

BUSINESS ANALYTICS

Data Mining Techniques

MACHINE LEARNING

- Machine learning teaches computers to do what comes naturally to humans: learn from experience.
- Machine learning algorithms use computational methods to “learn” information directly from data without relying on a predetermined equation as a model.
- The algorithms adaptively improve their performance as the number of samples available for learning increases.

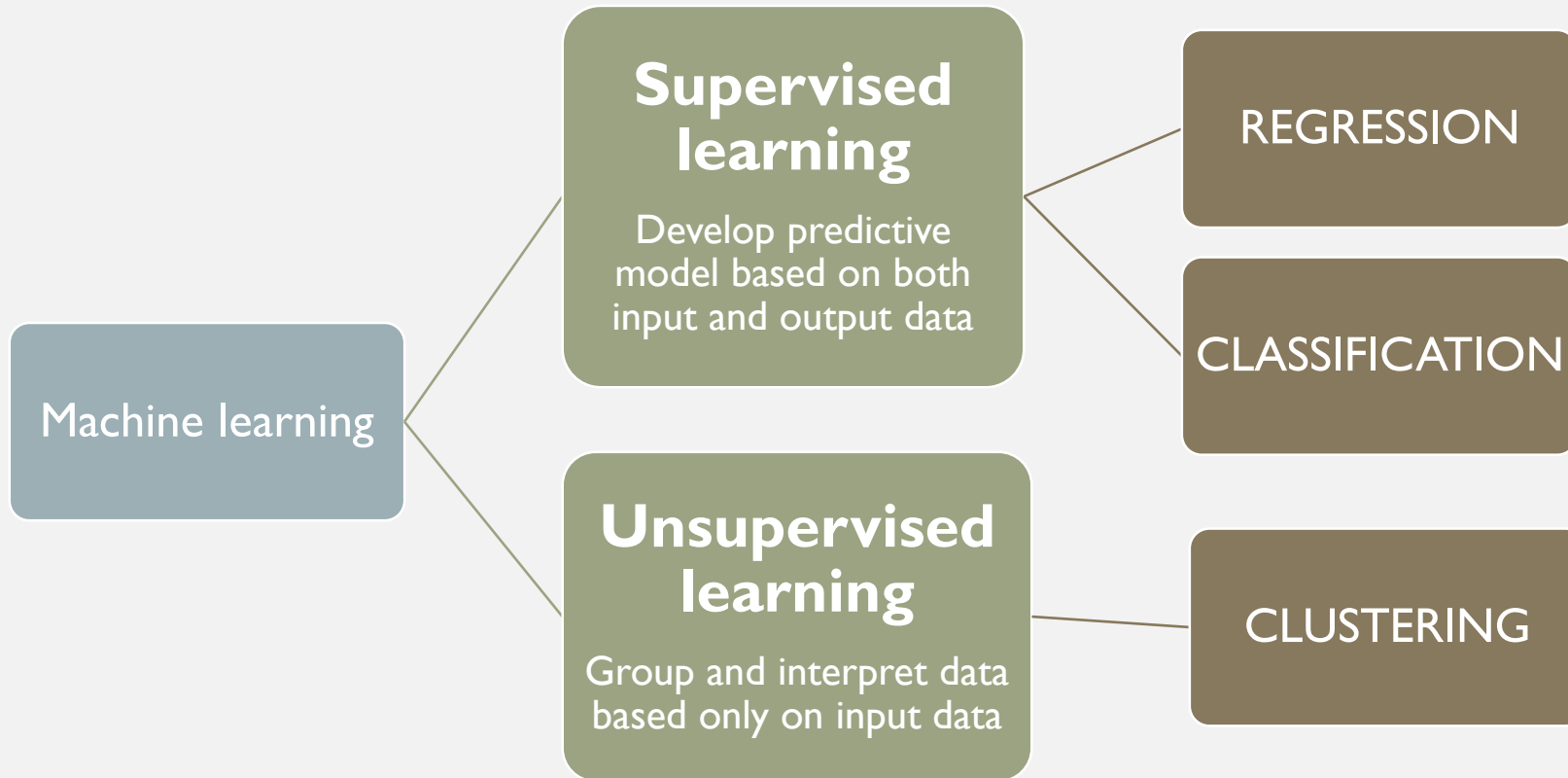
REAL WORLD APPLICATIONS

- Computational finance, for credit scoring and algorithmic trading.
- Image processing and computer vision, for face recognition, motion detection, and object detection.
- Computational biology, for tumor detection, drug discovery, and DNA sequencing.
- Energy production, for price and load forecasting.
- Automotive, aerospace, and manufacturing, for predictive maintenance.
- Natural language processing

MACHINE LEARNING

- An algorithm that 'learns' studies the features of a data set and comes up with its own prediction rule in order to determine classification.
- Two kinds of machine learning algorithms: Supervised and Unsupervised
- Supervised: provides the data set with correct labels, machine classifies them
- Unsupervised: provides the data set and machine clusters them based on studying the features.

MACHINE LEARNING TECHNIQUES



SUPERVISED LEARNING

- The aim of supervised learning is to build a model that makes predictions based on evidence in the presence of uncertainty.
- A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data.
- Classification techniques predict a discrete response: email is genuine or spam, tumor is cancerous or benign.
- Regression techniques predict continuous responses: changes in temp, fluctuations in demand.

UNSUPERVISED LEARNING

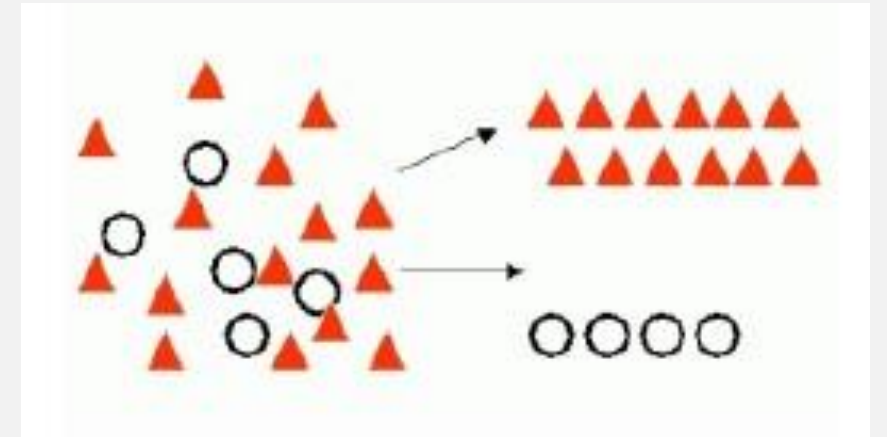
- Unsupervised learning finds hidden patterns or intrinsic structures in data.
- It is used to draw inferences from datasets consisting of input data without labeled responses.
- Clustering is the most common unsupervised learning technique.
- It is used in exploratory data analysis to find hidden patterns or groupings in data.
- Applications include gene sequence analysis, market research, and object recognition.

CLASSIFICATION

- Statistical classification is studied in machine learning.
- It is a type of supervised learning.
- Categories are predefined.
- It is used to categorise new observations into a predefined category.
- It is a means of predicting which group an observation belongs to.

BINARY CLASSIFICATION

- For binary classification there are only two classes.
 - Yes/No
 - Success/Failure
- Applications:
 - Credit card fraud transaction detection
 - Medical diagnosis
 - Spam detection



HOW DOES CLASSIFICATION WORK?

- bank loan application:
 - A bank loan officer wants to analyse the data in order to know which customer (loan applicant) is risky and which is safe.
- Data classification process include two steps:
 - Building the classifier or model
 - Using the classifier for classification

BUILDING THE CLASSIFIER OR MODEL

- This step is the learning step or the learning phase.
- The classification algorithms build the classifier.
- The classifier is built from the training set of data, made up of database records and their associated class labels.
- Once a predictive model has been trained it is needed to evaluate its predictive power on new data that have not been seen before.



MODEL TESTING

- This process determines if the predictive model is good enough to be moved into the production phase.
- The purpose of testing analysis is to compare the responses of the trained predictive model against the correct predictions for every instance in the training set.
- As these cases have not been used before to train the predictive model, the results of this process can be used as a simulation of what would happen in a real world situation.

CONFUSION MATRIX

- The confusion matrix is an $m \times m$, where m is the number of classes to be predicted. For binary classification problems, the number of classes is 2, thus the confusion matrix will have 2 rows and 2 columns.
- The rows represent the target classes
- The columns represent the output classes
- The diagonal cells show the number of cases correctly classified.
- The off-diagonal cells show the number of misclassified cases.

CONFUSION MATRIX

| | | | |
|-----------|---|--|---------------------|
| | | CLASSIFICATION | |
| | |  Positive | Negative |
| CONDITION | + | TRUE POSITIVE 2 | FALSE NEGATIVE 1 |
| | - | FALSE POSITIVE 4 | TRUE NEGATIVE 13 |
| | |  | |

| Instance | Target | Output | Instance | Target | Output |
|----------|--------|--------|----------|--------|--------|
| 1 | 1 | 0.99 | 11 | 1 | 0.41 |
| 2 | 1 | 0.85 | 12 | 0 | 0.40 |
| 3 | 1 | 0.70 | 13 | 0 | 0.28 |
| 4 | 1 | 0.60 | 14 | 0 | 0.27 |
| 5 | 0 | 0.55 | 15 | 0 | 0.26 |
| 6 | 1 | 0.54 | 16 | 0 | 0.25 |
| 7 | 0 | 0.53 | 17 | 0 | 0.24 |
| 8 | 1 | 0.52 | 18 | 0 | 0.23 |
| 9 | 0 | 0.51 | 19 | 0 | 0.20 |
| 10 | 1 | 0.49 | 20 | 0 | 0.10 |

Output-Target table.

- The matrix is helpful for choosing a decision threshold in order to label the instances as positive or negative.
- If I decided that my decision threshold is an output greater than .54 then we construct a confusion matrix to reflect that threshold.

| Instance | Target | Output | Instance | Target | Output |
|----------|--------|--------|----------|--------|--------|
| 1 | 1 | 0.99 | 11 | 1 | 0.41 |
| 2 | 1 | 0.85 | 12 | 0 | 0.40 |
| 3 | 1 | 0.70 | 13 | 0 | 0.28 |
| 4 | 1 | 0.60 | 14 | 0 | 0.27 |
| 5 | 0 | 0.55 | 15 | 0 | 0.26 |
| 6 | 1 | 0.54 | 16 | 0 | 0.25 |
| 7 | 0 | 0.53 | 17 | 0 | 0.24 |
| 8 | 1 | 0.52 | 18 | 0 | 0.23 |
| 9 | 0 | 0.51 | 19 | 0 | 0.20 |
| 10 | 1 | 0.49 | 20 | 0 | 0.10 |

Output-Target table.

| | Predicted Positive | Predicted Negative |
|-----------------|--------------------|--------------------|
| Actual Positive | 5 | 3 |
| Actual Negative | 1 | 11 |

CONFUSION MATRIX

- True positives (TP) which are instances that are positives and are classified as positives.
- False positives (FP) which are instances that are negatives and are classified as positives.
- False negatives (FN) which are instances that are positives and are classified as negatives.
- True negatives (TN) which are instances that are negative and are classified as negatives.

