CSCE 590-1: From Data to Decisions with Open Data: A Practical Introduction to AI

Lecture 13: Time Series Analysis

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

23RD FEB, 2021

Carolinian Creed: "I will practice personal and academic integrity."

Organization of Lecture 13

- Introduction Segment
  - Review of Quiz2
  - Recap of Lecture 12
-
- Main Segment
  - AutoAI paper
  - Time Series Analysis
- Concluding Segment
  - About Next Lecture – Lecture 14
  - Ask me anything

Introduction Segment

Recap of Lecture 12

- Generating explanations
  - LIME
  - AIX 360
  - Which methods work under what conditions?

- AutoAI
  - For removing mundane steps
  - Improving model performance

Recap Quiz 2

- Classification
- Clustering
- Bonus question

Main Segment

Auto AI Paper

- Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools
  - https://arxiv.org/abs/1908.05557, 2019

Reading Group Allocation

a	Brendan	Curran
b	Sarayu	Das
c	Marilyn	Gartley
d	Vedant	Khandelwal
e	Vishal	Pallagani
a	Avineet Kumar	Singh
b	Ahad Hasan	Tanim
c	James	Thompson
d	Rohit	Naini
e	Terric	Taylor

Auto-AI Tools Compared

Tool	Platform	Input data sources		Data pre-processing	Data types detected					Feature engineering				ML Tasks		Model selection and Hyperparameter optimization					Quick start / early stop			Model evaluation / Result analysis/ Visualization		
		Spreadsheet datasets	Image, text		Numerical	Categorical	Datetime	Time-series	Other (Hierarchical types) (7*)	Datetime, categorical processing	Imbalance, missing values	Feature selection, reduction	Advanced feature extraction (8*)	Supervised learning (9*)	Unsupervised learning (10*)	Ensemble	Genetic algorithm	Random search	Bayesian search	Neural architecture search	Quick finding of starting model	Allow maximum limit search time	Restrict time consuming combination of components	Model dashboard	Feature importance	Model explainability and interpretation, and reason code (11*)
TransmogrifAI	Apache Spark	Y	N	Y(*)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N	N			Y	Y	
H2O-AutoML	AWS, GCP, Azure	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	N	Y	N	Y	N	Y	N	N	N	Y	Y	Y	Y	Y
Darwin (+)	GCP	Y	N	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y	N	Y	Y	Y
DataRobot (+)	AWS, GCP, Azure	Y	Y	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	Y(12*)	Y		Y	Y	Y
Google AutoML (+)	Google Cloud	N	Y	Y						N	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Auto-sklearn		Y	N	N	N	N	N	N	N	Y(2*)	Y	Y	Y	Y	N	Y	N	Y	Y	N	Y	Y	Y	Y	Y	Y
MLjar (+)	MLJAR Cloud	Y(3*)	N	Y	Y	Y	N	N	N	Y	Y(4*)	N	N	Y(5*)	N	Y	N	Y	N	N	N	N	N	Y	Y	N
Auto_ml		Y	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N	N	N	N	Y	Y	Y
TPOT		Y	N	N	N	N	N	N	N	N	Y	N	Y	Y	N	Y	Y	N	N	N	N	Y	N	Y	Y	N
Auto-keras		Y	Y	N	N	N	N	N	N	N	Y	Y	N	Y	N	N	N	Y	Y	Y	Y	Y	N	Y	N	Y
Ludwig		Y	Y	Y(*)	Y	Y	N	Y	Y	N	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	N	N	Y	Y	N
Auto-Weka		Y	N	N	Y	Y	N	N	N	N	Y	Y	N	Y	N	Y	N	Y	Y	N	N	Y	Y	Y	N	N
Azure ML (+)	Azure	Y	Y	Y(6*)	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y	N	Y	Y	N		Y	Y	Y	Y	
H2O-Driverless AI (+)	AWS, GCP, Azure	Y(3*)	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	Y	Y	Y	Y

Fig. 2. Comparison table of functionality for AutoML tools. (+): commercialized tools; (*): the function is not very stable, it fails for some datasets; (2*): categorical input must be converted into integers; (3*): datasets have to include headers; (4*): missing values must be represented as NA; (5*): multiclass classification not provided; (6*): need some users' input for dataset description such as column types; (7*): ability to detect primitive data types and rich data types such as: text (id, url, phone), numerical (integer, real); (8*): advanced feature processing: bucketing of values, removing features with zero variance or features with drift over time; (9*): supervised learning includes binary classification, multiclass classification, regression; (10*): unsupervised learning includes clustering and anomaly detection; (11*): model interpretation and explainability refers to techniques such as LIME, Shapley, Decision Tree Surrogate, Partial Dependence, Individual Conditional Expectation, Lift chart, feature fit, prediction distribution plot, accuracy over time, hot spot and reason codes; (12*): confirmed by a company spokesperson, we could not find public documentation at the time of publication; In a few empty cells, it is not clear that the functionality is provided from documentations of the tools, to the best of our knowledge.

Source: Ava: From Data to Insights Through Conversation
CIDR 2017

Suggested Reading

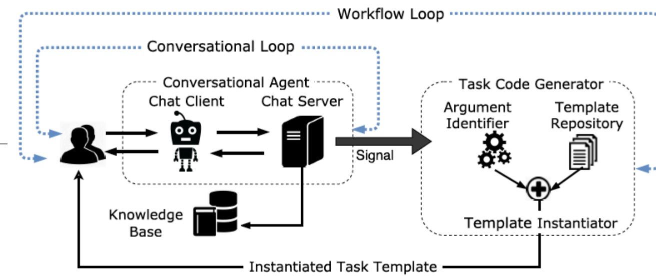
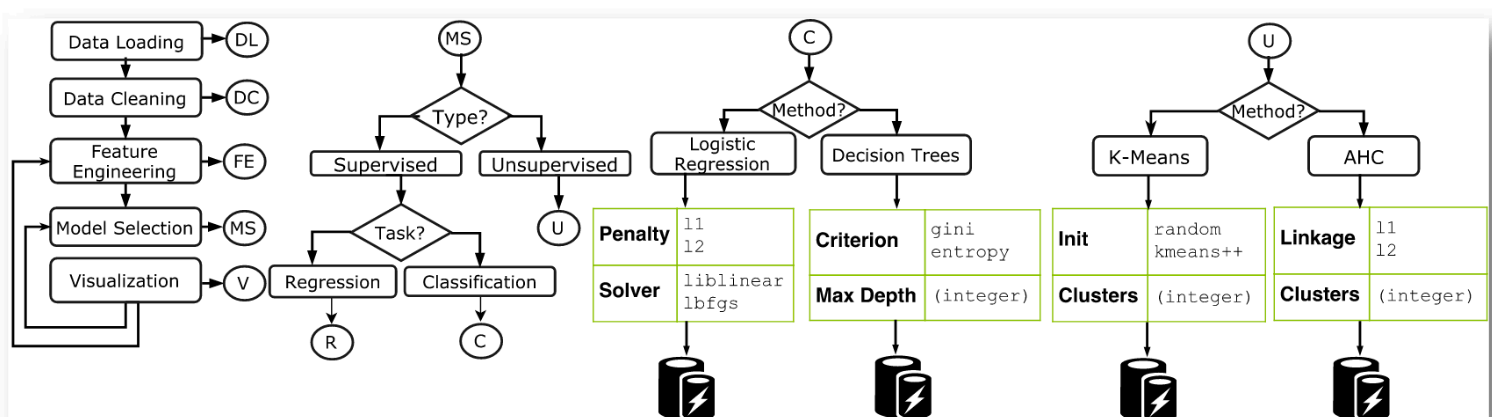


Figure 4: Overview of Ava's control flow architecture.

The Ava Storyboard Concept …



Time and Representation

What is the Time ?

Time: year, month (or week), day, hour, minute, second, and fraction of second

- Year: [YYYY}
- Month/ Week
  - week 01:
    - the week with the year's first Thursday in it (the formal ISO definition),
    - the week with 4 January in it,
    - the first week with the majority (four or more) of its days in the starting year, and
    - the week starting with the Monday in the period 29 December – 4 January.
- Problem of multiple formats

Reference: https://en.wikipedia.org/wiki/ISO_8601

Basic Time Operations

- Converting time to time zones
  - By default, refers to local time
- Time difference
- Code example
  - https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l13-timeseries/1-Basic%20Time.ipynb

Time Difference - Algebra

precedes	meets	overlaps	finished by	contains	starts	equals	started by	during	finishes	overlap-ped by	met by	preceded by
p	m	o	F	D	s	e	S	d	f	O	M	P

Table 1. Allen's thirteen basic relations

Relation		Converse	
precedes	(p)	(P)	preceded by
meets	(m)	(M)	met by
overlaps	(o)	(O)	overlapped by
finished by	(F)	(f)	finishes
contains	(D)	(d)	during
starts	(s)	(S)	started by
equals (e)			

Table 2. Converses of Allen's basic temporal relations

Allen's Algebra –

Allen, James F. "Maintaining knowledge about temporal intervals". *Communications of the ACM* **26**(11) pp.832-843, Nov. 1983

Source: https://www.ics.uci.edu/~alspaugh/cls/shr/allen.html

Time Difference – Algebra - Example

Relation			Converse
precedes	(p)	(P)	preceded by
meets	(m)	(M)	met by
overlaps	(o)	(O)	overlapped by
finished by	(F)	(f)	finishes
contains	(D)	(d)	during
starts	(s)	(S)	started by
equals (e)			

Table 2. Converses of Allen's basic temporal relations

"John was **not** in the room when I touched the switch to turn on the light”

- *a* be the time John was in the room,
- *b* be the time I touched the light switch, and
- *c* be the time the light was on.

a	(pmMP)		*b*
"John was in the room"	p	a b	"I touched the light switch"
	m	a b	
	M	a b	
	P	a b	
b	**(mo)**		***c***
"I touched the light switch"	m	b c	"The light was on"
	o	b c	

Table 3. Example "Turn on the light"

Source: https://www.ics.uci.edu/~alspaugh/cls/shr/allen.html

Time Series Analysis - Examples

- Long-term or trend movements: trend curve or line
- Cyclic movements or variations
- Seasonal movements or variations: recurring based on calendar
- Irregular or random variations: sporadic movement due to chance events

Series Analyses Examples

- Datasets
  - Energy consumption
  - COVID-19

- Code example
  - https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l13-timeseries/2-Time%20with%20Pandas.ipynb

Advanced Analysis

- Autoregressive Integrated Moving Average (ARIMA)
  - autoregressive models: AR(p)
  - moving average models: MA(q)
  - mixed autoregressive moving average models: ARMA(p, q)
  - integration models: ARIMA(p, d, q)
  - seasonal models: SARIMA(P, D, Q, s)

- Code example
  - https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l13-timeseries/3-ARIMA%20model.ipynb

Sources:
- https://en.wikipedia.org/wiki/Autoregressive_integrated_moving_average
- https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima.model.ARIMA.html

Beyond Statistical Analysis of Time

- Events analysis
  - Event: Label + timestamp
  - Detecting and processing event
- Causal reasoning
  - Establishing cause-and-effect
  - **Not** the same as correlation

Lecture 13: Concluding Comments

- Reviewed AutoAI paper
- Looked at time
  - Representation
  - Time difference analysis
  - Simple temporal analysis
  - Auto-regressive models
- A broad area for generating insights

Concluding Segment

About Next Lecture – Lecture 14

Lecture 14: Invited Lecture

- Understanding Timing Errors in Datasets
  - Sandeep Sandha, PhD Candidate, UCLA

    This lecture will introduce the different reasons for data timestamping errors present in the edge devices with a focus on Smartphones. We will discuss how the poor kernel design in Android is one of the main contributors to timestamping errors, with experiments showing the magnitude of error as high as 5 seconds across modern devices. To understand the impact of timestamping errors, the use-case of multimodal fusion for human activity recognition and cooking detection will be presented.
    Our analysis will highlight significant degradation in classifier accuracy due to the timing errors.

    As a solution approach to reducing the impact of timing errors on ML applications, we will introduce a new form of data augmentation called time-shift data augmentation. **As a motivational scenario, we will analyze the cooking dataset from the open-source domain.** Further, we will briefly look at the missing samples and the variable sampling rates in the data. Finally, we will conclude by presenting system techniques that can be deployed to enable robust data collection on the edge and introduce our open-source library used by UCLA's David Geffen School of Medicine to **timestamp data collected from the brain implants of patients**.

- **Keywords**: Sensor data, time analysis, AI on the edge