Machine Learning Techniques for Text

Module 1: Intro to Machine Learning

Dr. Nikos Tsourakis

Course outline

‹packt›
Machine Learning Techniques for Text
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- Module 0: Python Crash Course
- **Module 1: Intro to Machine Learning**
- Module 2: Detecting Spam Emails
- Module 3: Classifying Topics of Newsgroup Posts
- Module 4: Extracting Sentiments from Product Reviews
- Module 5: Recommending Music Titles
- Module 6: Teaching Machines to Translate
- Module 7: Summarizing Wikipedia Articles
- Module 8: Detecting Hateful and Offensive Language
- Module 9: Generating Text in Chatbots
- Module 10: Clustering Speech-to-Text Transcriptions

Overview

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- Inspired by the expressive power of human texts, this introductory module sets the scene for the discussion in the following ones, where we examine how to teach machines to extract meaningful interpretations from text corpora

- Building machines that learn from observations is becoming the dominant paradigm due to the ever-increasing amount of data that cannot be processed using traditional methods

- Text data is produced in vast quantities through social network interactions, scientific publications, and transcribing multimedia streams, among other things

- We introduce the main techniques for ***machine learning*** (ML) for text, the relevant terminology, and the implications while using text corpora

Module objectives

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

After completing this module, you should be able to:

- Introducing the peculiarities of human language
- Understanding how machines learn
- Identifying the basic taxonomy of machine learning algorithms
- Learning the basic terminology
- Understanding the importance of visualization and evaluation techniques

Machine Learning Techniques for Text

Section 1: Introduction

The language phenomenon

‹packt›
Machine Learning Techniques for Text
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- Human language is a structured communication system based on grammar and vocabulary with distinctive features:
  - Combine or recombine sets of words and create new sentences with little effort (***compositional***)
  - Can refer to people, objects, or situations (***referential***)
  - Use the auditive modality while speaking, the tactile modality in the Braille system and the visual modality for the sign language (***modality-independent***)
- Linguistics is the main field for studying human languages and applying scientific methods to questions about their nature and function
- Many of these questions overlap with other fields in the life sciences, social sciences, and humanities, making the study of languages multidisciplinary
- In the new machine age era, delegating the effort of analyzing human language to a computer is an attractive option

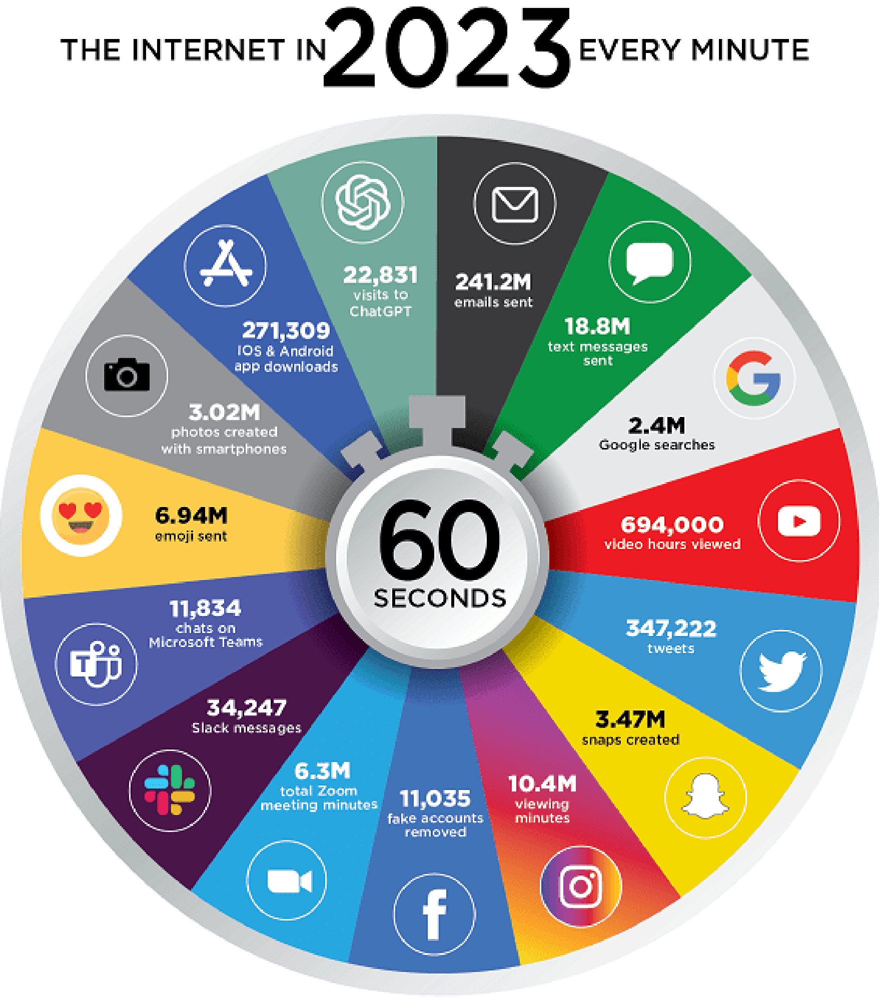
The data explosion

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- We live in a data-driven world that steadily becomes even more data-driven
- The innate tendency of humans to impart information, especially in written form, has caused an abundance of data for various languages and domains
- Besides people's willingness to share information, advances in computer connectivity and storage have paved the way for an explosion in the volume of text data



Created by: eDiscovery Today & LTMG

The era of AI

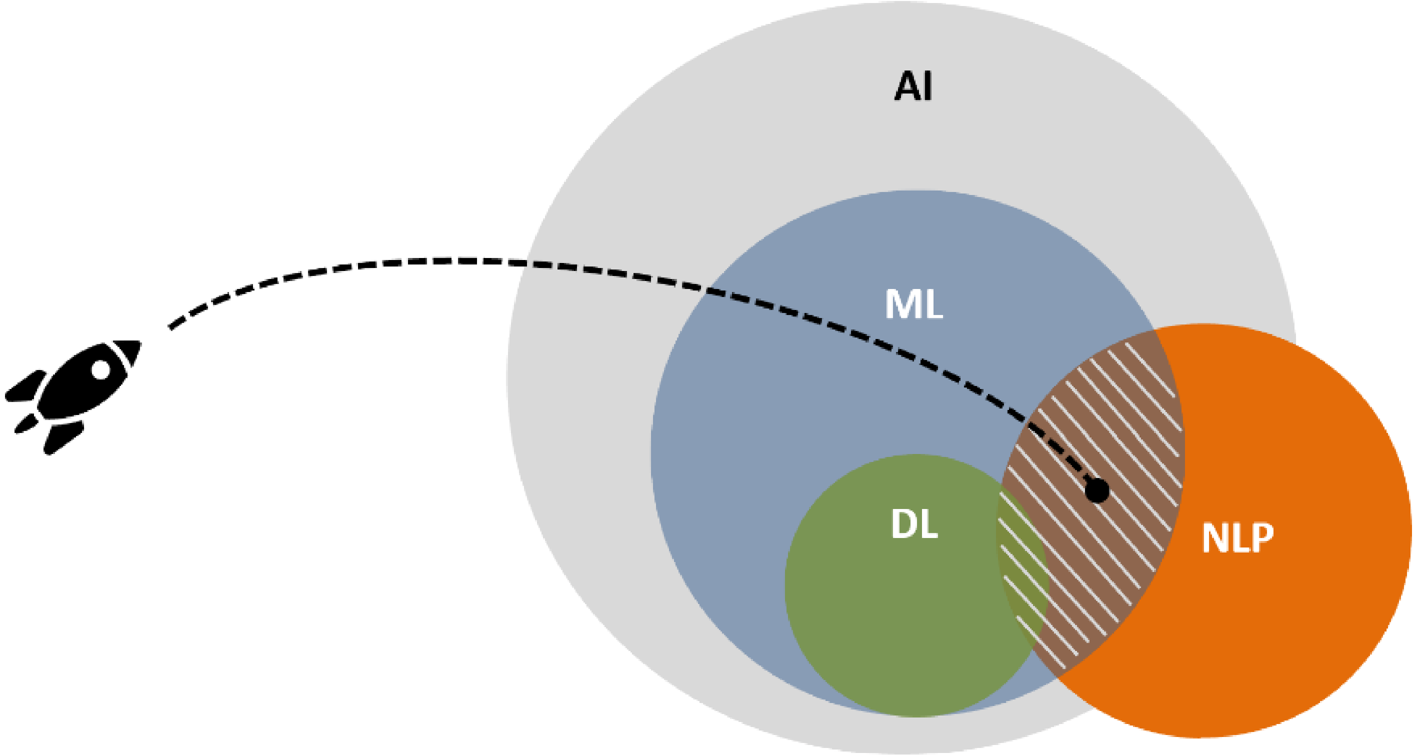
‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- AI has been around for decades, why did it just start taking off now?
- Three main reasons were the driving forces for this situation
  - **Data availability**: Digital devices, such as laptops and smartphones, are now an extension of the human body, generating vast amounts of data we can feed our learning algorithms – for example, email text, tweet posts, or video and audio transcriptions
  - **Computational scale**: The advancement in hardware permitted the creation of intelligence models that are big enough to take advantage of the huge datasets currently available
  - **New algorithms**: The AI community has grown significantly, which has led to the creation of more powerful algorithms

Our landing point



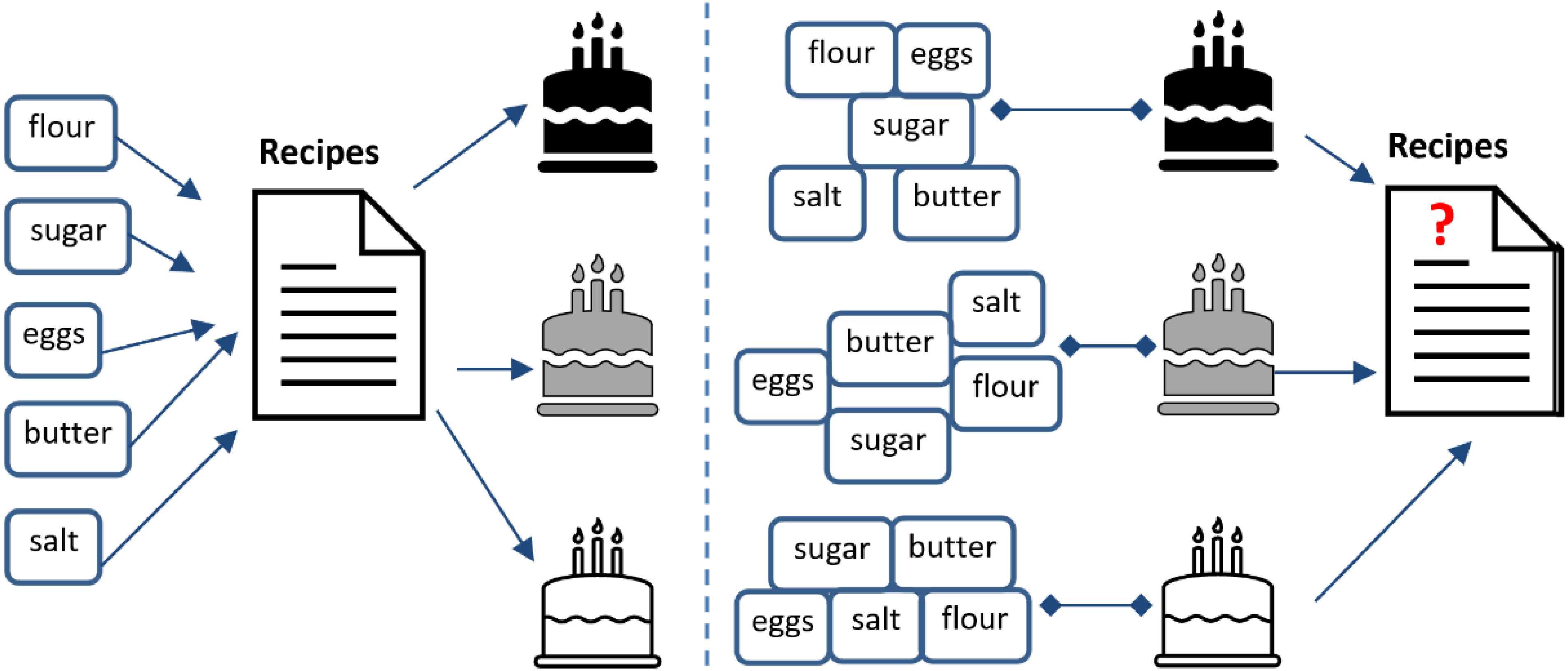
AI – Artificial intelligence, **ML** – Machine Learning, **DL** – Deep Learning, **NLP** – Natural Language Processing

The machine learning paradigm

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation



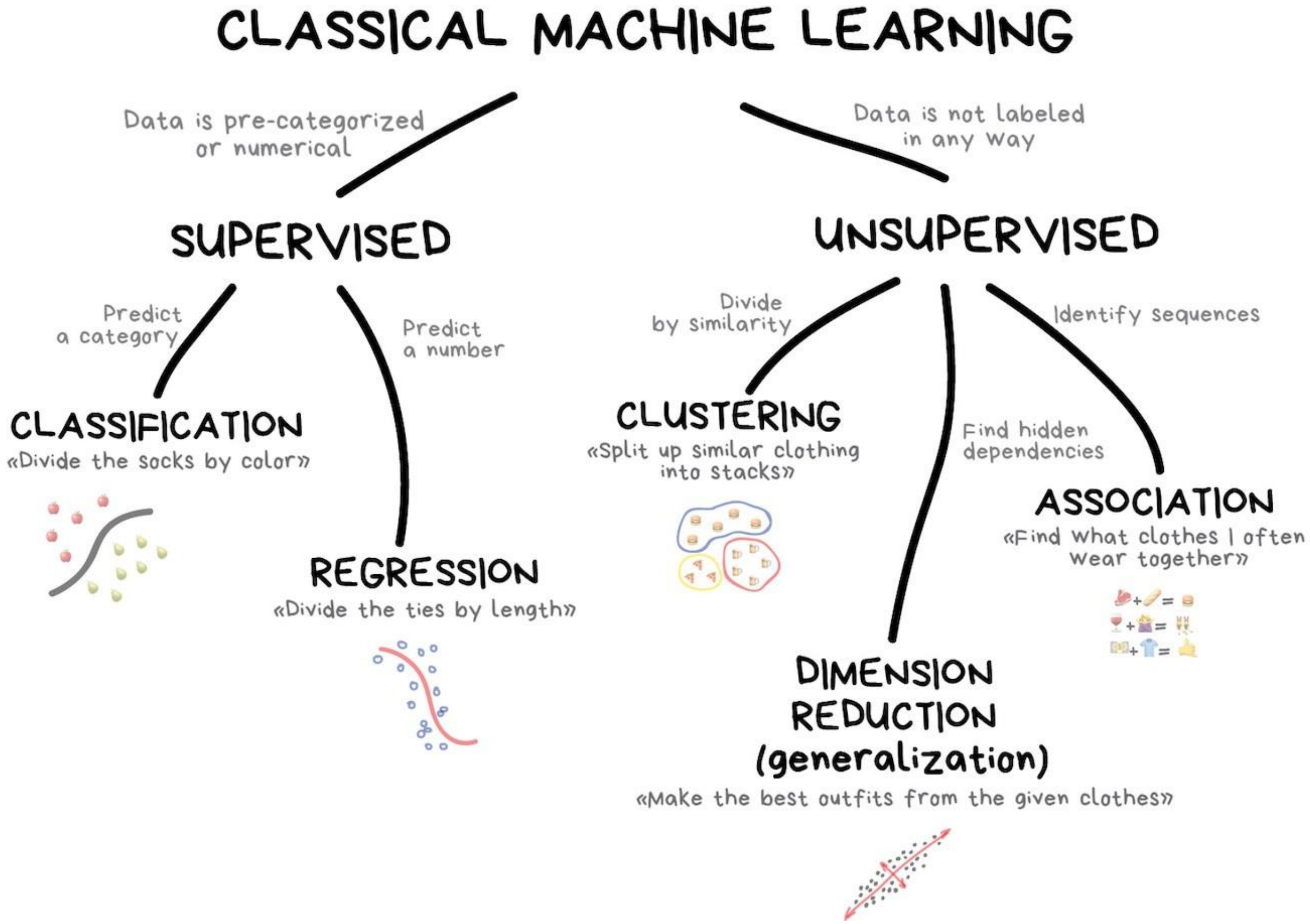
Traditional versus new software development paradigm

Taxonomy of machine learning techniques

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

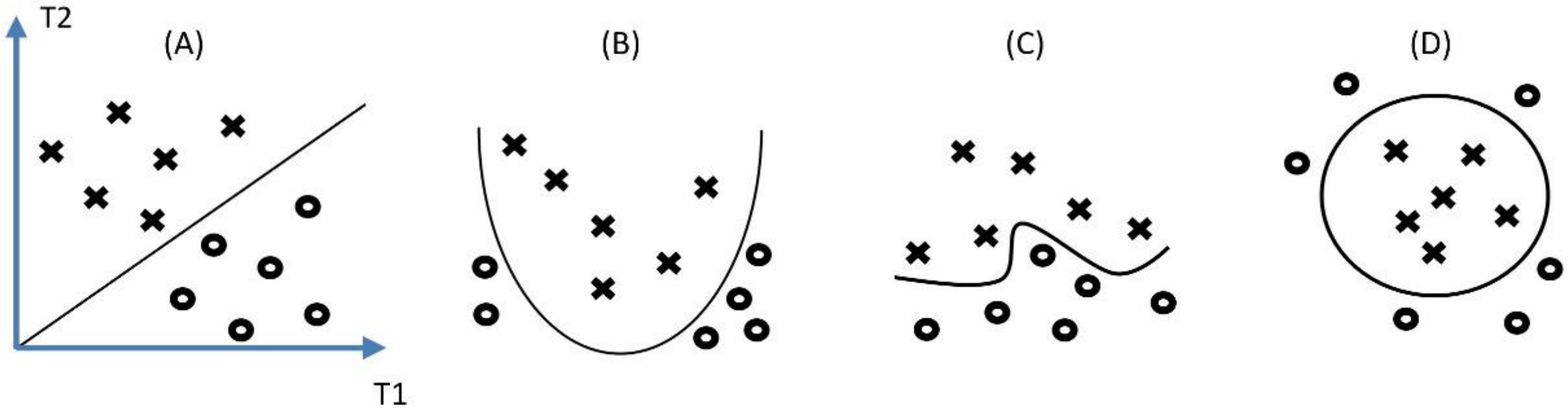

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
Find hidden dependencies
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

source: https://vas3k.com/blog/machine_learning/

Classification

‹packt›
Machine Learning Techniques for Text
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- Data points in 2-D space. E.g. ***spam***/***non-spam*** emails
- A machine learning algorithm can learn the equation of the lines that separate the data



Classify two groups of people

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

What is a possible separation line?

● Group A

○ Group B

$x = 60$

$\mathbf{y} = ax + b$

height					
1.7 - 1.8 m	○	○	●	●	●
1.6 - 1.7 m	○	○	●	●	●
1.5 - 1.6 m	○	○	●	●	●
	40 – 50 Kg	50 – 60 Kg	60 – 70 Kg	70 – 80 Kg	80 – 90 Kg

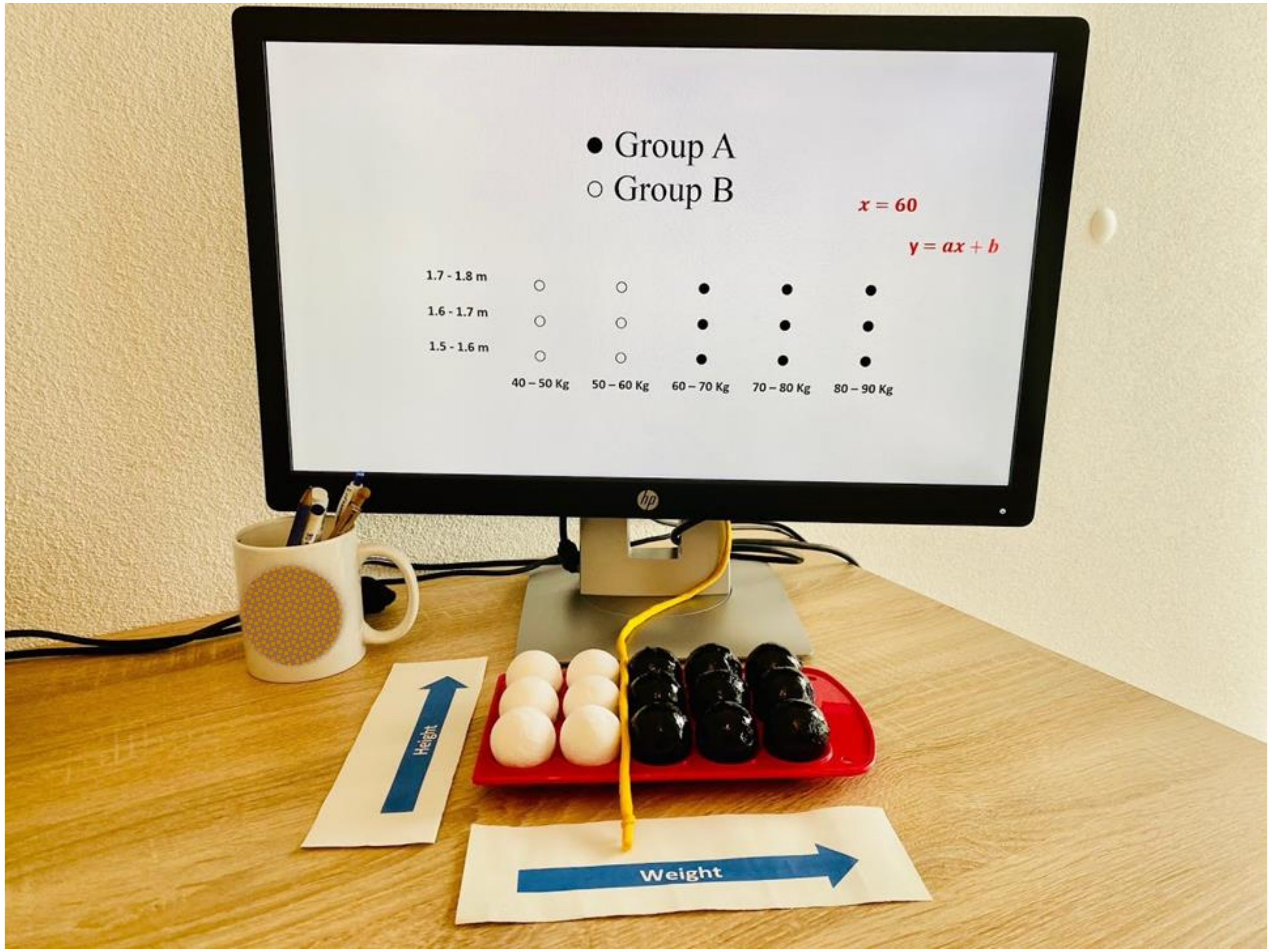
weight

Classify two groups of people

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation



Classify two groups of people

‹packt›
Machine Learning Techniques for Text
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

What is a possible separation line?

● Group A

○ Group B

$\mathbf{y} = \boldsymbol{sin}(\boldsymbol{x})$

height

	40 – 50 Kg	50 – 60 Kg	60 – 70 Kg	70 – 80 Kg	80 – 90 Kg
1.7 - 1.8 m	●	○	●	○	●
1.6 - 1.7 m	●	○	●	○	●
1.5 - 1.6 m	●	○	●	○	●

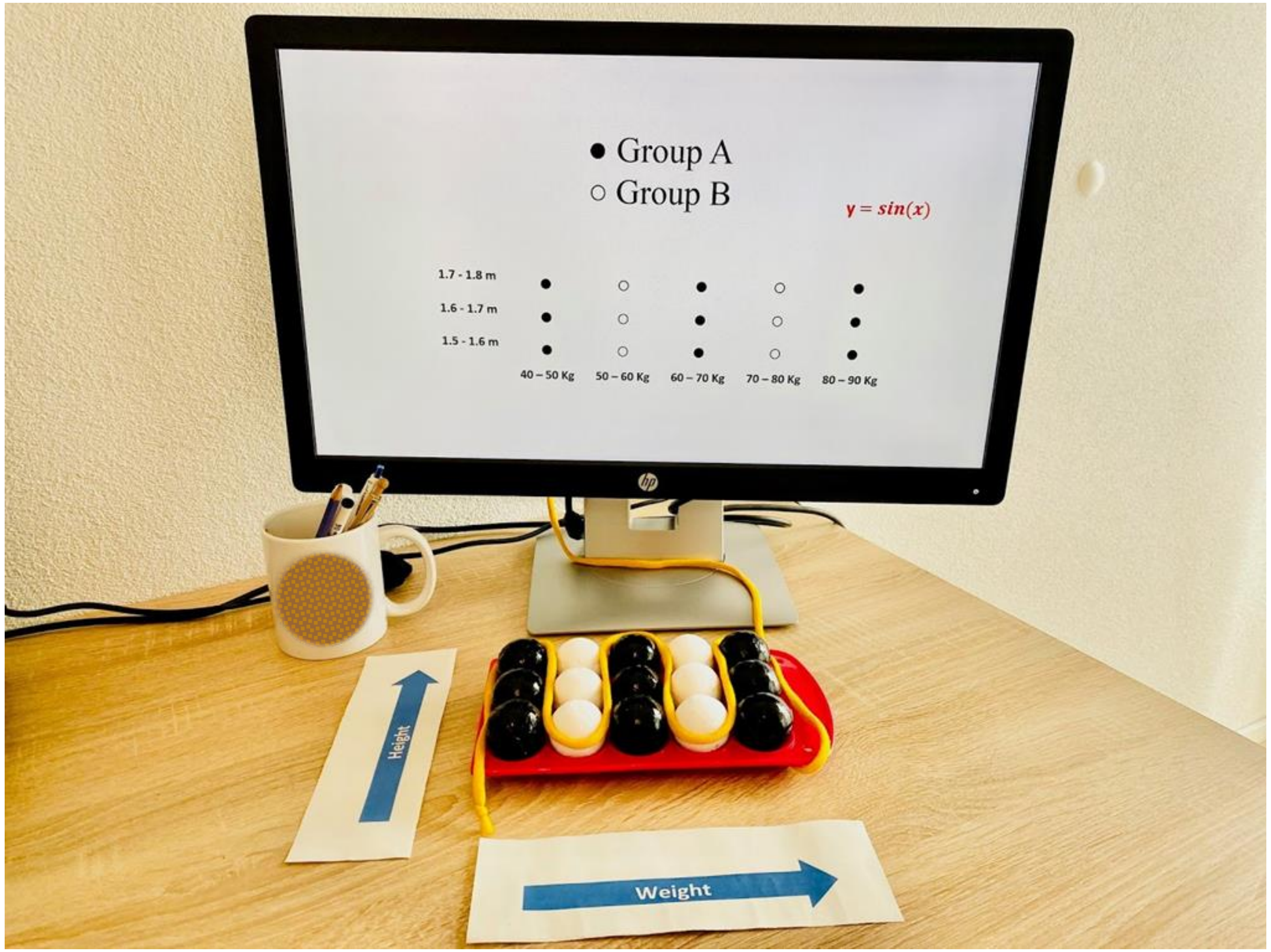
weight

Classify two groups of people

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation



Classify two groups of people

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

What is a possible separation line?

● Group A
○ Group B

?

height	40 – 50 Kg	50 – 60 Kg	60 – 70 Kg	70 – 80 Kg	80 – 90 Kg
1.7 - 1.8 m	○	●	○	●	○
1.6 - 1.7 m	●	○	●	○	●
1.5 - 1.6 m	○	●	○	●	○

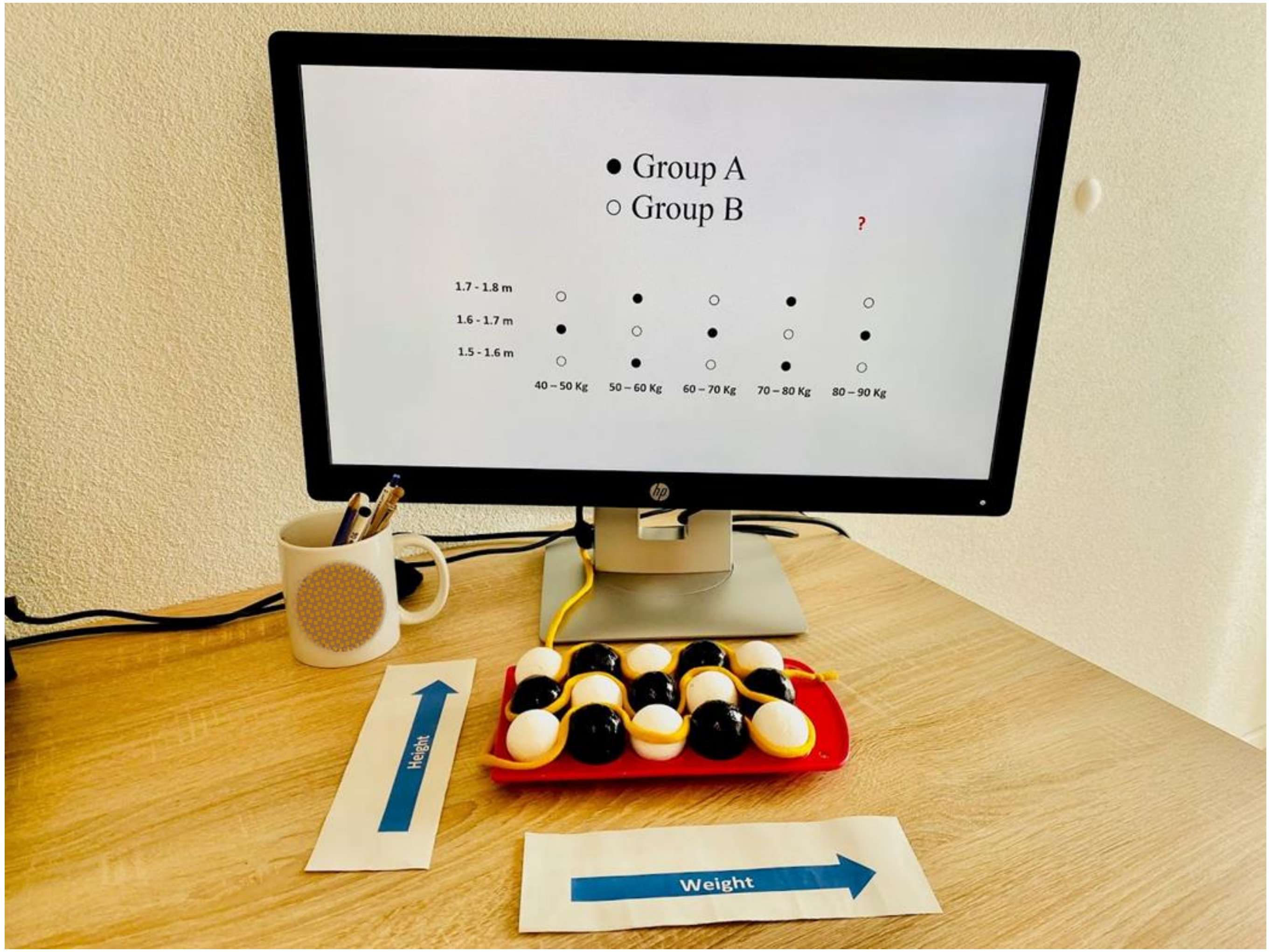
weight

Classify two groups of people

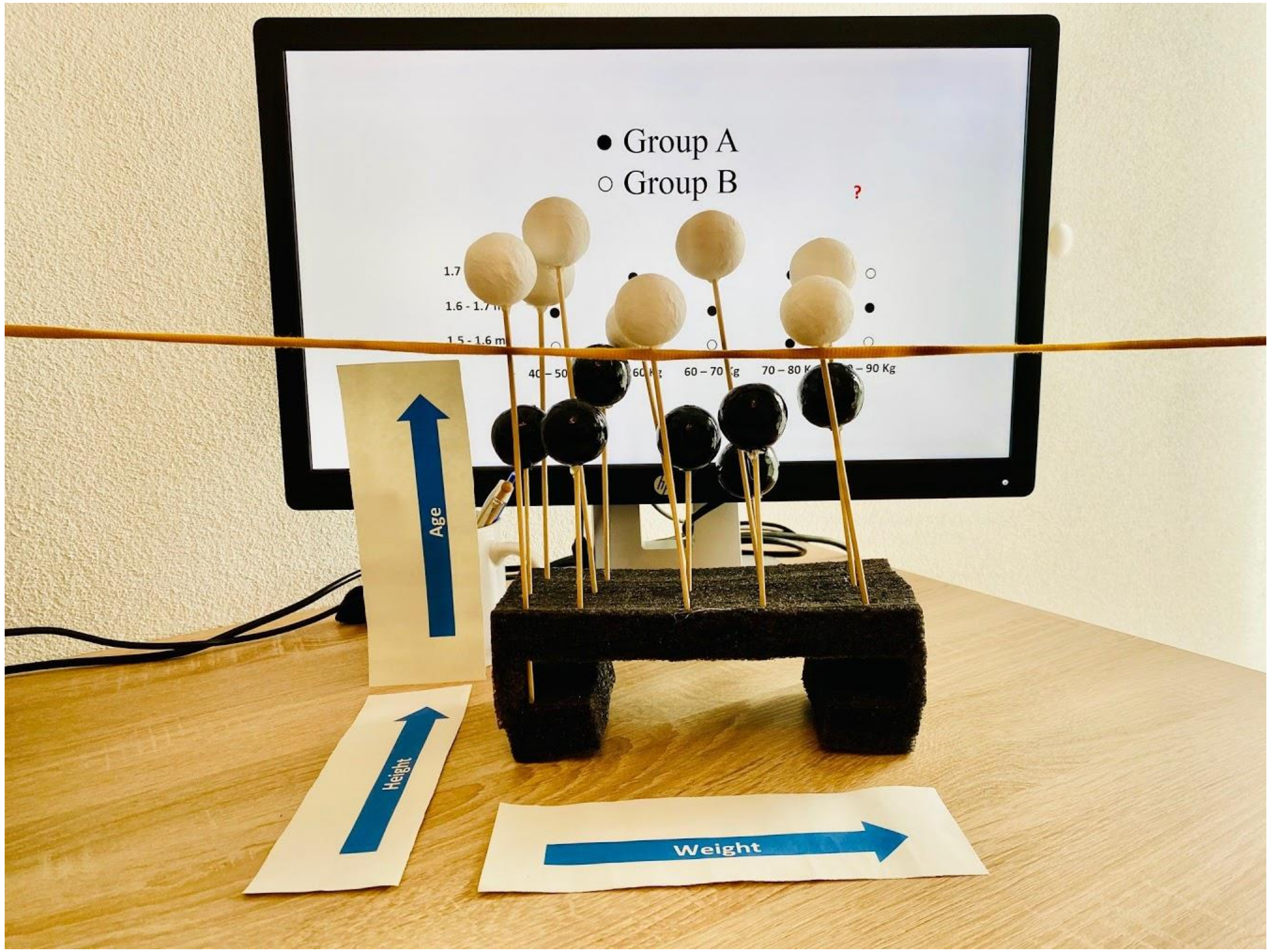
‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

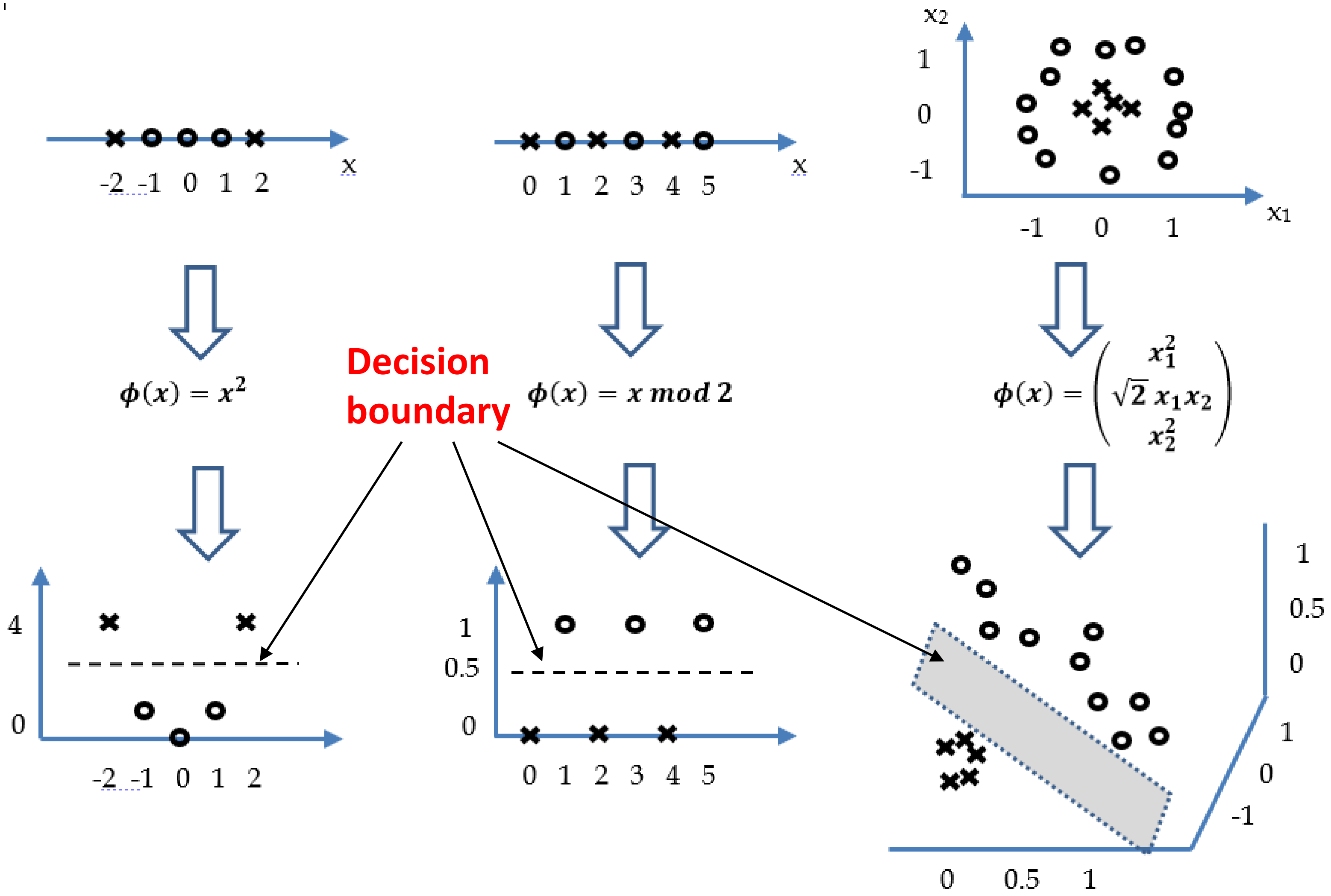


Classify two groups of people



Transform to higher dimensional spaces

‹packt›
Machine Learning Techniques for Text
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation



$$\phi(x) = x^2$$

$$\phi(x) = x \bmod 2$$

$$\phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{pmatrix}$$

Decision boundary

Higher dimensional spaces

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- Normally ML algorithms work in higher dimensions
- The human brain is wired to think in 3D it is difficult to visualize how data is transformed in a multi-dimensional space
- Intuitive analogy:
    - People can be classified according to five personality traits: extraversion, agreeableness, openness, conscientiousness, and neuroticism
    - We can be described as a single point in a 5D space

The ***Big Five*** *personality traits is a taxonomy for human personality and psyche*

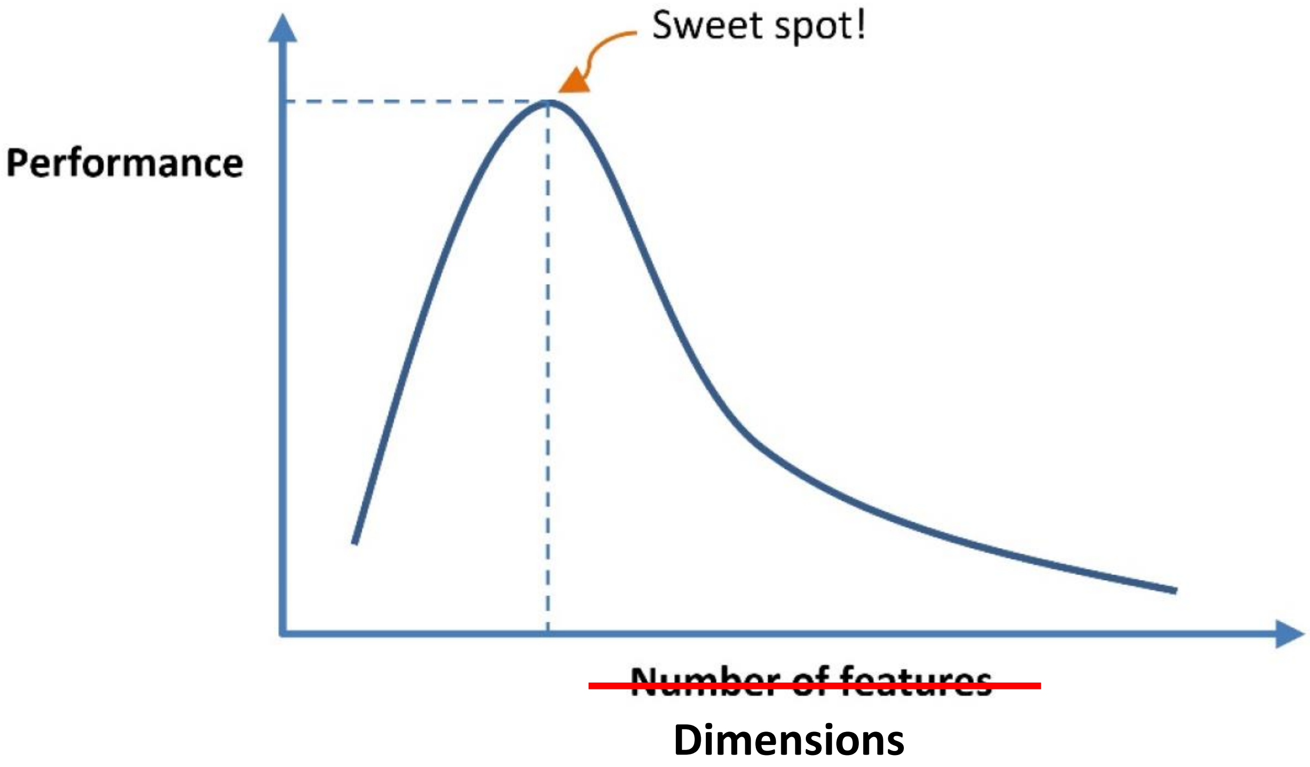
Peaking phenomenon

‹packt›

Machine Learning Techniques for Text

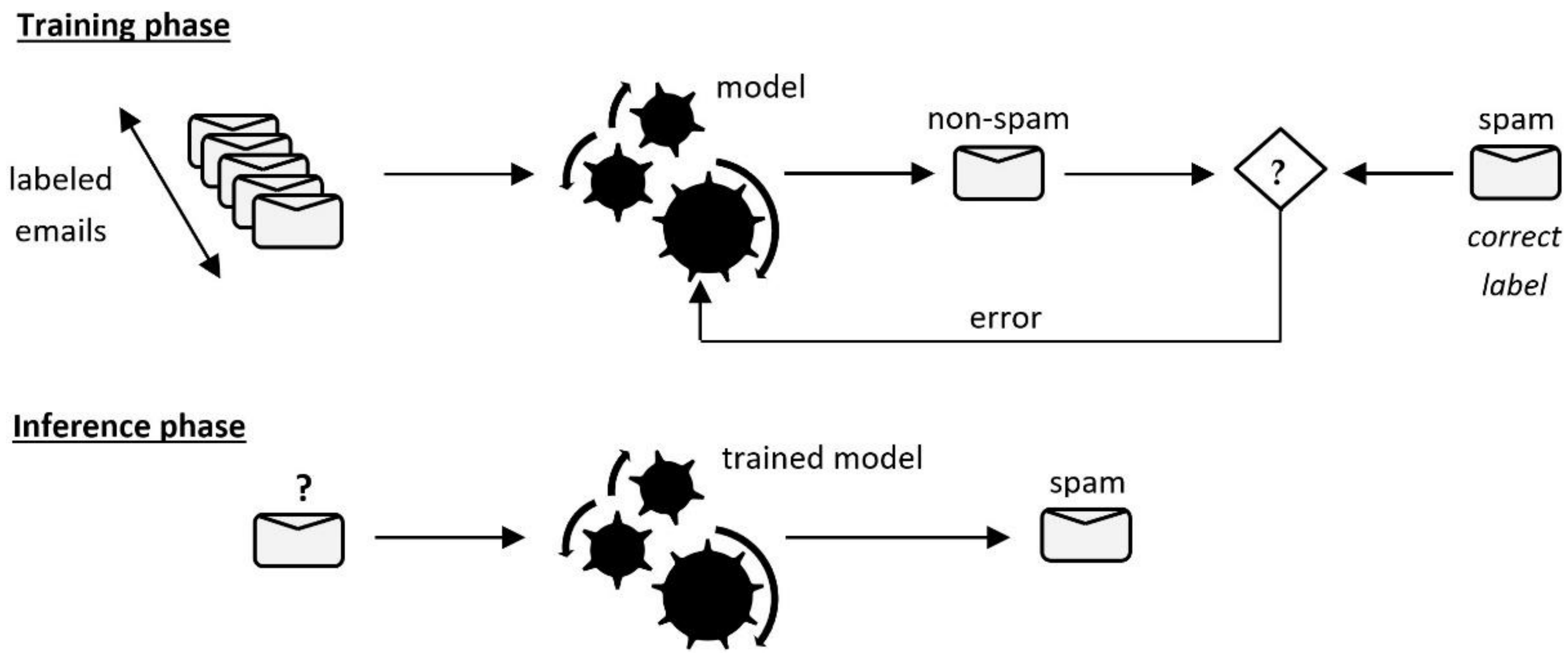
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- The performance of a model steadily increases as more features (or dimensions) are added and starts to deteriorate after a certain threshold is reached



Classification

- The process is split into two phases; namely ***training*** and ***inference***
- ***Supervised learning***: dataset with labeled data is used



Is it a classification, clustering or regression problem?

Classification	➢ Given an email, decide if it is spam or not
Classification	➢ Given a handwritten character, identify it as one of the known characters
Regression	➢ Predict the value of a house in 10 years
Clustering	➢ Segment customers according to their purchase habits
Regression	➢ Find the relationship between students’ GPA and the overall study hours
Clustering	➢ Identify players that are similar to each other
Regression	➢ Identify the revenue based on the money spent on online advertising
Classification	➢ Given a set of images of fruits, determine if they represent an orange, apple, or pear

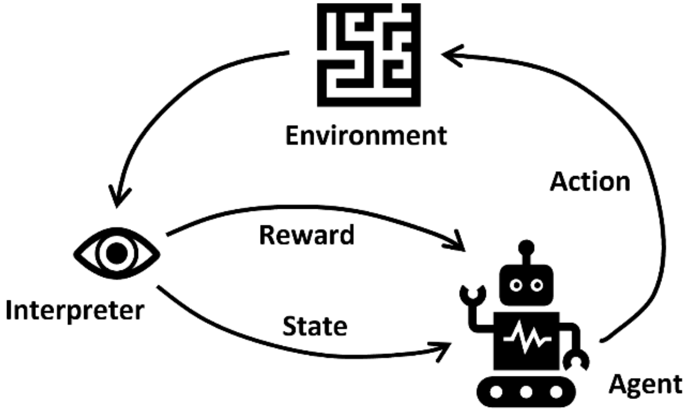
Reinforcement learning

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- ***Reinforcement learning*** problems are markedly different from the ones of supervised, unsupervised, and semi-supervised categories
- Reinforcement learning is the task of learning through trial and error, having an agent take actions within an environment
- The agent is most commonly an algorithm that must discover through interaction with its environment which sequence of actions is the best to accomplish a given goal



https://youtu.be/spfpBrBjntg?si=b8v-cr4vuHIpPwgn

Visualization of the data

‹packt›

Machine Learning Techniques for Text

Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- The vast majority of all human communication is visual; we are wired to understand images instantly while we need time to process text
- Visuals make it much easier to spot patterns and identify anomalies, which is critical to people working with data
- A good visualization should encompass three characteristics: being ***trustworthy***, ***accessible***, and ***elegant***
- In the exercises we use visualization techniques from the data-driven categories extensively to attack two main problems
  - first, to extract some evident information or identify possible problems with the data before resorting to analysis
  - and second, to report on the performance of the implemented systems

Evaluation of the results

‹packt›
Machine Learning Techniques for Text
Apply modern techniques with Python for text processing, dimensionality reduction, classification, and evaluation

- Determining the value or worth of something in terms of quantity and quality is the process of ***evaluation***
- The increasing sophistication of text systems necessitates evaluation frameworks that measure the stated objectives and anticipated results
- Make use of various ***computer***- and ***human-centered*** metrics, most commonly known as objective and subjective evaluation
- It is not uncommon that during the optimization of one metric, performance has deteriorated on another one
- Another crucial aspect is prioritizing the errors based on their severity for the given problem

Machine Learning Techniques for Text

Questions?