

Primer on Machine Learning

attilagk

Lazy8

Contents

Overview

Concepts

Demo: NY/SF Home

Machine Learning is Everywhere

- ▶ pattern detection and recognition (iphone touch ID, face ID, word autocomplete, speech to text)
- ▶ history based recommendation (youtube, facebook, google search, amazon,...)
 - ▶ products for customers
 - ▶ customers for providers
- ▶ email filtering and classification (gmail)

Machine Learning and Artificial Intelligence

1950s *[getting] machines to exhibit behavior,
which if done by humans, would be
assumed to involve the use of intelligence¹*

now *computational methods to automatically
learn and to improve with experience²*

ML³ statistical (“statistical learning”)
AI analytical (knowledge, logic)

¹Arthur Samuel, 1983

²<http://www.mlplatform.nl/what-is-machine-learning/>

³may mean Maximum Likelihood: abbrev. not widely used

Machine Learning Now

- ▶ big data
 - ▶ data science, data mining, ...
 - ▶ myth: machine learning needs big data⁴
- ▶ fast computers
- ▶ emerging new methods
 - ▶ deep learning, reinforcement learning, ...

⁴see NY/SF homes toy data

General Outline

1. formulate task (EN, ST)
2. collect data (EN, ST)
3. “engineer” data (EN, PR, DSS, HPC)
4. modeling (ST, EN)
5. fit/learn/train model(s) on data (LIB, PR, HPC)
6. validate, select (LIB, ST, PR, HPC)
7. apply model to test data (LIB)
8. interpret results (EN, ST)
9. report (DOC)

skill set

EN	expert knowledge
ST	statistics
PR	programming
DSS	domain spec. software
HPC	high performance computing
LIB	ML libraries ^a
DOC	L ^A T _E X, Web

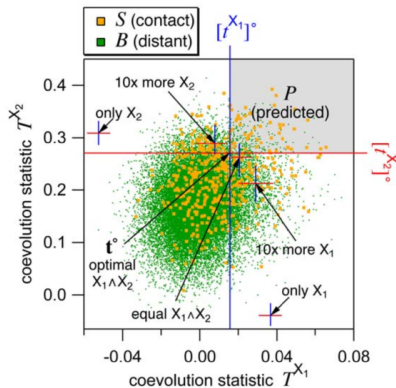
^aPython, R, Java, Julia, Scala, Matlab

My Story with Machine Learning

Myth: it's like cooking

skill	2006	2017
expert knowledge	?	?
statistics	-	+
programming	-	+
domain spec. softw.	-	+
high perform. comp.	-	?
ML libraries	-	?
L ^A T _E X, Web.	-	+

PLoS One. 2012;7(5):e36546.



Contents

Overview

Concepts

Demo: NY/SF Home

Tasks

1. supervised learning: training *and* test data
 - ▶ prediction, classification
 - ▶ pattern recognition
 - ▶ business, medical, ... predictions & decisions
2. unsupervised learning: *only* training data
 - ▶ structure discovery
 - ▶ social, biol., tech. networks, associations,...
 - ▶ probabilistic expert systems
 - ▶ hypothesis testing, feature subset selection
 - ▶ research, marketing
 - ▶ matrix completion (imputation)
 - ▶ recommendation systems

The “Home” Data for Classification

Useless except for demonstration

observation i home	input features / variables				output
	x_{i1}	x_{i2}	...	x_{ip}	y_i
	price/sqft	elevation	...	beds	city

training data

1	999	10	...	2	NY
2	1939	0	...	2	NY
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
491	764	163	...	1	SF
492	762	216	...	3	SF

test data

493	1800	120	...	2	?
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots

Home Data Overview

`figures/2017-10-17-classification/splom-1.pdf`

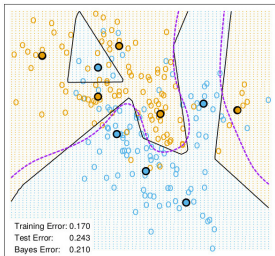
Informative Features

2 input features: 2D plots

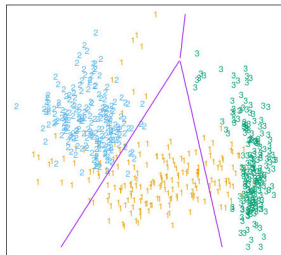
figures/2017-10-17-classification/elevation-price-per-sqft

Models for Classification

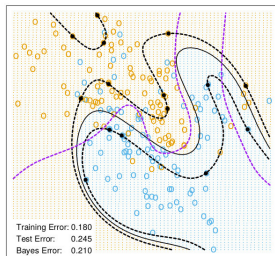
K-means classifier



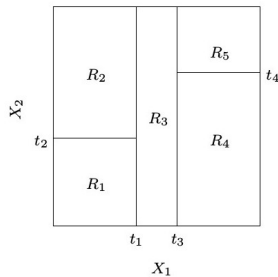
generalized linear regression



support vector machine



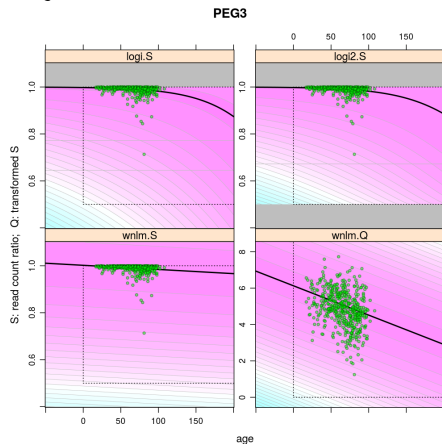
decision tree (inferior)



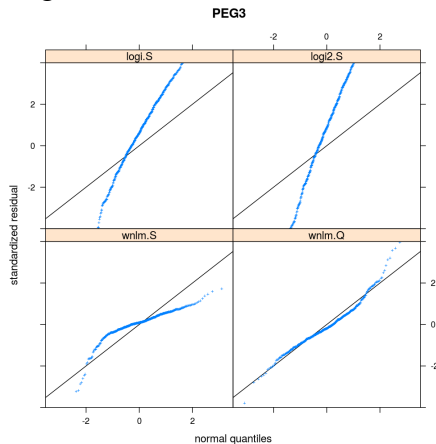
Which Model to Select? The One That Fits the Best!⁵

Information theory, optimality, model ensemble

subjective evaluation



diagnostics



⁵should also be: simple, interpretable, fast to fit,...

Contents

Overview

Concepts

Demo: NY/SF Home

The “Home” Data for Classification

Useless except for demonstration

observation i home	input features / variables				output
	x_{i1}	x_{i2}	...	x_{ip}	y_i
	price/sqft	elevation	...	beds	city

training data

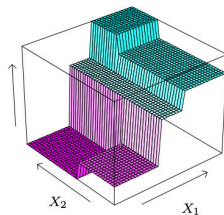
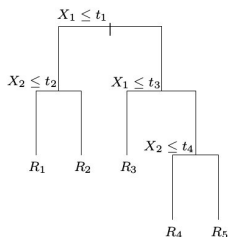
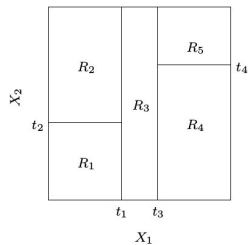
1	999	10	...	2	NY
2	1939	0	...	2	NY
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
491	764	163	...	1	SF
492	762	216	...	3	SF

test data

493	1800	120	...	2	?
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots


Decision Tree is Ideal for Demo

Tree Grows as We Partition Recursively



Why R?

- ▶ created by and for biostatisticians
- ▶ functional programming paradigm
- ▶ open source
- ▶ mature
- ▶ lots of machine learning packages
- ▶ R2D3⁶

⁶<http://www.r2d3.us/visual-intro-to-machine-learning-part-1/> 

Fitting and Interpreting Decision Tree

Using the rpart R library

figures/2017-10-17-classification-tree.pdf

Optimality Criteria

figures/2017-10-17-classification/complexity-1.pdf

Prediction

What We Sidestepped Here

- ▶ real World task
- ▶ data collection/engineering
- ▶ model selection
- ▶ validation, performance

Machine Learning and You

- ▶ doing it
 - ▶ Hello World! is easy but useless
 - ▶ obtaining skills takes years
 - ▶ rewarding
- ▶ understanding it
 - ▶ learn concepts not cooking
 - ▶ collaboration, interpretation