university of groningen

Introduction to Intelligent Systems

Learning (Part I)

Introduction: Unsupervised Learning

Supervised Learning

Prototype based learning

Unsupervised Learning : Vector Quantization (VQ)

competitive learning / K-means algorithm

relation to clustering

Supervised Learning : Learning Vector Quantization (LVQ)

relevance learning: adaptive distance measures

Validation Generalization performance

Cross-validation and related schemes

Model selection, over-fitting etc.

A clustering problem

Unsupervised Learning







data driven **feature selection** (color, shape,…) defines **similarity**














Unsupervised Learning

some issues in unsupervised learning:

- find a reasonable **criterion**, set of **features** to group data
- how many **clusters** or groups ? (model selection)
- how to define **similarity** / dissimilarity ?
- what is the aim ?

  reduction: represent large amount of data by few prototypes

  compression: represent all data by few features

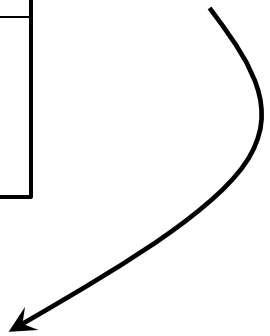
  clustering: identify groups of similar data

- aim may define **cost function** or optimization strategy = **learning**

  but there is no “right” or “wrong”, no feedback

Supervised) Learning from Examples

input (data)	4	13	6	11	8
output (class label)	0	1	0	1	0



question	7
prediction	?

consistent hypotheses:

odd numbers → 1

numbers > 10 → 1

Supervised Learning from Examples

input (data)	4	13	6	11	8	2
output (class label)	0	1	0	1	0	1

question	7
answer	?

consistent hypotheses:

~~odd numbers → 1~~

~~numbers > 10 → 1~~

prime numbers → 1

space of consistent hypotheses shrinks with increasing # of examples

Feature Selection

input (data)

output (class label)

				
pear	apple	apple	pear	apple

question

prediction

apple

selected feature: color

Feature Selection

input (data)						
output (class label)	pear	apple	apple	pear	apple	apple

question	
answer	pear

selected feature: shape

some issues in supervised learning:

- complexity of the **hypothesis space** (model selection)
- parameterization of hypothesis

  e.g. neural networks, prototype based classifiers, …
- how to handle **noise**, e.g. wrong *labels* in example data
- suitable **similarity** / **distance** measures
- construction / selection of **relevant features**
- the aim is clear:

  correct prediction of novel input data, i.e. **generalization**
- difficult: optimization of generalization ability

  → heuristic criteria, cost functions, **learning algorithms**
- **validation**: can we test/predict performance / generalization ability

  with respect to new unseen data ?