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

Lecture 5: Supervised Machine Learning

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

28TH JAN 2021

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

Organization of Lecture 6

- Introduction Segment
  - Recap of Lecture 6
- Main Segment
  - Terminology, objective and setup
  - Metrics to measure performance
  - Linear methods
- Quiz 1
- Concluding Segment
  - About Next Lecture – Lecture 7
  - Ask me anything

Introduction Segment

Recap of Lecture 5

- We looked at how to improve the quality of data
- Data preparation: handling missing values
- Importance of annotation and methods
  - Glossary
  - Taxonomy, Is-a relationship
  - Ontology

Main Segment

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

Nomenclature

Column, Attribute, Feature

Row, Item

	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
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

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

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

- **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**)
- **Prediction problem**
  - Learning value of a continuous variable

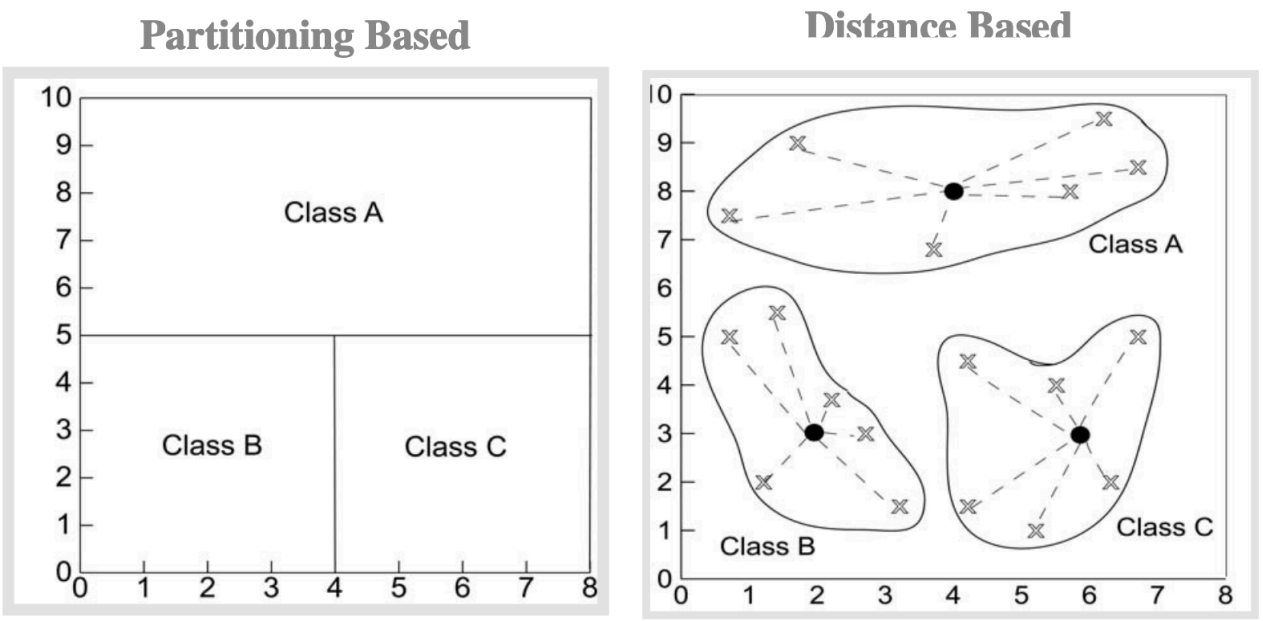
Reference: https://machinelearningmastery.com/difference-test-validation-datasets/
https://www2.seas.gwu.edu/~bell/csci243/lectures/classification.pdf

Sample Learning Task

- COVID-19 data

- Notebook: https://github.com/biplav-s/course-d2d-ai/blob/main/sample-code/l6-l7-supervised-ml/Supervised-Regression.ipynb

Methods for Classification



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

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?*

Example: Detecting Spam in Email

- **Effectiveness**: what the **user** of a system sees, primarily cares about
  - *How many spams identified?*
  - *How many spams missed?*
- **Efficiency**: what the **executor** in a system sees, primarily cares about
  - *How fast were spams detected?*
  - *How much memory was used per million emails processed ?*

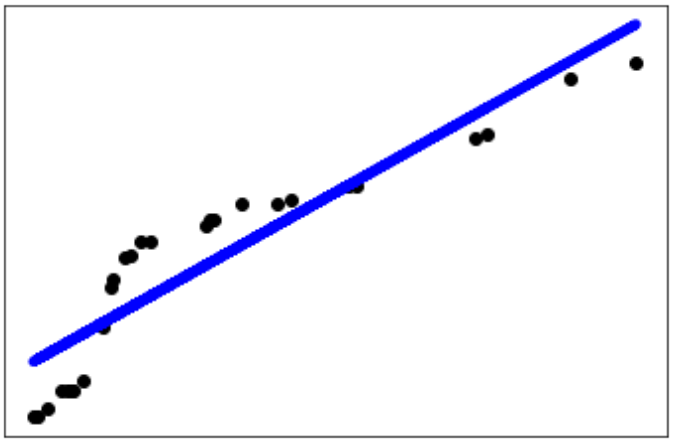
Comparing Classification Methods

- 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-d2d-ai/blob/main/sample-code/l6-l7-supervised-ml/Supervised-Regression.ipynb

Metrics: Accuracy, Precision, Recall

		Predicted class	
		Class = Yes	Class = No
Actual Class	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)

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 6: Concluding Comments

- We looked at
  - Supervised learning task
  - Concepts related to data characteristics and quality
  - Evaluation approach
- Also investigate regression method

Quiz 1: About Open Data

Concluding Segment

About Next Lecture – Lecture 7

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

Lecture 7: Analysis

- Review Quiz 1
- Structured Data: Supervised Methods
  - Decision trees/ random forest
  - The variety of methods
  - Choosing a method that works

- Reading material:
  - “Which ML to Use” with title: Data-driven advice for applying machine learning to bioinformatics problems
  - “10 tips with title”: Ten quick tips for machine learning in computational biology