AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

CSCE 590-1: Trusted AI

Lecture 4: AI: Supervised Machine Learning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

31ST AUG 2021

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

Organization of Lecture 4

- Introduction Segment
  - Recap from Lecture 3
  - Project discussion
  - Coding guidelines
- Main Segment
  - Introduction to Machine Learning
  - Methods and tools
  - Data preparation
- Concluding Segment
  - About next lecture – Lecture 5
  - Ask me anything

Introductory Segment

Recap of Lecture 3

- Explored data in detail
  - By structure: structured, semi—structured, unstructured
  - By source: enterprise, social, open, sensor; our focus will be open
  - By types: text, audio, image, video
- Discussed 5-star data open standard
- Looked at data access via APIs
- Discussed internal data representation spectrum – glossary to knowledge graph

Project Discussion

- Information to be shared by students
  - Go to Google sheet: https://docs.google.com/spreadsheets/d/1VAX8ntb5zBQ-vOdsMHMhvEdwoaCZtuBaO4kJdkSA4eQ/edit?usp=sharing
  - Create a Google drive called “CSCE 590-1 Trusted AI (<YourName>)” and share with instructor: firstname.lastname@gmail.com
    - Put shared url in Column E
    - Put project title in column G
    - Create a folder in shared directory call project. Under it, have a Google doc called “Project Description”. In it, have the following as bullets with associated details: Problem, User, AI Method, Data, Reliability: Testing, Holding Human Values, Human-AI interaction. See next slide for framework and guidance on what to put.
  - Put Github location for your code in F
    - Create one repository
    - For each quiz, project, etc, create a sub-folder

Course Project

- **Framework**
  1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
  2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
  3. (AI Method) Think of what the solution will do to help the primary user
  4. (Data) Explore the data for a solution to work
  5. (Reliability:Testing) Think of the evaluation metric we should employ to establish that the solution will works? (e.g., 20% reduction in patient deaths)
  6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
  7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Minimum Coding Guidelines

UoSC (Gamecocks Coding Guidelines!)

- Have a project plan with details of tasks, deadlines and status
- Code should have:
  - Documentation
    - Report: specification of what it does, a test plan to how to see it works
    - Comments before every function of what it does
  - Organization: doc (documentation), data and code should in separate folders
  - Version control: the code and report should be in version control or git/ bitbucket, and be replicable
  - Test program: a stand-along program to demonstrate the code works
- A report or presentation should be created that is shared with instructor.

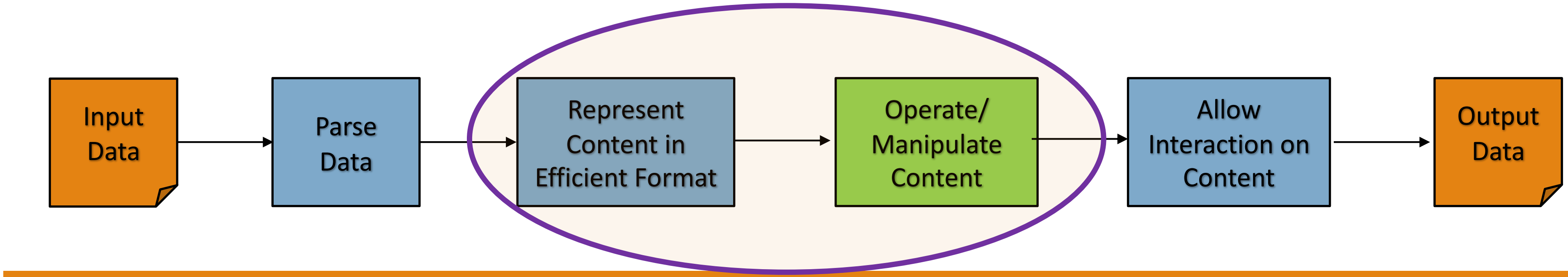
Good to have
Follow language-specific coding convention.

- General: https://en.wikipedia.org/wiki/Coding_conventions
- Python - PEP8, Java - https://www.python.org/dev/peps/pep-0008/

Main Segment

JOHN'S
WEATHER FORECASTING STONE

CONDITION	FORECAST
Stone is Wet	Rain
Stone is Dry	Not Raining
Shadow on Ground	Sunny
White on Top	Snowing
Can't See Stone	Foggy
Swinging Stone	Windy
Stone Jumping Up & Down	Earthquake
Stone Gone	Tornado

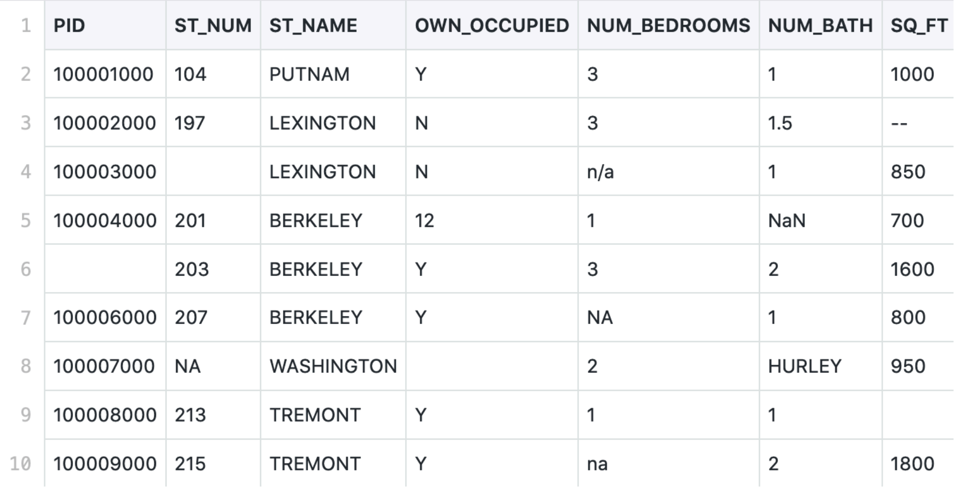


Input Data → Parse Data → Represent Content in Efficient Format → Operate/ Manipulate Content → Allow Interaction on Content → Output Data

Nomenclature

Column, Attribute, Feature

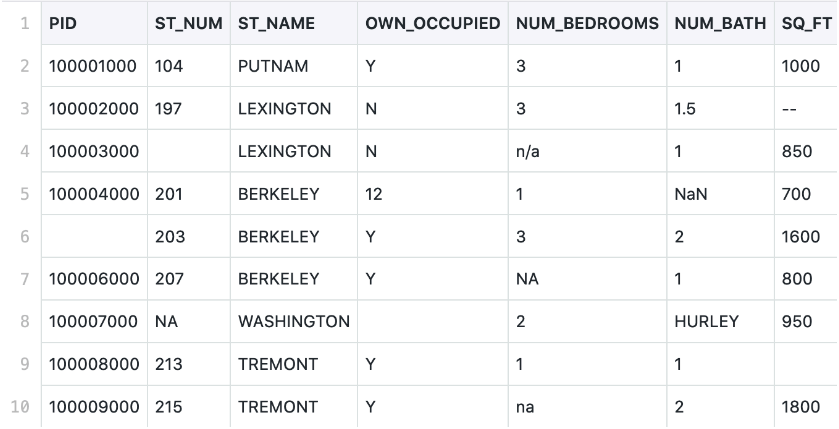
Row, Item



1	PID	ST_NUM	ST_NAME	OWN_OCCUPIED	NUM_BEDROOMS	NUM_BATH	SQ_FT
2	100001000	104	PUTNAM	Y	3	1	1000
3	100002000	197	LEXINGTON	N	3	1.5	--
4	100003000		LEXINGTON	N	n/a	1	850
5	100004000	201	BERKELEY	12	1	NaN	700
6		203	BERKELEY	Y	3	2	1600
7	100006000	207	BERKELEY	Y	NA	1	800
8	100007000	NA	WASHINGTON		2	HURLEY	950
9	100008000	213	TREMONT	Y	1	1	
10	100009000	215	TREMONT	Y	na	2	1800

Types of Attributes/ Columns

- Numeric: has number as value in computational sense; all mathematical functions are valid.
  - Example: SQ_FT
- Categorical: has distinct values
  - Nominal: each value is incomparable with other
    - Example: OWN_OCCUPIED, ST_NAME
  - Ordinal: the values can be ordered
    - Example: ST_NUM, NUM_BEDS
- Comment:
  - Q: what type is a binary variable?
  - A: depends on the semantics – nominal (gender), ordinal (number basements).



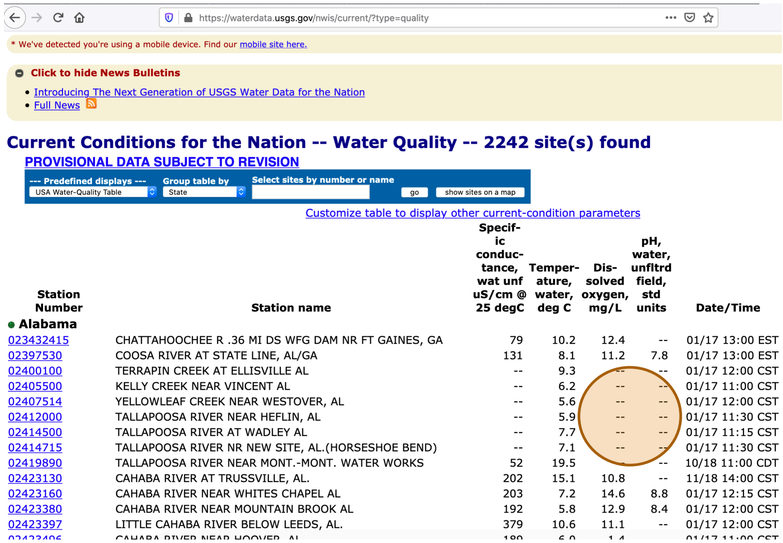
PID	ST_NUM	ST_NAME	OWN_OCCUPIED	NUM_BEDROOMS	NUM_BATH	SQ_FT
100001000	104	PUTNAM	Y	3	1	1000
100002000	197	LEXINGTON	N	3	1.5	--
100003000		LEXINGTON	N	n/a	1	850
100004000	201	BERKELEY	12	1	NaN	700
	203	BERKELEY	Y	3	2	1600
100006000	207	BERKELEY	Y	NA	1	800
100007000	NA	WASHINGTON		2	HURLEY	950
100008000	213	TREMONT	Y	1	1	
100009000	215	TREMONT	Y	na	2	1800

Why is Type of Variable Important

- Handling of missing values
- Distance between
  - Values
  - Data items
- Used for measuring accuracy, error
- Guiding the learning process
  - Selection of algorithms

Common Problem: Missing Value

- Occurrence
  - Missing completely at random
  - Missing at random (a group not wanting to participate)
  - Missing not at random (a group not able to participate)
- What does it mean?
  - The value was not provided
  - The value does not exist or has no practical interpretation
  - The value is being hidden (redaction)
  - Others: The value is not reliable, …



* We've detected you're using a mobile device. Find our mobile site here.

Click to hide News Bulletins

- Introducing The Next Generation of USGS Water Data for the Nation
- Full News

Current Conditions for the Nation -- Water Quality -- 2242 site(s) found

PROVISIONAL DATA SUBJECT TO REVISION

--- Predefined displays ---: USA Water-Quality Table
Group table by: State
Select sites by number or name

Customize table to display other current-condition parameters

Station Number	Station name	Specif-ic conduc-tance, wat unf uS/cm @ 25 degC	Temper-ature, water, deg C	Dis-solved oxygen, mg/L	pH, water, unfltrd field, std units	Date/Time
● **Alabama**						
023432415	CHATTAHOOCHEE R .36 MI DS WFG DAM NR FT GAINES, GA	79	10.2	12.4	--	01/17 13:00 EST
02397530	COOSA RIVER AT STATE LINE, AL/GA	131	8.1	11.2	7.8	01/17 13:00 EST
02400100	TERRAPIN CREEK AT ELLISVILLE AL	--	9.3	--	--	01/17 12:00 CST
02405500	KELLY CREEK NEAR VINCENT AL	--	6.2	--	--	01/17 11:00 CST
02407514	YELLOWLEAF CREEK NEAR WESTOVER, AL	--	5.6	--	--	01/17 12:00 CST
02412000	TALLAPOOSA RIVER NEAR HEFLIN, AL	--	5.9	--	--	01/17 11:30 CST
02414500	TALLAPOOSA RIVER AT WADLEY AL	--	7.7	--	--	01/17 11:15 CST
02414715	TALLAPOOSA RIVER NR NEW SITE, AL.(HORSESHOE BEND)	--	7.1	--	--	01/17 11:30 CST
02419890	TALLAPOOSA RIVER NEAR MONT.-MONT. WATER WORKS	52	19.5	--	--	10/18 11:00 CDT
02423130	CAHABA RIVER AT TRUSSVILLE, AL.	202	15.1	10.8	--	11/18 14:00 CST
02423160	CAHABA RIVER NEAR WHITES CHAPEL AL	203	7.2	14.6	8.8	01/17 12:15 CST
02423380	CAHABA RIVER NEAR MOUNTAIN BROOK AL	192	5.8	12.9	8.4	01/17 12:00 CST
02423397	LITTLE CAHABA RIVER BELOW LEEDS, AL.	379	10.6	11.1	--	01/17 12:00 CST
02423496	CAHABA RIVER NEAR HOOVER, AL	189	6.0	1.4		01/17 11:00 CST

- How to detect it?
  - By checking for specific values: NA, Not applicable, out-of-range value, 0, -1, “”.

Missing Value – Handling

- Ignoring missing value (Omission)
  - Reduces available data
- Impute new value (Imputation)
  - Mean or median
  - Default value
- Analysis techniques which are robust against missing value
  - Expectation maximization

Code Examples

- Basic concepts: **DataPreparation-Numeric.ipynb**

- An illustration: **Clean-RealSample.ipynb**

- Code: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l5-dataprep/Clean-RealSample.ipynb

Code Examples

- COVID-19 data exploration
  - New York Times collected data for US
  - Focus on South Carolina as well as Richland county
  - Aggregate as well as daily counts of cases and deaths

  - https://github.com/biplav-s/course-tai/blob/main/sample-code/l4-l5-supervised-ml/CovidExploration.ipynb

Concepts

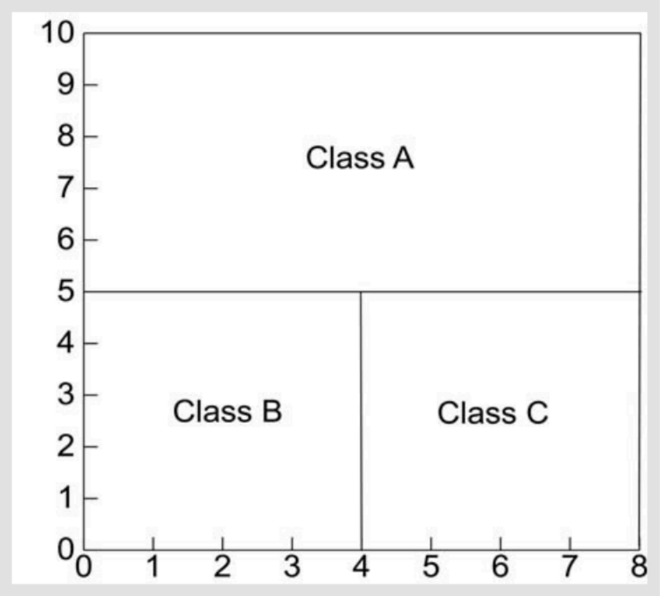
- **Input data**: data available
  - **Training data**: used for training a learning algorithm and get a model
    - [Optional] **Validation data**: used to tune parameters
  - **Test data**: used to test a learning model

- **Prediction problem**
  - Learning value of a continuous variable
- **Classification problem**
  - Separating data into classes (also called labels, categorical types)
  - One of the attributes is the class label we are trying to learn
  - Class label is the **supervision**
- **Clustering problem**
  - We are trying to learn grouping of data
  - There is no attribute indicating membership in the groups (hence, **unsupervised**)

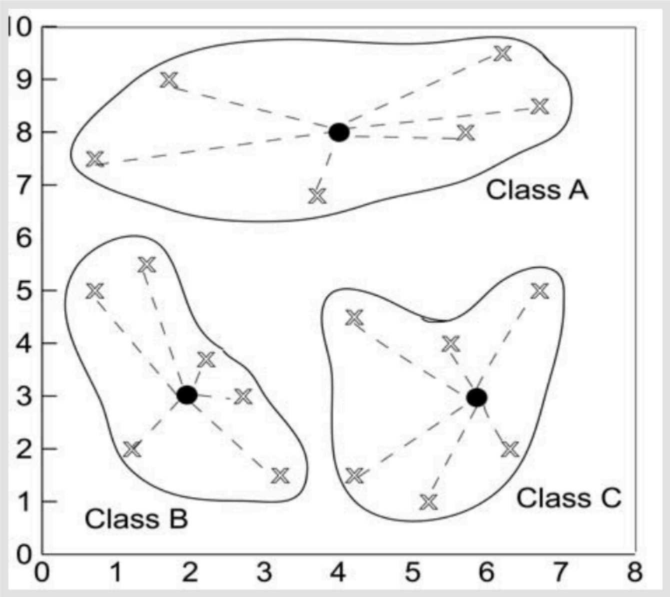
Reference: https://machinelearningmastery.com/difference-test-validation-datasets/
https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf

Methods for Classification

Partitioning Based



Distance Based



Source: https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf

Linear Methods

Assumption: target value (y) is expected to be a linear combination of the features (Xj).

Function estimate (linear)
W: weight, b: bias

$$f(X_j) = X_j W + b$$

Error Term (mean squared error)

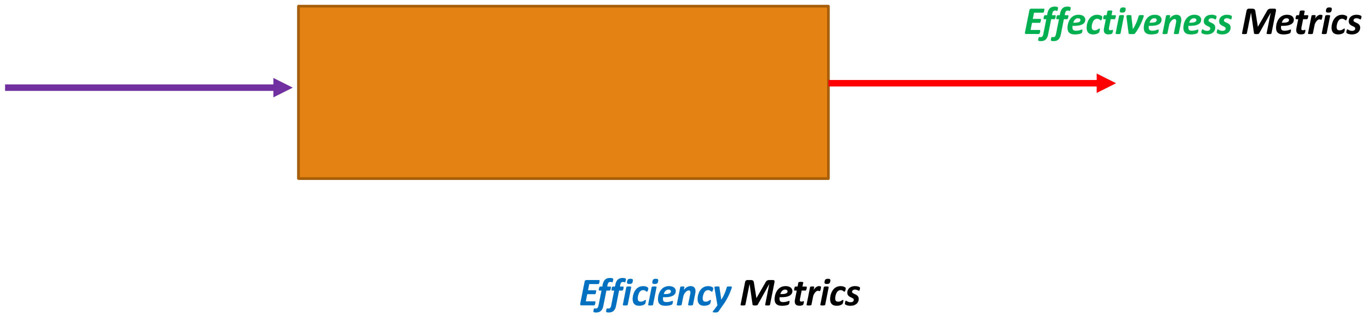
$$MSE = \frac{1}{n} \sum_{j=1}^{n} \left[f(X_{j\cdot}) - y_j \right]^2$$

Many variants depending on the nature of error being minimized: overfitting (Ridge), number of non-zero coefficients (Lasso), …

- Reference: https://scikit-learn.org/stable/modules/linear_model.html

Metric Types

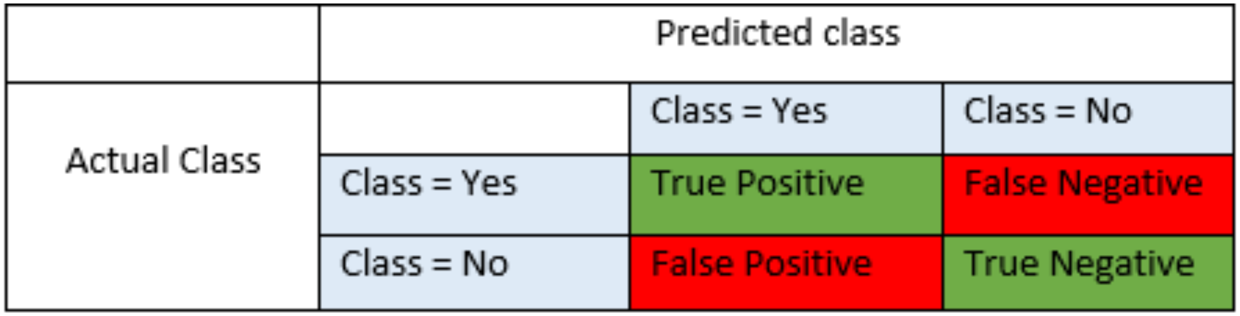
- **Effectiveness**: what the **user** of a system sees, primarily cares about
- **Efficiency**: what the **executor** in a system sees, primarily cares about



Example: Predicting COVID cases

- **Effectiveness**: what the **user** of a system sees, primarily cares about
  - *How accurate (high) is the prediction?*
  - *How low is the error?*
- **Efficiency**: what the **executor** in a system sees, primarily cares about
  - *How low is the error?*
  - *How fast was prediction made?*
  - *How stable is the prediction to change in data?*

Metrics: Accuracy, Precision, Recall



	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

Accuracy =
(TP+TN)/
(TP+FP+FN+TN)

Precision =
(TP)/
(TP+FP)

Recall =
(TP)/
(TP+FN)

F1 Score: *Harmonic Mean*

1/F1 = 1/Precision + 1/Recall

F1 = 2*(Recall * Precision) /
(Recall + Precision)

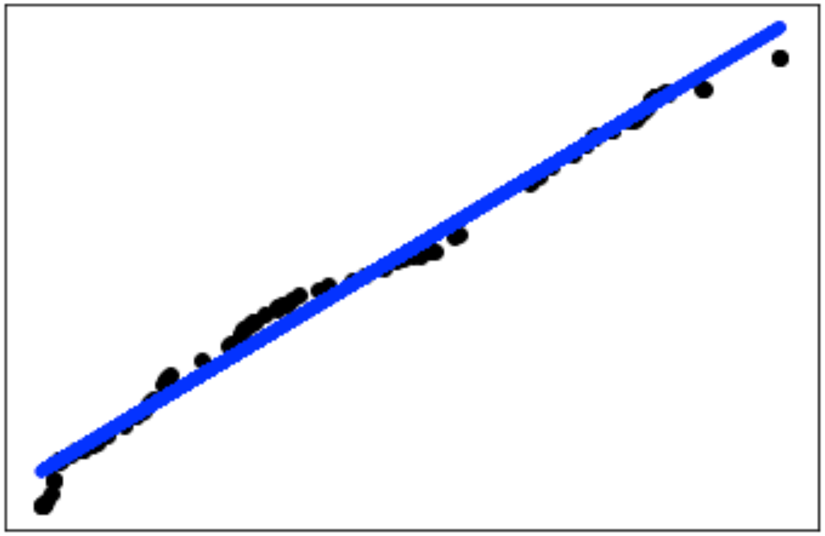
Comparing Classification Methods

- Predictive accuracy
- Interpretability: providing insight
- Robustness: handling noisy data

- Speed
- Scalability: large volume of data

Source: Data Mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber

Linear Regression



Notebook: https://github.com/biplav-s/course-tai/blob/main/sample-code/l4-l5-supervised-ml/Supervised-Regression-Classification.ipynb

Machine Learning – Insights from Data

- Descriptive analysis
  - Describe a past phenomenon
  - **Methods**: classification, clustering, dimensionality reduction, anomaly detection, neural methods
- Predictive analysis
  - Predict about a new situation
  - **Methods**: time-series, neural networks
- Prescriptive analysis
  - What an agent should do
  - **Methods**: simulation, reinforcement learning, reasoning

- New areas
  - Counterfactual analysis
  - Causal Inferencing
  - Scenario planning

Sample Learning Task

- COVID-19 data

- Notebook: https://github.com/biplav-s/course-tai/blob/main/sample-code/l4-l5-supervised-ml/Supervised-Regression-Classification.ipynb

Reference and Demo

- Data: UCI Datasets - https://archive.ics.uci.edu/ml/datasets.php
- Tools:
  - Weka - https://www.cs.waikato.ac.nz/ml/weka/

References

- Blogs: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/
- Google: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc
- Insead:
  - Description: https://inseaddataanalytics.github.io/INSEADAnalytics/CourseSessions/Sessions67/ClassificationAnalysisReading.html
  - Data analytics for Business: https://inseaddataanalytics.github.io/INSEADAnalytics/

Lecture 4: Concluding Comments

- Did an overview of Machine learning
- Looked at data processing and cleaning
- Looked at supervised learning problem
- Worked with COVID data

Concluding Segment

About Next Lecture – Lecture 5

Lecture 5: Supervised ML Continued

- More Classification Methods: Linear, Decision Tree, Random Forest
- Choosing between methods
- Tools: weka
- Problems beyond COVID-19