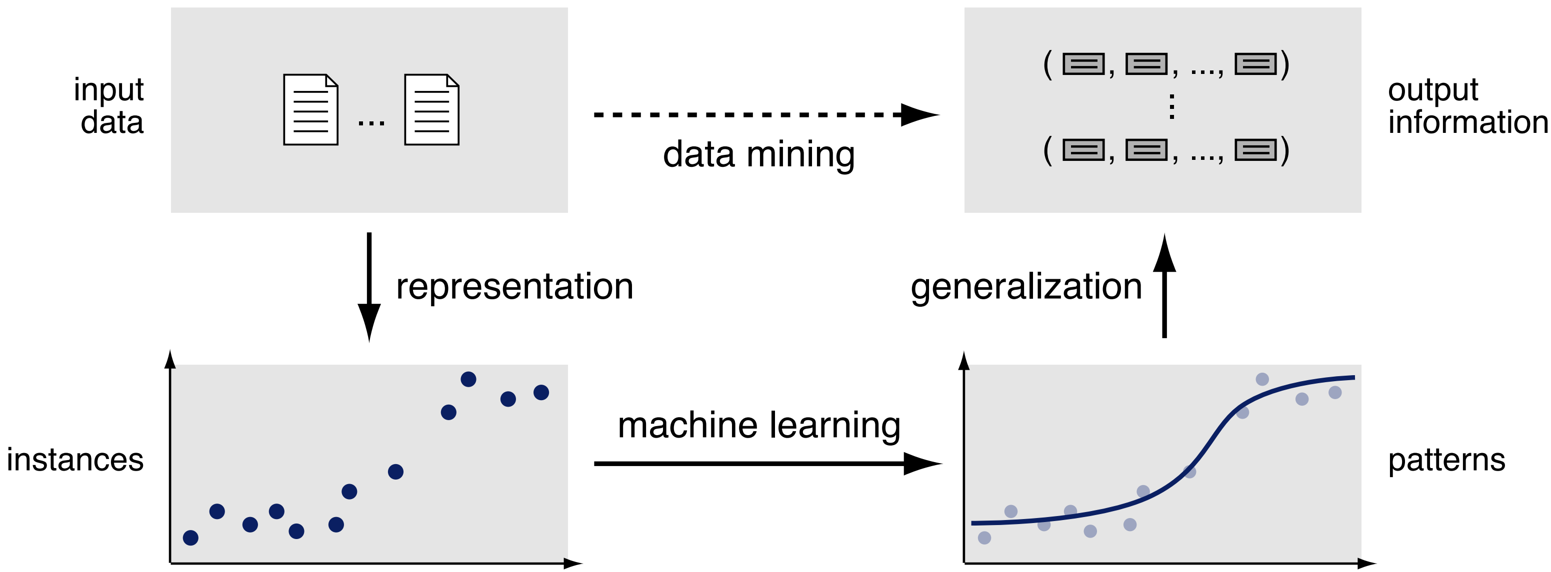
NLP as data mining

- Input data. A text corpus, i.e., a collection of texts to be processed
- Output information. Annotations of (spans of) the texts, or new texts
- The *representation* step is what makes NLP specific.

Data Mining

Process

The data mining process

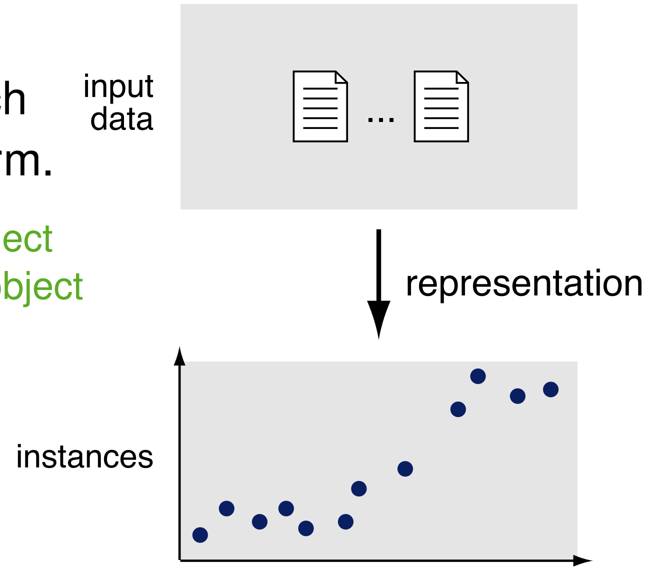


Data Mining
Representation

Instance representation

- Given a task to predict some type C, each object $o_i \in O$ is mapped to a common form.

Sentiment analysis: each text (span) is one object
Entity recognition: each candidate entity is one object



Feature representation

- Map o_i to a (sparse) vector of values $\mathbf{x}^{(i)}$.
- This is done with a function $\alpha : O \rightarrow X$.
- What X covers is defined by humans.

(Neural) Embedding representations

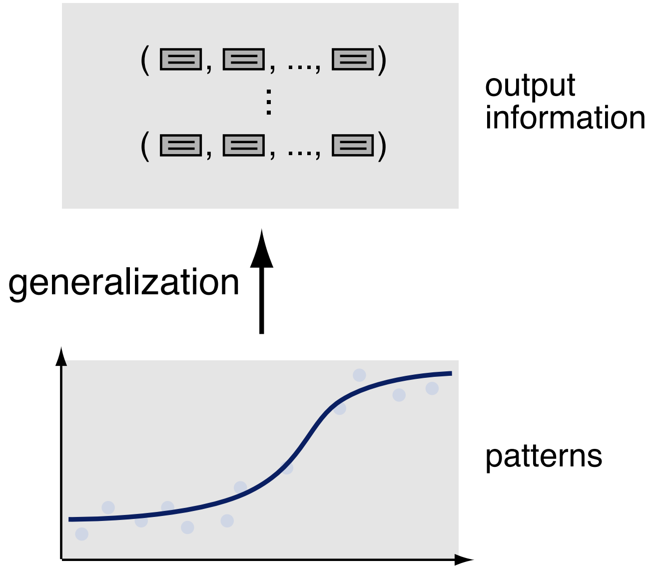
- Map o_i to one or more (dense) vectors of values $\mathbf{v}^{(i)}$.
- This is learned from the distributional representation of inputs.
- Neural models obtain features from vectors via self-learned functions.

Data Mining

Generalization

Generalization

- Application of the learned model y to unseen data to infer new information.
- How well y generalizes depends on how well it fits the target function γ.
- Generalization is mainly decided by the training process (see above).



Bias in training

- The training process explores a large space of models $Y = \{y_0, y_1, \ldots\}$.
- An important training decision is how much to bias the process wrt. the complexity of the model y to be learned.

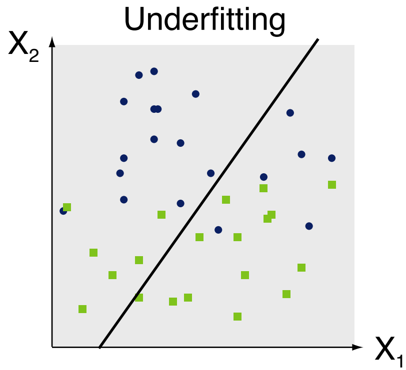
Data Mining

Underfitting and Overfitting

Simple vs. complex models

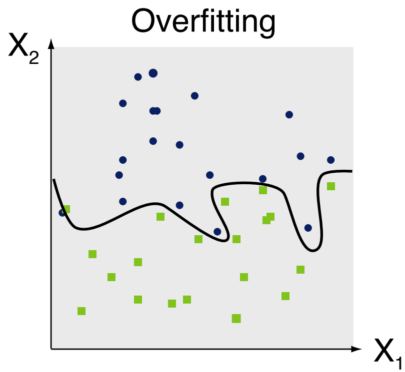
- Simple. Induce high bias to avoid noise; may underfit the input data
- Complex. Induce low bias to fit the input data well; may capture noise

  Simple models may, e.g., be linear functions, complex models high polynomials.



Underfitting (too high bias)

- A model y generalizes too much, not capturing all relevant properties of the training data.
- y is too simple and will have limited effectiveness.



Overfitting (too high variance)

- A model y captures both relevant and irrelevant properties of the training data.
- y is too complex and will thus not generalize well.

Data Mining

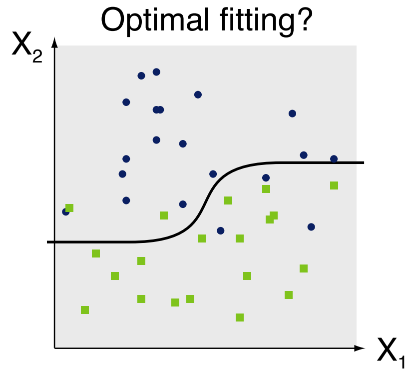
Optimal Fitting and Regularization

Avoiding underfitting and overfitting

- The best way to avoid both is to achieve an *optimal fitting*.
- Overfitting can also be countered through *regularization*.

Optimal fitting

- A model y perfectly approximates the complexity of γ based on the training data.
- In general, the right complexity is unknown.



Regularization

- Refrain from making y complex, unless it significantly reduces the loss.
- This is done by adding a term to the loss function that forces the feature weights to be small.

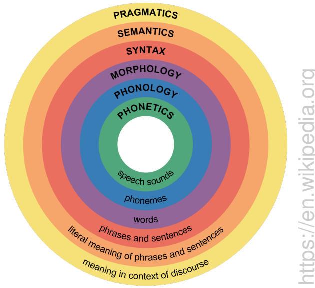
  More on regularization in a later part of this course.

Conclusion

Conclusion

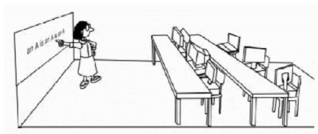
NLP and linguistics

- Linguistic knowledge from phonetics to pragmatics
- Ambiguity exists across linguistic levels
- Techniques from simple rules to language models



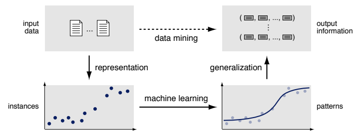
NLP and machine learning

- Learning target functions in analysis or synthesis tasks
- Inferring models from statistical patterns in training sets
- Focus here on supervised and unsupervised learning



NLP and data mining

- Inference of output information from huge input data
- Representation, machine learning, and generalization
- NLP can be seen as data mining on text



References

Much content taken from

- **Ng (2018).** Andrew Ng. Machine Learning. Lecture slides from the Stanford Coursera course, 2018. https://www.coursera.org/learn/machine-learning
- **Stein and Lettmann (2010).** Benno Stein and Theodor Lettmann. Machine Learning. Lecture Slides, 2010. https://webis.de/lecturenotes/slides.html#machine-learning
- **Wachsmuth (2015).** Henning Wachsmuth. Text Analysis Pipelines — Towards Ad-hoc Large-scale Text Mining. LNCS 9383, Springer, 2015.
- **Wachsmuth (2023).** Henning Wachsmuth. Introduction to Natural Language Processing. Lecture slides, 2023. https://www.ai.uni-hannover.de/en/teaching/courses/inlp
- **Witten and Frank (2005).** Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, San Francisco, CA, 2nd edition, 2005.

Other references

- **Mitchell (1997).** Tom M. Mittchell. Machine Learning. McGraw Hill, 1997.
- **Samuel (1959).** Arthur Samuel. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development, 44:206?226, 1959.