Data Mining and Business Intelligence

ITA5007

PROF. E.P.EPHZIBAH

Topics

SUPERVISED AND UNSUPERVISED LEARNING

summary of terms used:

Algorithm: Refers to a specific procedure used to implement a particular data mining technique: classification tree, discriminant analysis, and the like.

Model: Refers to an algorithm as applied to a dataset, complete with its settings (many of the algorithms have parameters that the user can adjust).

Observation: Is the unit of analysis on which the measurements are taken (a customer, a transaction, etc.); also called case, record, pattern, or row. (Each row typically represents a record; each column, a variable.)

Important terms

Pattern: Is a set of measurements on an observation (e.g., the height, weight, and age of a person).

Prediction: The prediction of the value of a continuous output variable; also called estimation.

Predictor: Usually denoted by X, is also called a feature, input variable, independent variable, or from a database perspective, a field.

Response: Usually denoted by Y, is the variable being predicted in supervised learning; also called dependent variable, output variable, target variable, or outcome variable.

Supervised Learning Refers to the process of providing an algorithm (logistic regression, regression tree, etc.) with records in which an output variable of interest is known and the algorithm "learns" how to predict this value with new records where the output is unknown.

Test Data (or test set) Refers to that portion of the data used only at the end of the model building and selection process to assess how well the final model might perform on additional data.

Training Data (or training set) Refers to that portion of data used to fit a model.

Unsupervised Learning Refers to analysis in which one attempts to learn something about the data other than predicting an output value of interest (e.g., whether it falls into clusters).

Important terms

Validation Data (or validation set) Refers to that portion of the data used to assess how well the model fits, to adjust some models, and to select the best model from among those that have been tried.

Variable Is any measurement on the records, including both the input (X) variables and the output (Y) variable.

Supervised Learning

Supervised learning algorithms are those used in classification and prediction.

We must have data available in which the value of the outcome of interest (e.g., purchase or no purchase) is known.

These training data are the data from which the classification or prediction algorithm "learns," or is "trained," about the relationship between predictor variables and the outcome variable.

Supervised Learning

Once the algorithm has learned from the training data, it is then applied to another sample of data (the validation data) where the outcome is known, to see how well it does in comparison to other models.

The model can then be used to classify or predict the outcome of interest in new cases where the outcome is unknown.

Supervised Learning

Simple linear regression analysis is an example of supervised learning

The Y variable is the (known) outcome variable and the X variable is a predictor variable.

A regression line is drawn to minimize the sum of squared deviations between the actual Y values and the values predicted by this line.

The regression line can now be used to predict Y values for new values of X for which we do not know the Y value.

Supervised Learning

The target attribute (class label) expresses for each record either the membership class or a measurable quantity.

Classification and regression models belong to this category.

In a supervised (or direct) learning analysis, a target attribute either represents the class to which each record belongs or expresses a measurable quantity, such as the total value of calls that will be placed by a customer in a future period.

As a second example of the supervised perspective, consider an investment management company wishing to predict the balance sheet of its customers based on their demographic characteristics and past investment transactions.

Supervised learning processes are therefore oriented toward prediction and interpretation with respect to a target attribute.

Unsupervised Learning

Unsupervised learning algorithms are those used where there is no outcome variable to predict or classify.

Hence, there is no “learning” from cases where such an outcome variable is known.

Association rules, dimension reduction methods, and clustering techniques are all unsupervised learning methods.

There is no target attribute that exists and consequently, the purpose of the analysis is to identify regularities, similarities, and differences in the data. It is also possible to derive association rules.

Alternatively one can determine groups of records, called clusters, characterized by similarity within each cluster and by dissimilarity among the elements of distinct clusters.

Sample dataset for classification

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Sample dataset for classification

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Sample dataset for prediction

X_1	X_2	X_3	Y
11	68.028164	0	21.46126
12	69.086446	0	23.28792
13	84.806730	1	18.71906
14	19.011313	1	18.50209
15	63.046323	1	22.37717
16	82.686964	1	24.19955
17	59.263664	1	20.64198
18	88.756598	0	29.50144
19	77.884304	1	24.49684
20	9.346073	0	27.15191

Sample dataset for prediction

ID	Gender	Score	Years in company	Division	Salary Increase
1	F	11	9	Production	22
2	F	89	1	Production	97
3	F	21	4	Production	47
4	F	81	1	Production	127
5	F	31	4	Research	65
6	F	71	1	Research	53
7	F	11	4	Sales	74
8	F	16	7	Production	18
9	M	20	6	Research	129
10	M	79	3	Sales	475
11	M	51	3	Research	342
12	M	69	2	Sales	329
13	M	30	7	Sales	185
14	M	71	7	Sales	332
15	M	39	1	Sales	268
16	M	89	6	Production	518
17	M	50	8	Production	390

Unsupervised learning.

Unsupervised (or indirect) learning analyses are not guided by a target attribute.

Therefore, data mining tasks in this case are aimed at discovering recurring patterns and affinities in the dataset.

As an example, consider an investment management company wishing to identify clusters of customers who exhibit homogeneous investment behavior, based on data on past transactions.

In most unsupervised learning analyses, one is interested in identifying clusters of records that are similar within each cluster and different from members of other clusters.

Data mining tasks under supervised and unsupervised learning

Seven basic data mining tasks can be identified:

- characterization and discrimination;
- classification;
- regression;
- time series analysis;

supervised learning

- association rules;
- clustering;
- description and visualization;

unsupervised learning

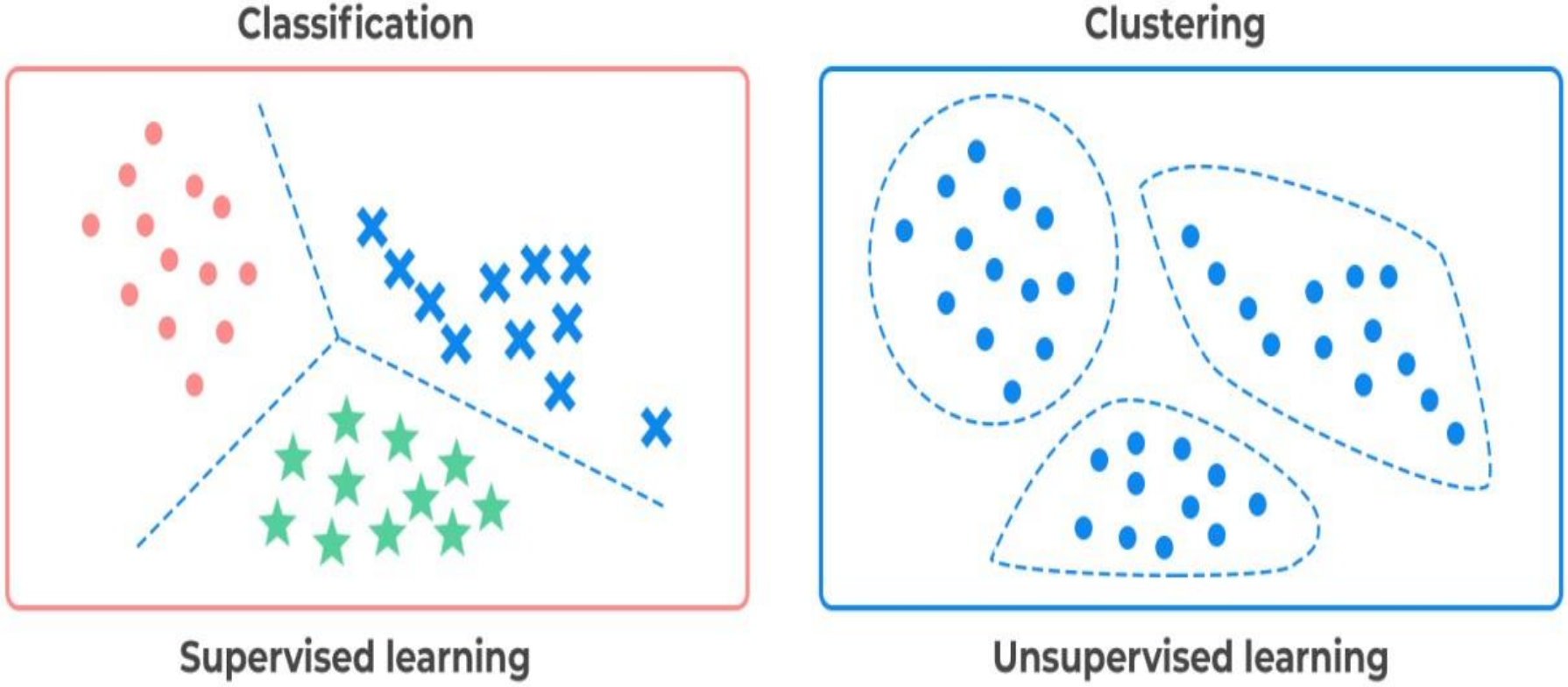
Sample dataset for Unsupervised learning

Transaction No	Products
1	beer, wine, cheese
2	beer, potato chips
3	eggs, flour, butter, cheese
4	eggs, flour, butter, beer, potato chips
5	wine, cheese
6	potato chips
7	eggs, flour, butter, wine, cheese
8	eggs, flour, butter, beer, potato chips
9	wine, beer
10	beer, potato chips
11	butter, eggs
12	beer, potato chips
13	flour, eggs
14	beer, potato chips
15	eggs, flour, butter, wine, cheese
16	beer, wine, potato chips, cheese
17	wine, cheese
18	beer, potato chips
19	wine, cheese
20	beer, potato chips

Sample dataset for Unsupervised learning

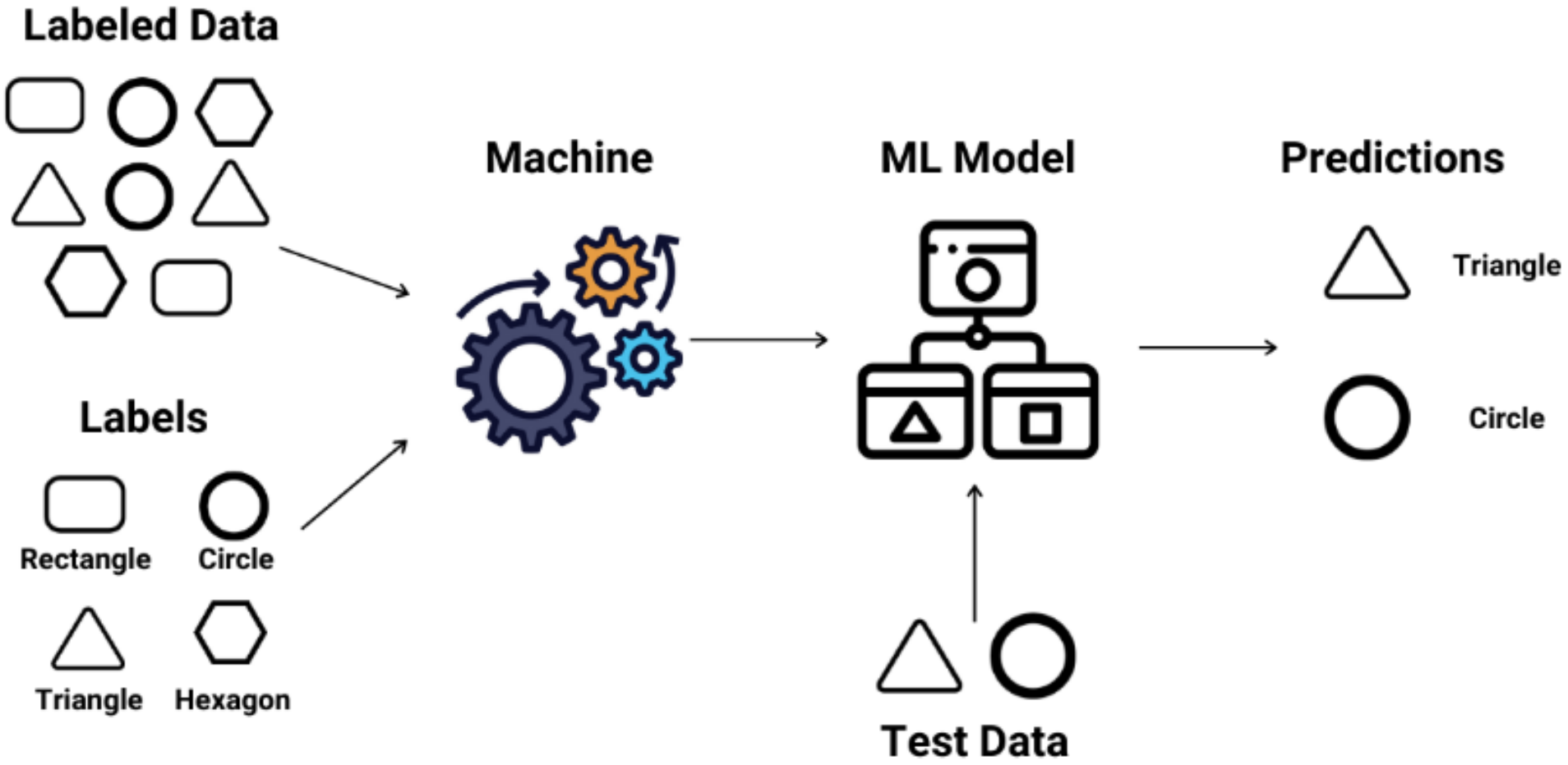
TID	items
T1	I1, I2, I5
T2	I2, I4
T3	I2, I3
T4	I1, I2, I4
T5	I1, I3
T6	I2, I3
T7	I1, I3
T8	I1, I2, I3, I5
T9	I1, I2, I3

Supervised vs. Unsupervised Learning

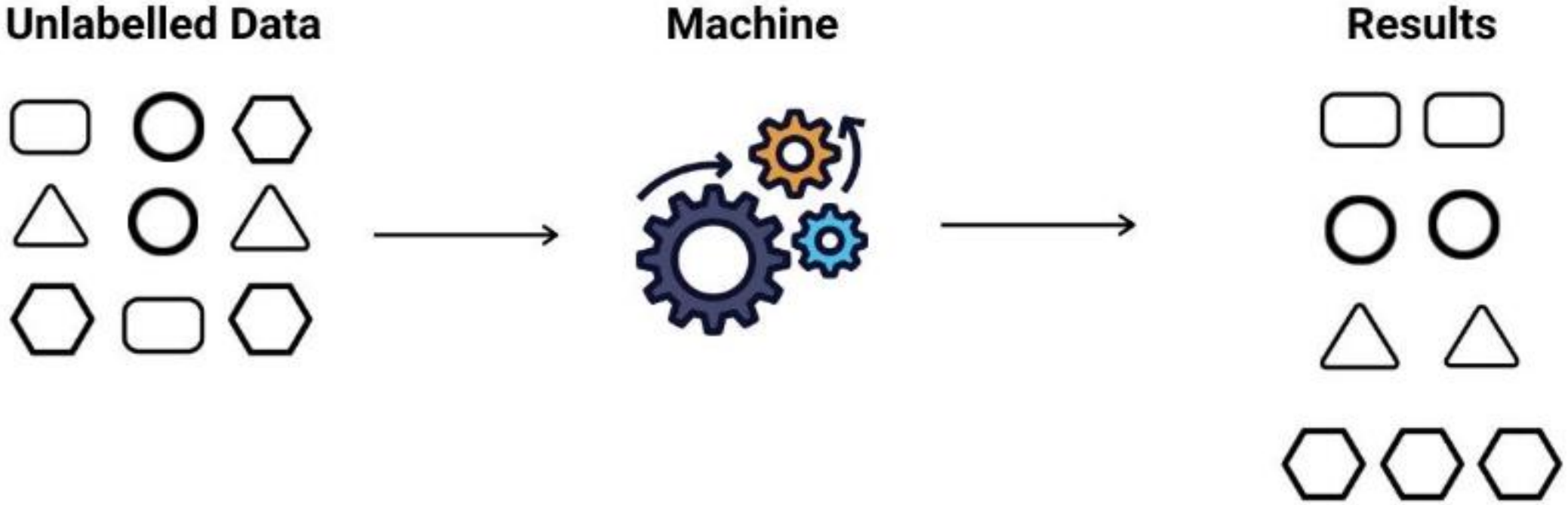


https://www.linkedin.com/pulse/supervised-vs-unsupervised-learning-whats-difference-smriti-saini

Supervised Learning



Unsupervised Learning



https://www.enjoyalgorithms.com/blogs/supervised-unsupervised-and-semisupervised-learning

Questions

2.1 Assuming that data mining techniques are to be used in the following cases, identify whether the task required is supervised or unsupervised learning.

a. Deciding whether to issue a loan to an applicant based on demographic and financial data (with reference to a database of similar data on prior customers).

b. In an online bookstore, making recommendations to customers concerning additional items to buy based on the buying patterns in prior transactions.

c. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison to other packets whose threat status is known.

d. Identifying segments of similar customers

e. Predicting whether a company will go bankrupt based on comparing its financial data to those of similar bankrupt and nonbankrupt firms.

f. Estimating the repair time required for an aircraft based on a trouble ticket.

g. Automated sorting of mail by zip code scanning.

h. Printing of custom discount coupons at the conclusion of a grocery store checkout based on what you just bought and what others have bought previously.