Chapter 04

Data Analysis Overview

Dr. Steffen Herbold
herbold@cs.uni-goettingen.de

Outline

- Overview
- Foundational Concepts
- Summary

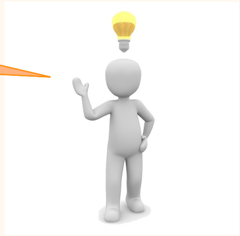
No Free Lunch Theorem

- d_m^y ordered sets of size m of cost values for $y \in Y$
- $f: X \to Y$ a function that is optimized
- $P(d_m^y | f, m, a)$ the conditional probability of getting d_m^y by m times running algorithm a on the function f

Theorem: For any pair of algorithms a_1 and a_2

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

All algorithms are equal



Implication of the NFL Theorem

"if an algorithm does particularly well on average for one class of problems then it must do worse on average over the remaining problems"

David H. Wolpert and William G. Macready: No Free Lunch Theorems for Optimization, IEEE Transactions on Evolutionary Computation, 1(1):67-82

No Silver Bullet

→There is no „one way“ to do data analysis
- But there are some standard techniques that often perform well

- Many factors influence the suitable techniques
  - Data
  - Problem to be solved
  - Available resources
  - …

→ Broad portfolio of data analysis techniques required

Categories of Data Analysis Techniques

Category	Techniques Covered	Problem to be solved
Association Rules	Apriori	Relationships between items
Clustering	K-Means Clustering DB Scan	Grouping of similar items Identification of structures
Classification	K-nearest Neighbor Decision Trees Random Forests Logistic Regression Naive Bayes Support Vector Machines Neural Networks	Assignment of labels to objects
Regression	Linear Regression Ridge Lasso	Relationship between outcome and inputs
Time Series Analysis	ARMA	Identification of temporal structures Forecasting of temporal processes
Text Mining	Bag-of-Words Stemming/Lemmatization TF-IDF	Analysis of textual data

Outline

- Overview
- **Foundational Concepts**
- Summary

Machine Learning

- Definition after Tom M. Mitchel [2]:
  - A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

- Relation to the data analysis techniques
  - Experience E: our data
  - Task T: clustering/association mining/classification/…
  - Performance Measure P: depend on tasks

T. M. Mitchel: Machine Learning, McGraw Hill, 1997

Description of a „Whale“ Picture

- How would you describe this picture with general concepts?

Features of Objects

Object

Feature map
$\phi: O \rightarrow \mathcal{F}$

Features

hasFin	=	true
shape	=	oval
colorTop	=	black
colorBottom	=	white
background	=	blue

- O is the object space
- ϕ is the feature map
- $\mathcal{F}$ is the feature space
  - $\mathcal{F} = \{\phi(o), o \in O\}$
- Example:
  - Five-dimensional space with dimensions as above
  - $\phi(\text{"whalepicture"}) = (true, oval, black, white, blue)$

Scales of Features

- Steven‘s levels of measurement

Categorical

Scale	Property	Allowed Operations	Example
Nominal	Classification or membership	$=, \neq$	Color as „black“, „white“ and „blue“
Ordinal	Comparison or levels	$=, \neq, >, <$	Size in „small“, „medium“, and „large“
Interval	Differences or affinities	$=, \neq, >, <, +, -$	Dates, temperatures, discrete numeric values
Ratio	Magnitudes or amounts	$=, \neq, >, <, +, -, \cdot, /$	Size in cm, duration in seconds, continuous numeric values

S. S. Stevens: On the Theory of Scales of Measurement, Science, 103(2684):677-680

Encoding Categorical Features

- Many algorithms can only work with numeric features

- Encode categorical features as binary numeric features
  - Example: $x \in \{\text{small}, \text{medium}, \text{large}\}$
  - Encode as three variables $x^{small}, x^{medium}, x^{large}$
  - $x^{small} = \begin{cases} 1 & if\ \ x = \text{small} \\ 0 & otherwise \end{cases}, \ldots$
  - Can also use one variable less, remaining case is encoded by all zeros

- This is called *One-Hot-Encoding*

Training Data

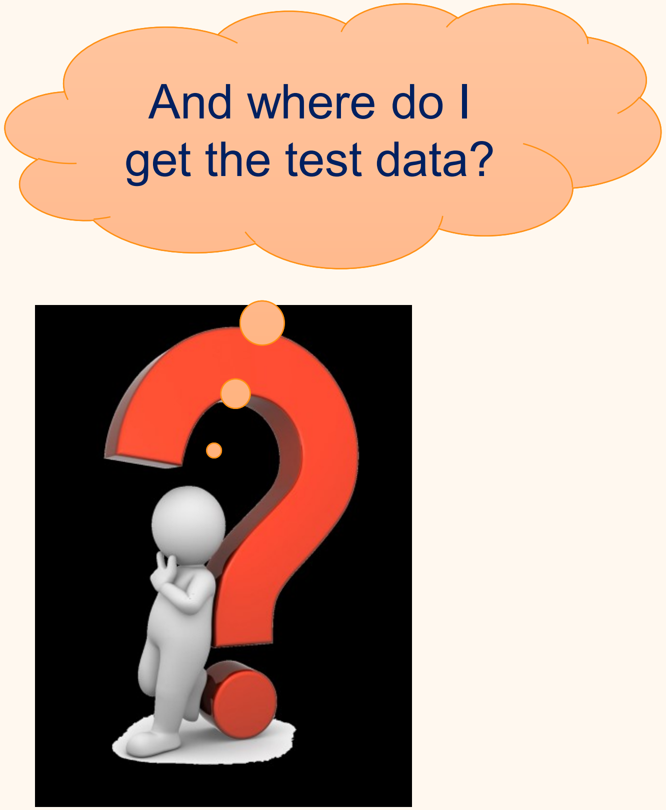
- *Instances* of objects described by their features

		$\phi(o)$			
hasFin	shape	colorTop	colorBottom	background	**value of interest**
true	oval	black	black	blue	**whale**
false	rectangle	brown	brown	green	**bear**
…	…	…	…	…	**…**

- *Supervised* learning if the value of interest is known
  → Classification, regression
- Otherwise *unsupervised learning*
  → Clustering, Association Rule Mining

The Test Data

- Data for the evaluation of analysis results
  - Same distribution as training data
- Training data ≠ Test data
  - Evaluate generalization
  - Avoid overfitting
    - Analysis results only valid on training data
    - Different and not working on unseen data
- Test data often difficult to obtain



Hold-out Data

- Data not used for training at all

- Commonly used hold out data sizes
  - 50% of all data
  - 33% of all data
  - 25% of all data in case a validation set is used

Depends a lot on available data!

- Example:
  - Nine months of customer transactions available
  - First six months as training data
  - Last three months as test data

k-fold Cross Validation

- Create k partitions of available data
- One partition for testing, all others for training
- Estimate performance by averaging over the iterations

	Test data	Training data			
Iteration 1	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
Iteration 2	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
Iteration 3	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
Iteration 3	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5
Iteration 5	Partition 1	Partition 2	Partition 3	Partition 4	Partition 5

Available data

Outline

- Overview
- Foundational Concepts
- **Summary**

Summary

- No generic algorithm for all problems
- Objects are described by features
- Features are used for learning about objects
- Data usually split into different sets for different purposes