LECTURER: Nghia Duong-Trung

DATA UTILIZATION

TOPIC OUTLINE

Introduction to Data Utilization	1
Pattern Recognition	2
Natural Language Processing (NLP)	3
Image Recognition	4
Detection and Sensing	5.1

TOPIC OUTLINE

Problem-Solving	5.2
Decision Support	6.1
Data Security and Data Protection	6.2

UNIT 2

PATTERN RECOGNITION

STUDY GOALS

On completion of this unit, you will have learned …

- ... definitions of pattern recognition and its real-world applications.
- ... definitions of user attitude, behavior, dissonance, and consonance.
- ... the concept of big data and its main characteristics.
- ... how to conduct research using online data such as Google Trends analytics.
- ... the predictive analytics process.
- ... well-known predictive algorithms such as regression, decision tree, and neural network.

EXPLAIN SIMPLY

1. What is pattern recognition and why is it needed in data analytics?
2. What are the different statistical and machine learning techniques available for predictive analytics? Are all of them reliable and accurate?
3. How can we design a neural network models' structure and evaluate the process?

- Automated recognition of patterns and regulations in data using **AI** and **machine learning** techniques.
- Used in **predictive and prescriptive analytics**.
- Predictive analytics with the aim of finding future trends, improves different sectors.
  - **Predicts future sales and customers' demands**
  - **Personalizes drug combinations to cure leukemia**
  - **Personalizes products and services for customers**

PREDICTING USER BEHAVIOR

- **Attitude:** feelings, tendencies, and beliefs determining how a person behaves
- **Behavior:** including actions taken in accordance to a person's attitude
- **Cognitive dissonance:** tension caused by inconsistency between one’s attitudes and behavior
- **Cognitive consonance:** relief offered by consistency between one’s attitudes and behavior

A Theory of Predictive Dissonance: Predictive Processing Presents a New Take on Cognitive Dissonance: https://www.frontiersin.org/articles/10.3389/fpsyg.2018.02218/full

CLICKSTREAM ANALYTICS

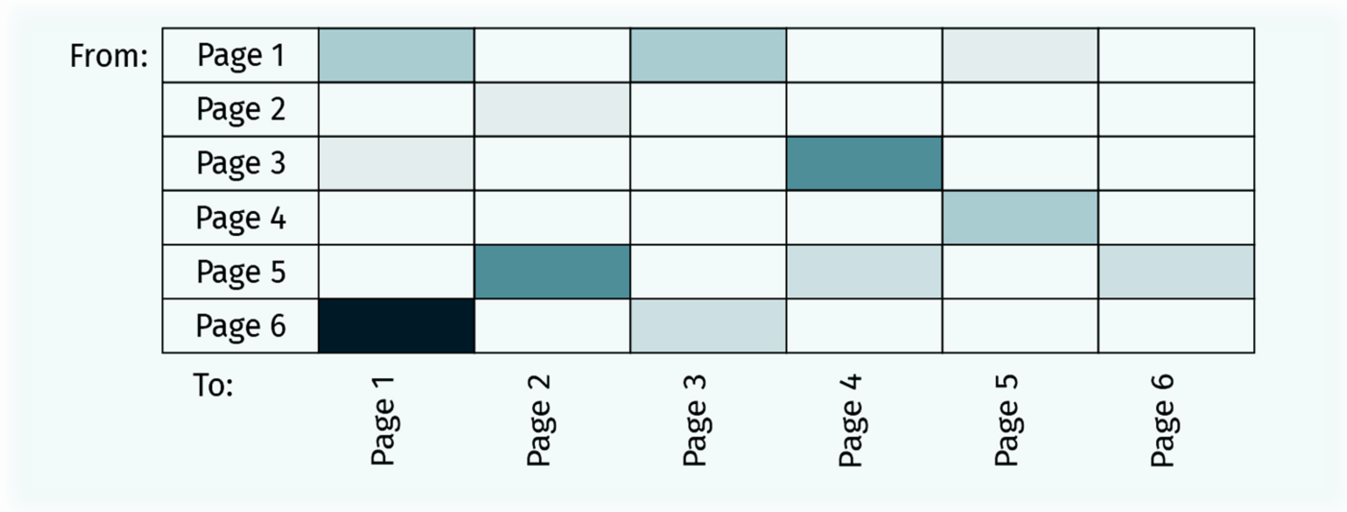
- Process of collecting, analyzing, and reporting aggregated data about user’s **website activity** and the sequence of web pages they have visited
- **Aim**: providing 360-degree view of the customer by identifying their interests
- **How**: clickstream analytics log and analyze all events occurring within the *time interval between user login and logout (e.g.,* user session)

Google Analytics:

https://analytics.google.com/analytics/web/provision/#/provision

CLICKSTREAM ANALYTICS TECHNIQUES

- Markov chain:
  - A **stochastic** model used to describe the sequence of events where the **probability** *is a function of the current state.*
  - *predicts user's next click.*
- Clustering:
  - **Finds users with similar patterns of visiting webpages**
- Big data analytics:
  - **Apache Spark, SAP HANA**

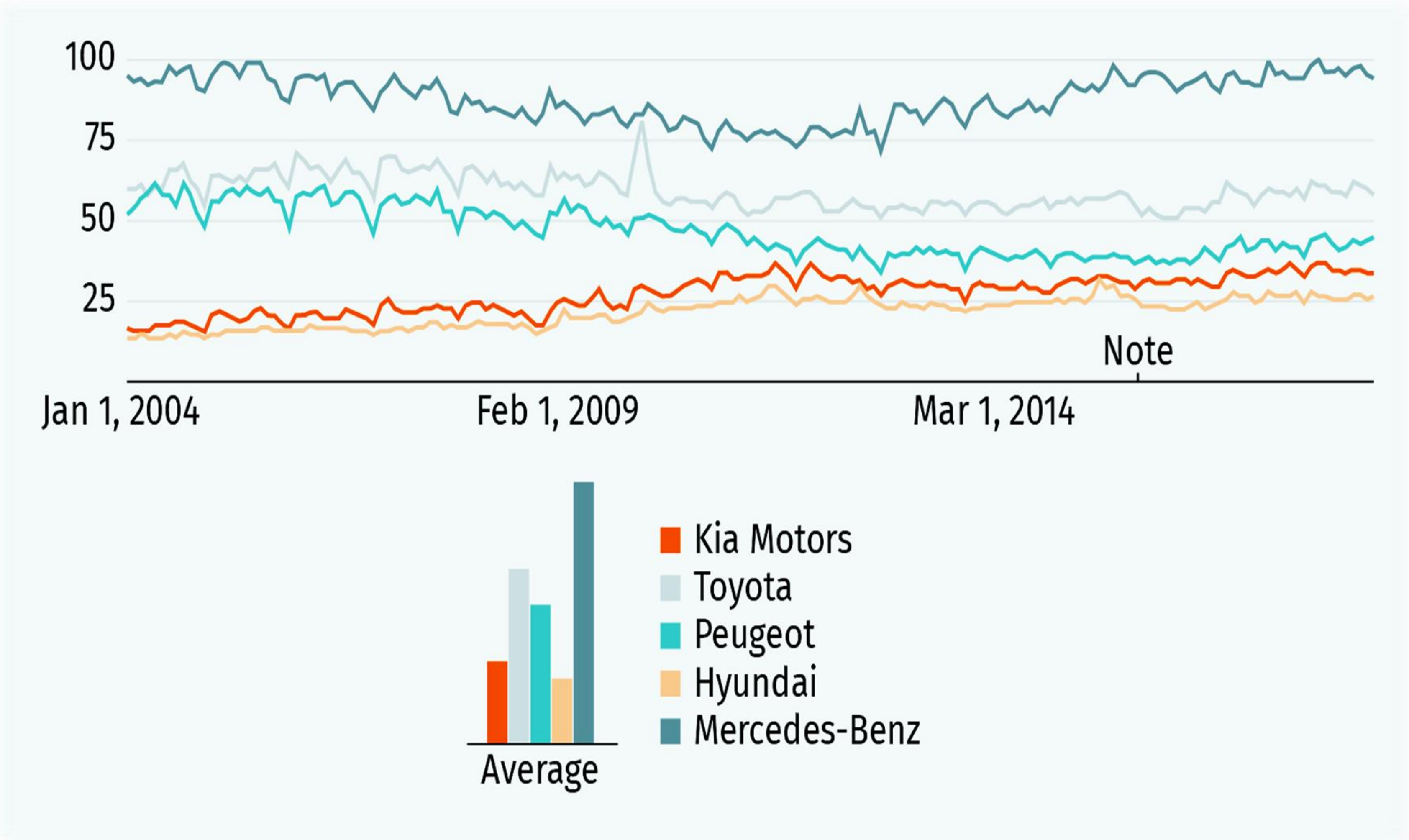


Source of the graphic: Course Book DLMBBD01, p. 37

CLICKSTREAM ANALYTICS APPLICATION AND TOOLS

- Clickstream analytics is a practical tool for understanding **customer's online activities**.
- **Google Trends** has made it possible to analyze the relative popularity of search enquiries in Google searches over time.
- Google Trends data for searches related to automobile sales were used to investigate the correlation between forward planning and GDP.

CLICKSTREAM ANALYTICS APPLICATION AND TOOLS

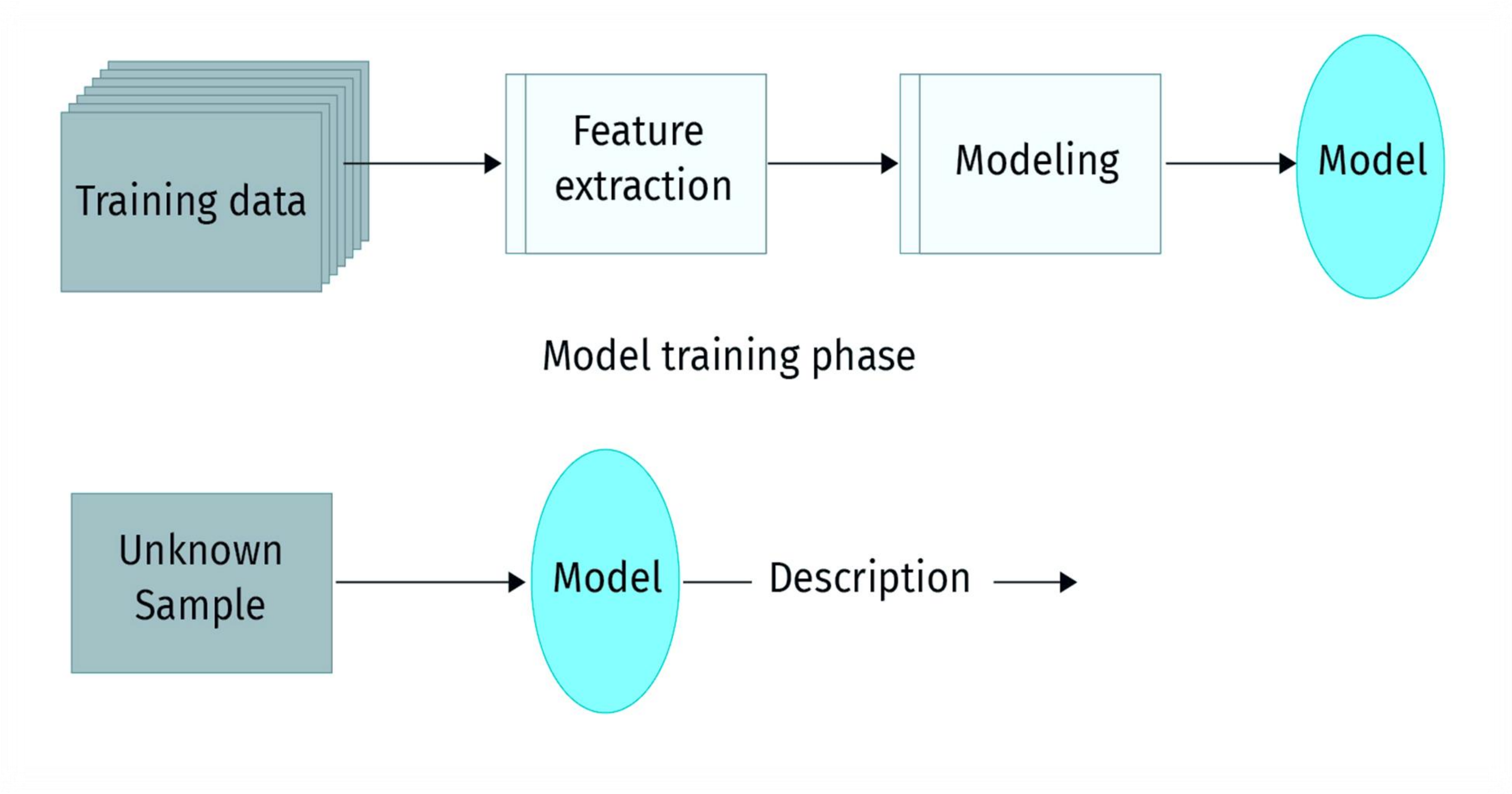


Source of the graphic: Course Book DLMBBD01, p. 38

PREDICTION TECHNIQUES

- **Prediction:** an estimation or forecast of future outcomes based on knowledge of the past.
- Requires identification of factors that have occurred in the past and appear to have influenced the outcome we are trying to predict.
- While developing a predictive model, one should consider the **impact** of each characteristic in the data and the **correlation** between them, all while **minimizing** the rate of misclassification.

BUILDING AND USING A DESCRIPTIVE MODEL



Source of the graphic: Course Book DLMBBD01, p. 38

– **Predictive analysis aim:** predict behavior of an entity

- Collecting Data: Model (imitating the user's behaviour) is built using data from previous experiments. Potential data sources should be identified to gather data.
- Pre-processing data: Gathered data have different formats, scales or units and include invalid or outlier items. This phase includes cleansing data, transforming format, eliminating outliers and erroneous data and unifying scales.
- Modeling: Classify similar behavioral reactions based on statistical similarities while differentiating dissimilar behavior.
- Deploying the model: After the model is built, it should be brought to production. The actual accuracy of the model can be verified only by deploying it in a real environment.
- Monitoring and evolving model: There is always a degree of uncertainty in real-world predictive models for two reasons : Existence of **lurking variables -** those that are not considered in the exploratory or observable variable but may affect the results. Secondly the changing of characteristics of the entity under investigation from the initial model.

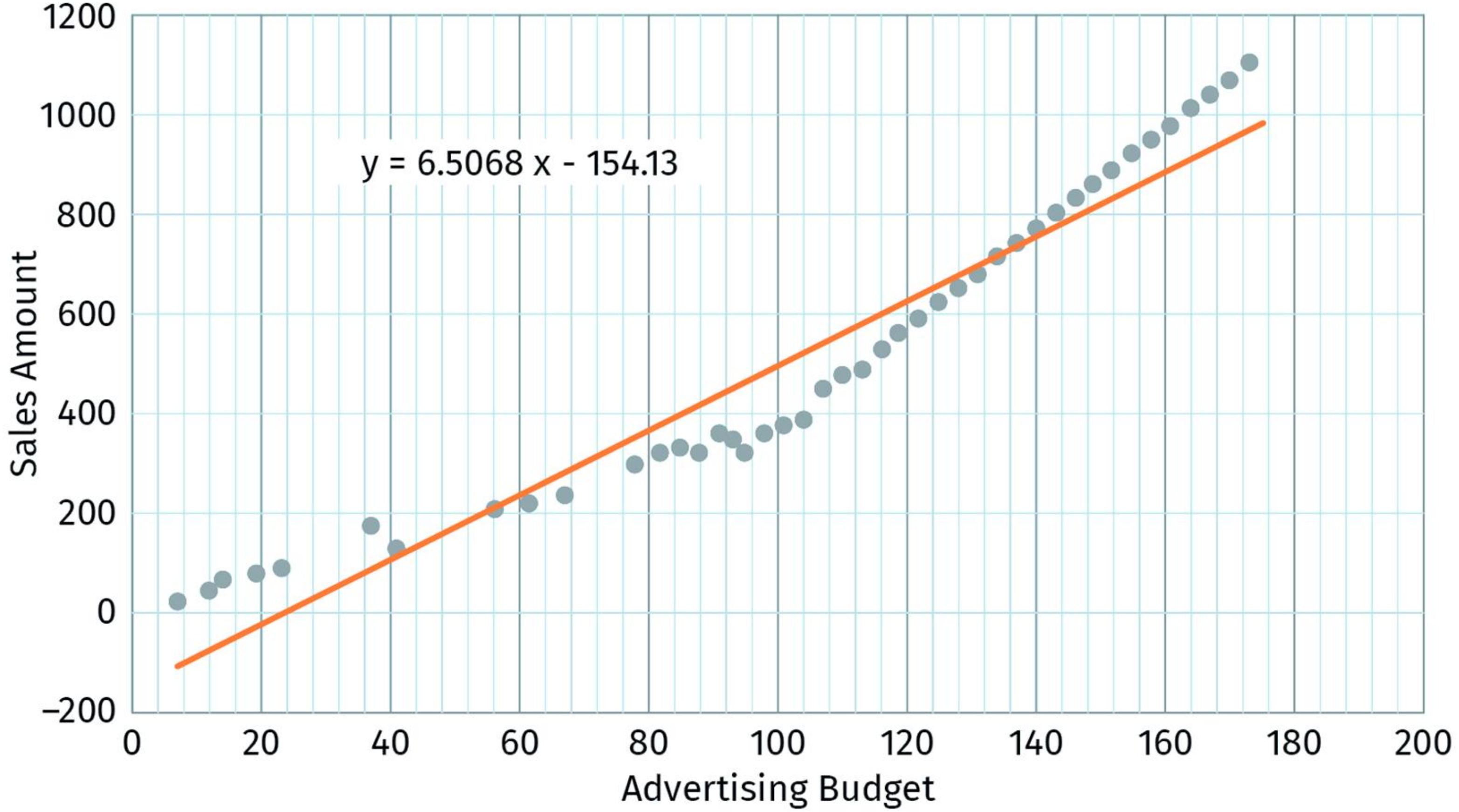
NAÏVE BAYESIAN MODELING

- A statistical model relying on the probability distribution of input values and conditional probabilities.
- The knowledge or information available prior to the experiment incorporating our backgrounds, or domain and knowledge, are considered as **a-priori information;** this information is the core part of the **Naïve Bayes** analysis.
- The data used for developing a model is a subset of a total set of observable data.

REGRESSION

- Determines the existing relationship between a dependant variable (target/responsive variable) and one or more independent variables/predictors.
- Predictor variables could be demographic, geographical, or domain-specific variables.
- Regression reveals how the value of a response variable changes when one of its predictors is varied while the other predictors remain fixed.

REGRESSION

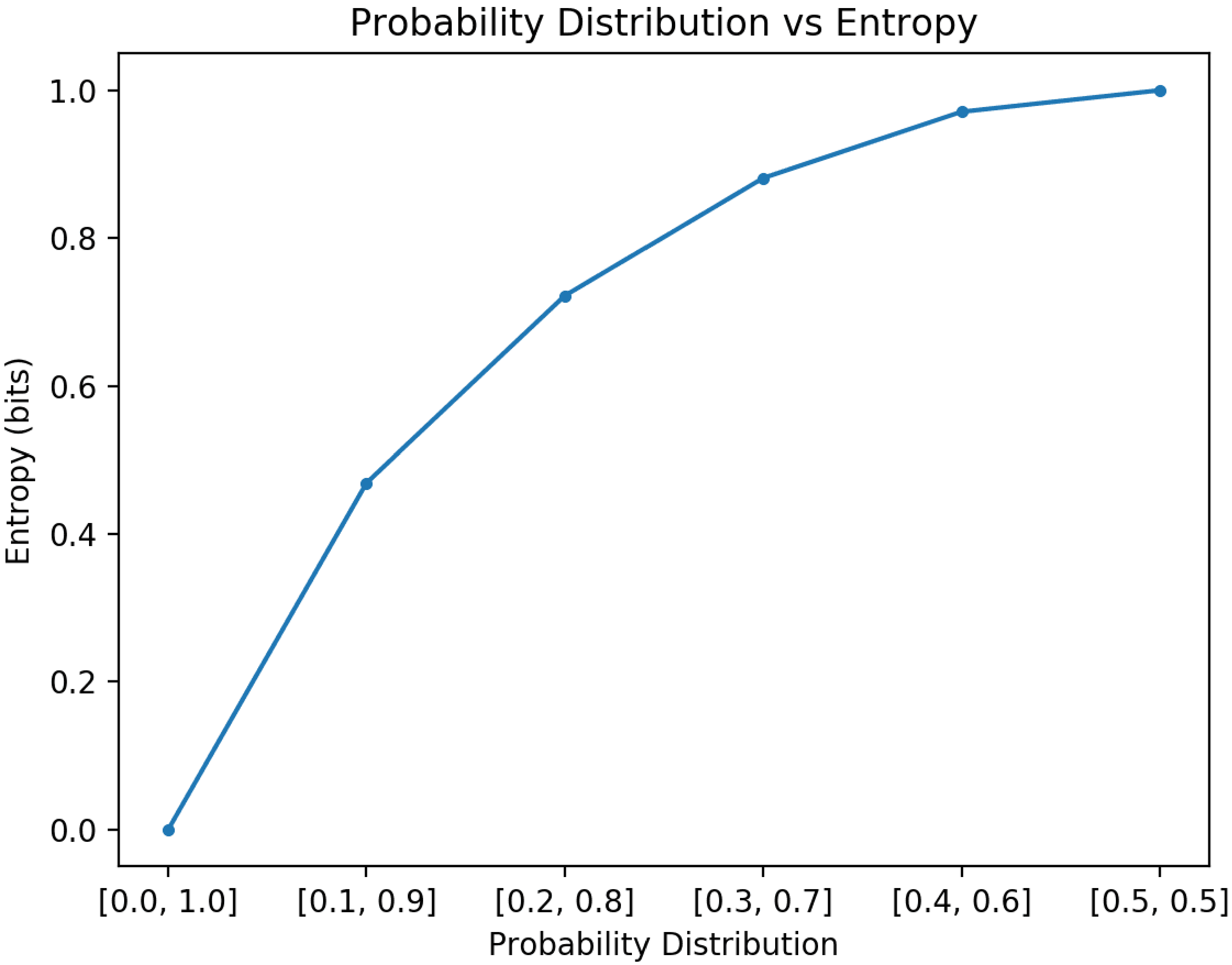


Source of the graphic: Erkeç, 2018.

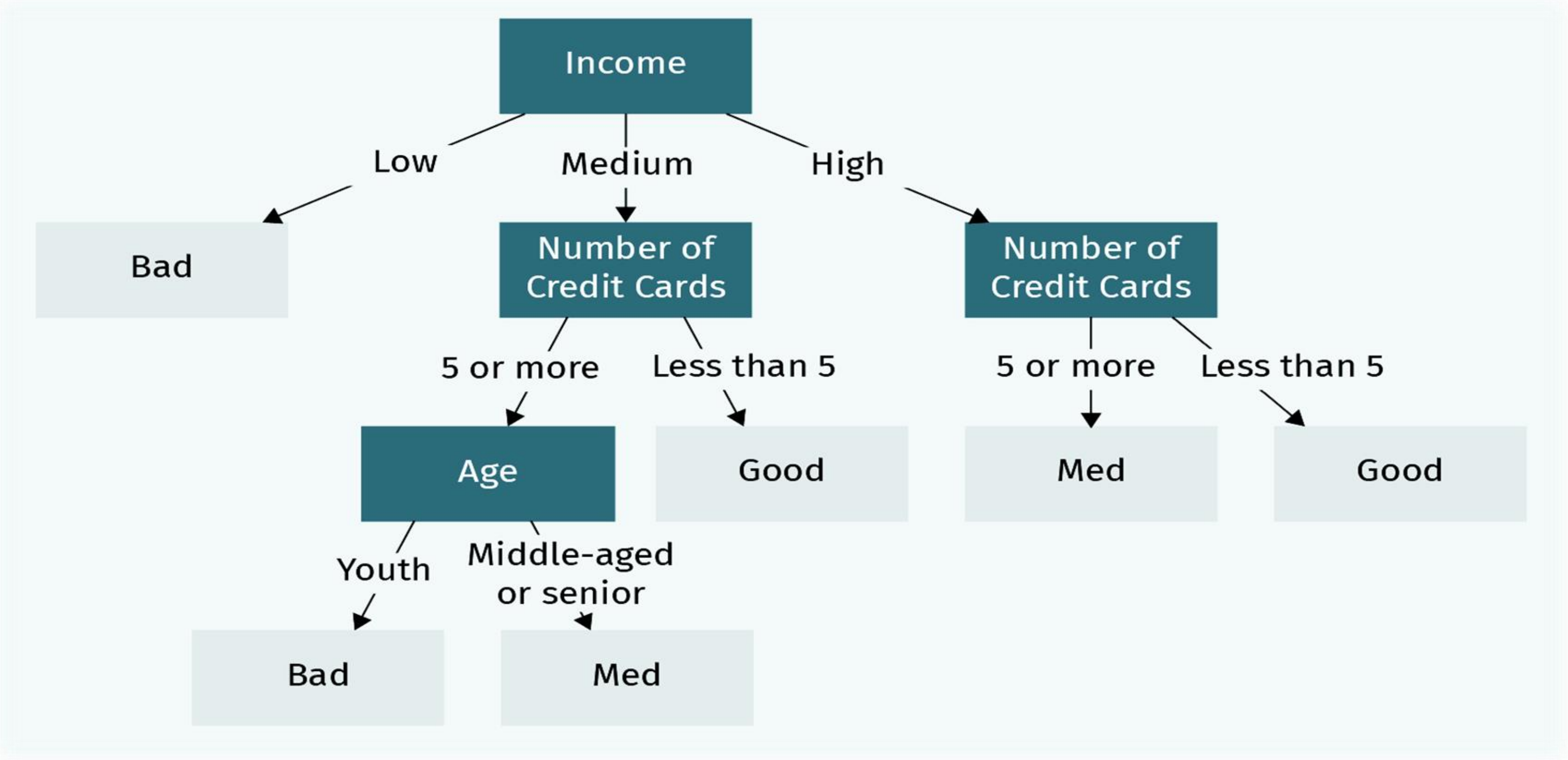
DECISION TREE

- Using a branching model to show each possible outcome of a decision.
- A classification method made up of decision nodes and branches.
- Beginning with the root node, an attribute or combination of attributes are evaluated at each following decision node.
- Sequence of one or more decision nodes leads to a terminating leaf node.

information entropy



Advantage: A tree provides an easy way to interpret the results by creating a rule-based prediction system.



Source of the graphic: Course Book DLMBBD01, p. 46

C4.5 Algorithm

- extension of ID3 algorithm
- Tests every variable at each level of the tree and selects the best splitter.
- using information gain or entropy

Classification and regression trees (CART)

- Construct binary decision tree (there are exactly two branches/ outcomes) at each decision node
- The best attribute at each level is selected based on the **Gini index**.

NEURAL NETWORK

- Inspired by biological neural systems in animal brains.
- Dense network of neurons with large numbers of interconnections that can solve complex learning problems.
- Features in a neural network are obtained throughout the learning process and are not provided manually.
- A typical neural net consists of several input neurons (nodes) arranged in the input layer, then connected to hidden units in the hidden layer.

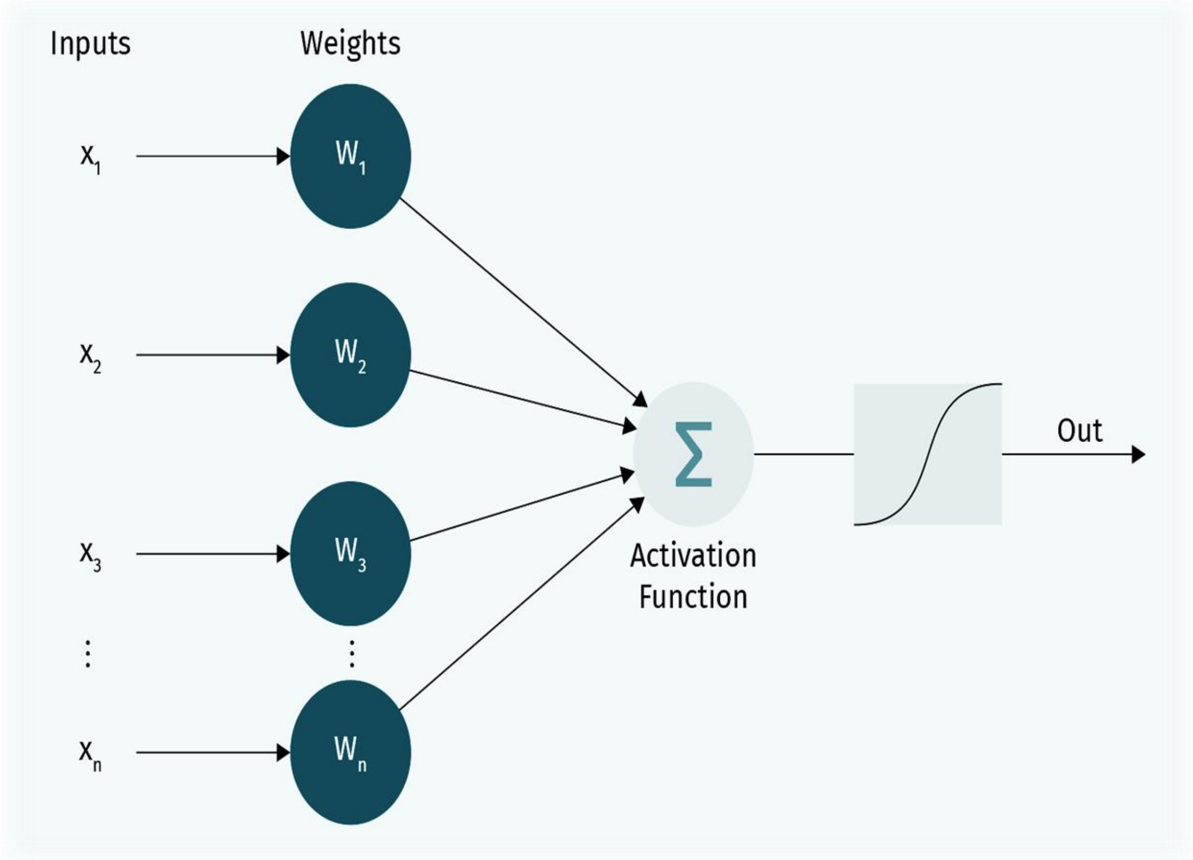
AN ARTIFICIAL NEURON

- Each link between a neuron and its input has a weight.

$$\mathrm{net} = \sum \mathrm{w_i x_i} = \mathrm{w_1 x_1} + \cdots + \mathrm{w_n x_n}$$

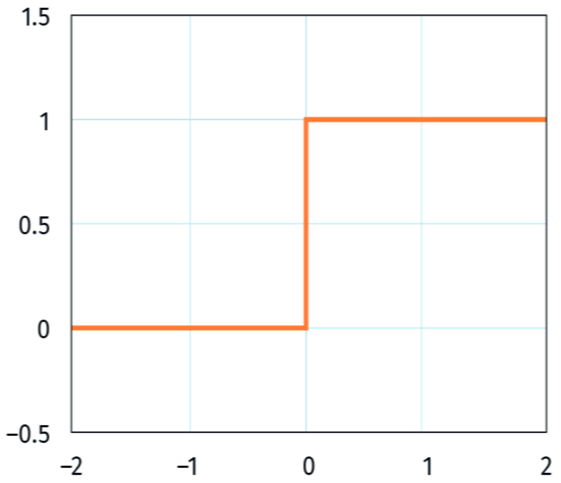
- An activation function is applied to determine the output.

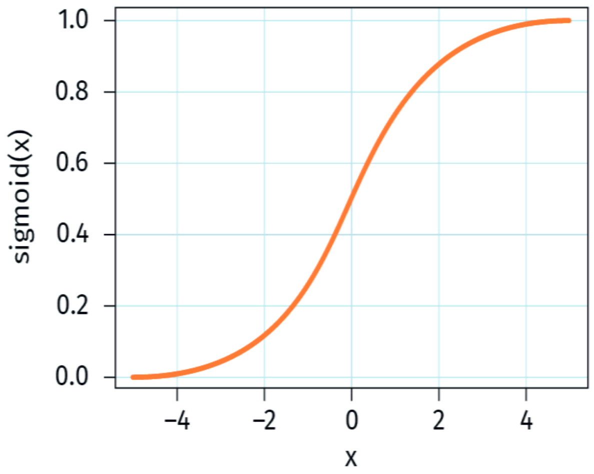
$$\mathrm{y} = \mathrm{f(net)}$$

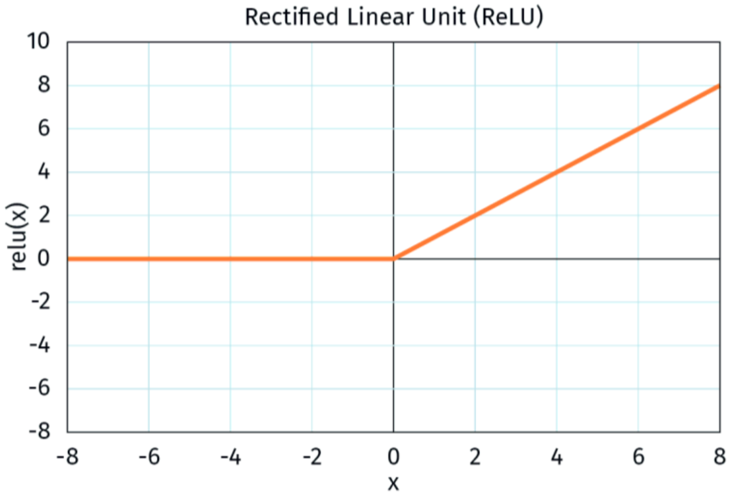


Source of the graphic: Course Book DLMBBD01, p. 48

ACTIVATION FUNCTIONS







Step Function

- Simulates the behavior of biological neurons.
- Defines a threshold for neural activation.

Logistic Sigmoid

- Nonlinear function (continuous version of step)

$$f(x) = \frac{1}{1 + e^{-x}}$$

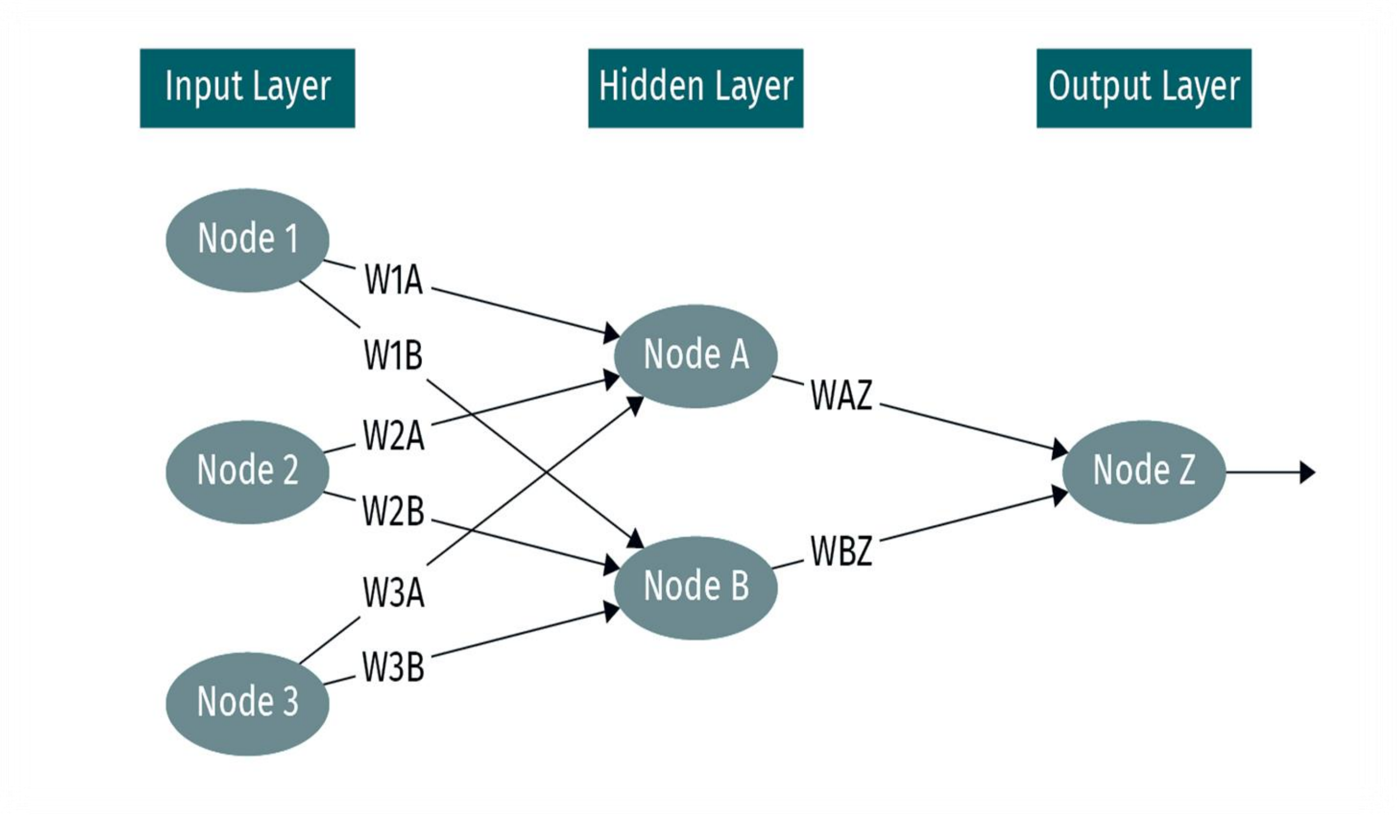
Rectified Linear Unit (ReLU)

$$\mathrm{ReLU}(x) = \max(x.0) = \begin{cases} x.x > 0 \\ 0.x \leq 0 \end{cases}$$

Source of the graphics: Course Book DLMBBD01, pp. 49-50

NETWORK ARCHITECTURE

- Fully-connected network (each neuron in a layer is connected to all neurons in the next layer)
- Feedforward network (signal flows in a forward direction, no closed loop)



Source of the graphic: Larose & Larose, 2014.

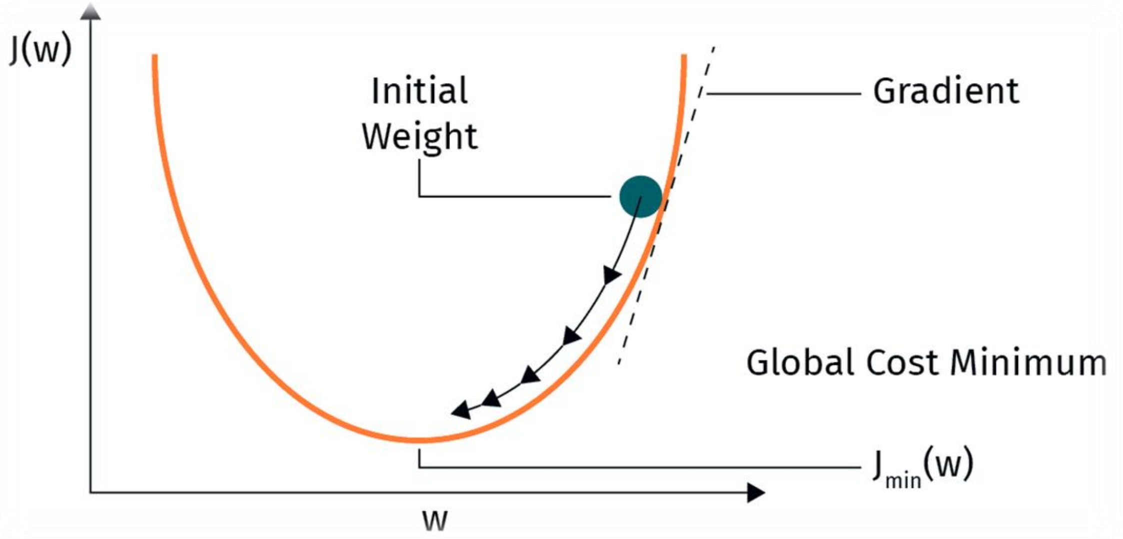
NEURAL NETWORK EVALUATION

- Learning = learning weights
- Weights are initialized with random values or with a special procedure like Glorot initialization.
- Error for training sample: the discrepancy between calculated output and the predetermined target value
- Methods of measuring discrepancy:
  - **sum of squares error (SSE)**
  - **cross entropy error function**

SSE

$$\text{SSE} = \sum\nolimits_{\text{instances}} \sum\nolimits_{\text{output nodes}} (\text{actual value} - \text{predicted value})^2$$

- Find a set of weights to minimize SSE
- Gradient descent method as the optimization method to specify the direction in which the weight should be adjusted



$$\nabla \text{SSE}(W) = \left[\frac{\partial\,\text{SSE}}{\partial\,w_0}, \frac{\partial\,\text{SSE}}{\partial\,w_1}, \ldots, \frac{\partial\,\text{SSE}}{\partial\,w_m}\right]$$

$$W^{(\tau+1)} = W^{(\tau)} - \propto \nabla \text{SSE}(W)$$

Source of the graphic: Course Book DLMBBD01, p. 53

LIMITATIONS OF PREDICTION

- Limitations of using data mining or machine learning techniques in predicting future events:
  - **dependence on historical data**
  - **dependence on the quality of data**
  - **bias**
  - **overfitting**
  - **selectivity**

REVIEW STUDY GOALS

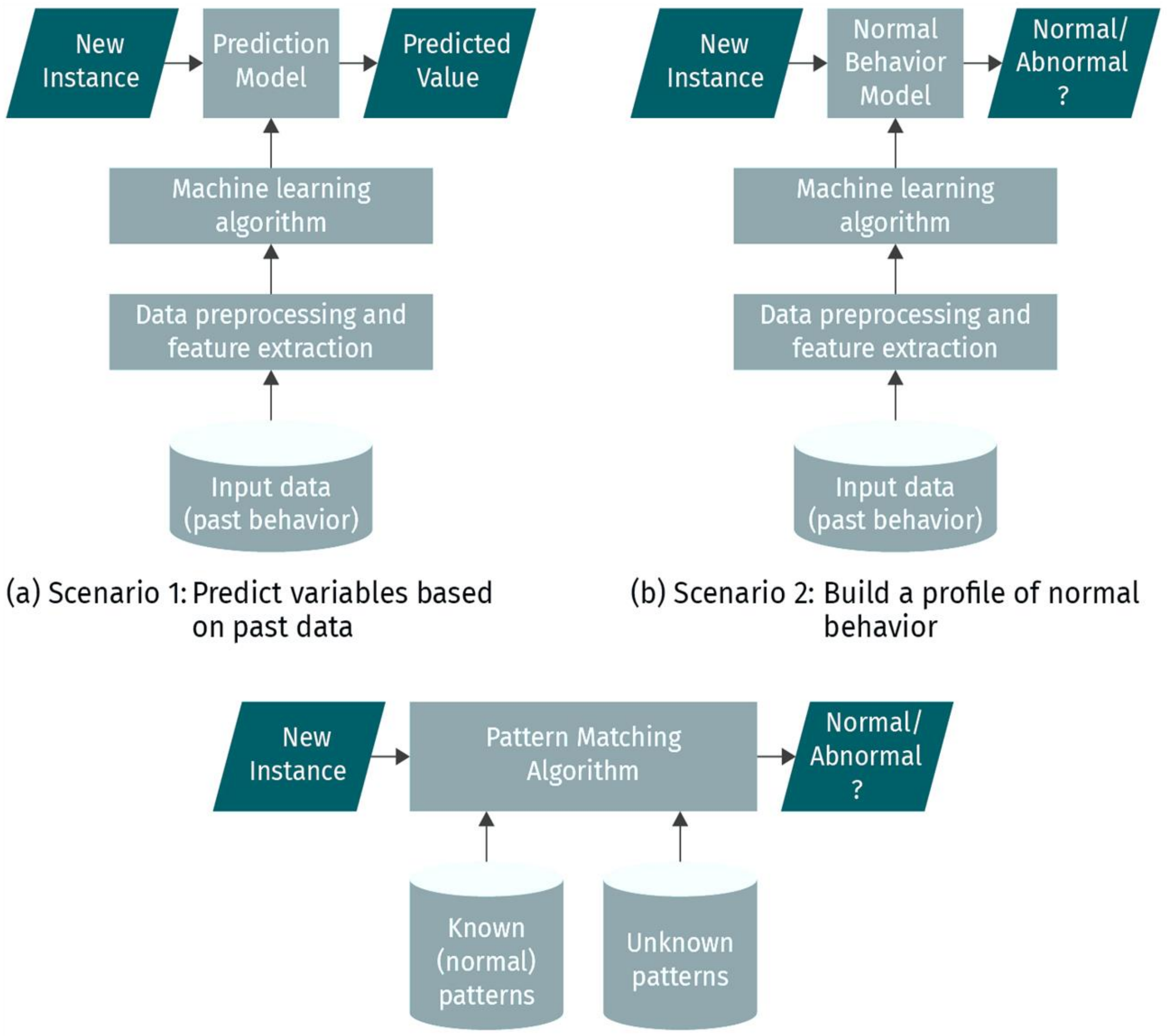
On completion of this unit, you will have learned …

- ... definitions of pattern recognition and its real world applications
- ... definitions of user attitude, behavior, dissonance, and consonance
- ... the concept of big data and its main characteristics
- ... how to conduct research using online data such as Google Trends analytics
- ... the predictive analytics process
- ... well-known predictive algorithms such as regression, decision tree, and neural network.

SESSION 2

TRANSFER TASK

Based on the three following proposed models of distinguishing normal events from abnormal events in an existing database, **try to explain** the how cyber security systems protect the systems from cybercrimes and reduce the associated costs.


New Instance
Prediction Model
Predicted Value
Machine learning algorithm
Data preprocessing and feature extraction
Input data (past behavior)
New Instance
Normal Behavior Model
Normal/ Abnormal ?
Machine learning algorithm
Data preprocessing and feature extraction
Input data (past behavior)
New Instance
Pattern Matching Algorithm
Normal/ Abnormal ?
Known (normal) patterns
Unknown patterns

(a) Scenario 1: Predict variables based on past data

(b) Scenario 2: Build a profile of normal behavior

(c) Scenario 3: Have an existing database of both normal and abnormal events

Source of the graphic: Course Book DLMBBD01, p. 56

TRANSFER TASK
PRESENTATION OF THE RESULTS

Please present your results.

The results will be discussed in plenary.

1. Which statement about decision trees is **incorrect**?

   a) All decision tree algorithms use information entropy to construct the tree.
   b) A decision tree is a classification technique.
   c) A single decision tree is easy to interpret.
   d) A set of decision trees (forest) produces more.

LEARNING CONTROL QUESTIONS

2. Which technique is used to minimize the error of a neural network?

a) sigmoid function
b) sum of squares error
c) gradient descent
d) randomized weights

LEARNING CONTROL QUESTIONS

3. Which of the following is a stochastic model used to describe a sequence of events where the probability of the next event is a function of the current state only.

a) regression model
b) Markov model
c) neural network model
d) entropy

LIST OF SOURCES

Erkeç, E. (2018, January 12). *Creating simple linear regression in Azure machine learning* [guide]. https://codingsight.com/creating-simple-linear-regression-azure-machine-learning

Larose, D. T. & Larose, C. D. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (3rd ed.). Wiley.

Preis, T., Reith, D. & Stanley, H. E. (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368*(1933). 5707-5719.

Rumelhart, David E., Hinton, G. E. & Williams, R. J. (1986).Learning representations by back-propagating errors. *Nature 323*(6088). 533-536.

Stephens-Davidowitz, S. I. (2012). The effects of racial animus on a black presidential candidate: using Google search data to find what surveys miss. *Available at SSRN 2050673.*

Sumeet, D. & Du, X. (2011). *Data mining and machine learning in cybersecurity*. CRC press, 2016.

Vujić, S. & Zhang, X. (2018). Does Twitter chatter matter? Online reviews and box office revenues. *Applied Economics 50*(34-35). 3702-3717.