Machine Learning (DLMDSML01)

Nghia Duong-Trung[1]

[1] **German Research Center for Artificial Intelligence (DFKI GmbH)**

Internal Use @ IU International University of Applied Science, Berlin Campus

SPONSORED BY THE

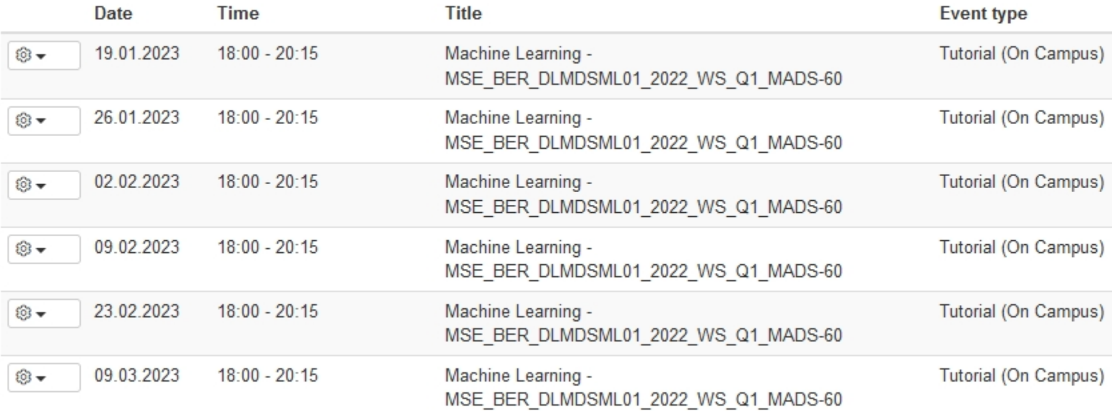
Federal Ministry
of Education
and Research

Who I Am

German Research Center for Artificial Intelligence

- Name: Nghia Duong-Trung
- Senior Researcher @ The German Research Center for Artificial Intelligence
  - BMBF Projects: `https://milki-psy.de/`, `https://tech4comp.de/`
- Freelance Lecturer @ IU
- PostDoc @ Technische Universität Berlin, Germany
  - BMBF Project: `https://kiwi-biolab.de/`
- PhD @ University of Hildesheim, Germany
- Profile: `https://sites.google.com/ismll.de/duongtrungnghia/`

Study Schedule



	Date	Time	Title	Event type
⚙ ▾	19.01.2023	18:00 - 20:15	Machine Learning - MSE_BER_DLMDSML01_2022_WS_Q1_MADS-60	Tutorial (On Campus)
⚙ ▾	26.01.2023	18:00 - 20:15	Machine Learning - MSE_BER_DLMDSML01_2022_WS_Q1_MADS-60	Tutorial (On Campus)
⚙ ▾	02.02.2023	18:00 - 20:15	Machine Learning - MSE_BER_DLMDSML01_2022_WS_Q1_MADS-60	Tutorial (On Campus)
⚙ ▾	09.02.2023	18:00 - 20:15	Machine Learning - MSE_BER_DLMDSML01_2022_WS_Q1_MADS-60	Tutorial (On Campus)
⚙ ▾	23.02.2023	18:00 - 20:15	Machine Learning - MSE_BER_DLMDSML01_2022_WS_Q1_MADS-60	Tutorial (On Campus)
⚙ ▾	09.03.2023	18:00 - 20:15	Machine Learning - MSE_BER_DLMDSML01_2022_WS_Q1_MADS-60	Tutorial (On Campus)

IU: New from Q12023

- Check attendance
  - Attendance or partial attendance
  - Excuse note (yes | no)
  - Absence reason (yes | no)
- Regularly take a screenshot in Zoom

Section 1

- Textbook: check it in MyCampus
- Brief introduction about machine learning
- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

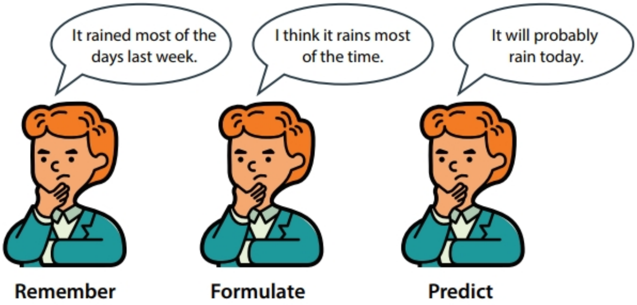
Brief machine learning introduction

ML, AI

- To define machine learning, first let's define a more general term: artificial intelligence.
    - The set of all tasks in which a computer can make decisions
    - A computer makes these decisions by mimicking the ways a human makes decisions
        - by using logic and reasoning
        - by using our experience
- The set of all tasks in which a computer can make decisions based on **data**
- Machine learning is a part of artificial intelligence

Making decisions with data

- Remember-formulate-predict framework
  - We **remember** past situations that were similar
  - We **formulate** a general rule
  - We use this rule to **predict** what may happen in the future


It rained most of the days last week.
I think it rains most of the time.
It will probably rain today.
Remember
Formulate
Predict

Models and algorithms

- We know that in machine learning, we get the computer to learn how to solve a problem using data. The way the computer solves the problem is by using the data to build a **model**
  - A set of rules that represent our data and can be used to make predictions
- An **algorithm** is the process that we used to build the model
  - Example: we looked at how many days it rained and realized it was the majority
  - A procedure, or a set of steps, used to solve a problem or perform a computation

Transfer Task

- in group, short presentation
  - How a human learn to recognize an object? And how we transfer it into data?

Spam detection 1

- We will detect spam and non-spam emails. Non-spam emails are also referred to as ham.

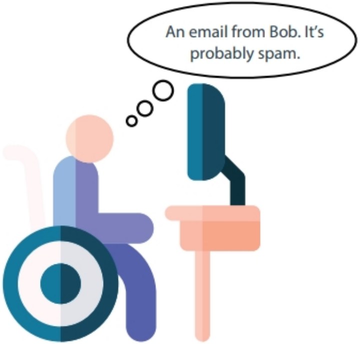
- Scenario: our friend Bob likes to send us email. A lot of his emails are spam, in the form of chain letters. We are starting to get a bit annoyed with him. It is Saturday, and we just got a notification of an email from Bob. Can we guess if this email is spam or ham without looking at it?

- To figure this out, we use the remember-formulate-predict method

- First, let us remember, say, the last 10 emails that we got from Bob. That is our data. We remember that six of them were spam, and the other four were ham. From this information, we can formulate the following models

Spam detection 1: model 1

- Model 1: Six out of every 10 emails that Bob sends us are spam
    - This rule will be our model. Note, this rule does not need to be true. It could be outrageously wrong. But given our data, it is the best that we can come up with



Spam detection 2

- Let’s look more carefully at the emails that Bob sent us in the previous month
  - Monday: Ham
  - Tuesday: Ham
  - Saturday: Spam
  - Sunday: Spam
  - Sunday: Spam
  - Wednesday: Ham
  - Friday: Ham
  - Saturday: Spam
  - Tuesday: Ham
  - Thursday: Ham

Spam detection 2: model 2

- Model 2?

Spam detection 3: model 3

- let’s say we continue with this rule, and one day we see Bob in the street, and he asks, "Why didn’t you come to my birthday party?" We have no idea what he is talking about. It turns out last Sunday he sent us an invitation to his birthday party, and we missed it!
- Why did we miss it? Because he sent it on the weekend, and we assumed that it would be spam
  - need a better model! -> How?

Spam detection 3: model 3

- Let's go back to look at Bob's emails and find a pattern
  - 1 KB: Ham
  - 2 KB: Ham
  - 16 KB: Spam
  - 20 KB: Spam
  - 18 KB: Spam
  - 3 KB: Ham
  - 5 KB: Ham
  - 25 KB: Spam
  - 1 KB: Ham
  - 3 KB: Ham
- What rule can we **formulate**?

Spam detection 3: model 3

- Model 3?
- A test case: We look at the email we received today from Bob, and the size is 19 KB. So, we conclude that it is ...

Feature

- notice that to make our predictions, we used the day of the week and the size of the email. These are examples of **features**
  - Any property or characteristic of the data that the model can use to make predictions
- Model 4: If an email is larger than 10 KB or it is sent on the weekend, then it is classified as spam. Otherwise, it is classified as ham.
- Model 5: If the email is sent during the week, then it must be larger than 15 KB to be classified as spam. If it is sent during the weekend, then it must be larger than 5 KB to be classified as spam. Otherwise, it is classified as ham

Spam detection 6: model 6

Model 6: Consider the number of the day, where Monday is 0, Tuesday is 1, Wednesday is 2, Thursday is 3, Friday is 4, Saturday is 5, and Sunday is 6. If we add the number of the day and the size of the email (in KB), and the result is 12 or more, then the email is classified as spam. Otherwise, it is classified as ham

Day	Size (KB)	Target
0	1	Ham
1	2	Ham
5	16	Spam
6	20	Spam
6	18	Spam
2	3	Ham
4	5	Ham
5	25	Spam
1	1	Ham
3	3	Ham

More features

- The goal is to make the computer think the way we think, namely, use the remember-formulate-predict framework. In a nutshell, here is what the computer does in each of the steps:
  - **Remember**: Look at a huge table of data.
  - **Formulate**: Create models by going through many rules and formulas, and check which model fits the data best.
  - **Predict**: Use the model to make predictions about future data

More features, more models

- Model 7
  - If the email has two or more spelling mistakes, then it is classified as spam.
  - If it has an attachment larger than 10 KB, it is classified as spam.
  - If the sender is not in our contact list, it is classified as spam.
  - If it has the words buy and win, it is classified as spam.
  - Otherwise, it is classified as ham.
- Model 8
  - If (size) + 10 (number of spelling mistakes) – (number of appearances of the word “mom”) + 4 (number of appearances of the word “buy”) > 10, then we classify the message as spam. Otherwise, we classify it as ham

More features, more applications

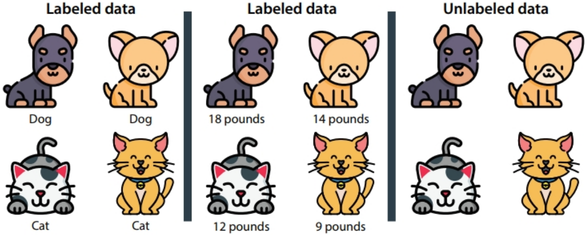
- Predicting house prices based on the house’s size, number of rooms, and location
- Predicting today’s stock market prices based on yesterday’s prices and other factors of the market
- Recognizing images as faces or animals, based on the pixels in the image
- Processing long text documents and outputting a summary
- Recommending videos or movies to a user (e.g., on YouTube or Netflix)
- Building chatbots that interact with humans and answer questions
- Training self-driving cars to navigate a city by themselves
- Diagnosing patients as sick or healthy
- Segmenting the market into similar groups based on location, acquisitive power, and interests
- Playing games like chess or Go

Types of machine learning

Labeled and unlabeled data

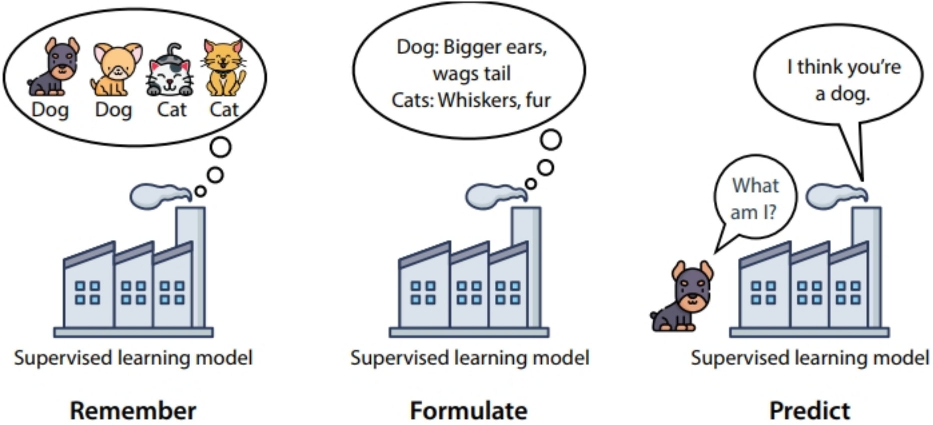
German Research Center for Artificial Intelligence

- **Data**: Any time we have a table with information, we have data
- **Features**: If our data is in a table, the features are the columns of the table
- **Labels/Targets**: what we want to predict
- **Predictions**: The guess that the model makes is called a prediction
- labeled and unlabeled data: Labeled data is data that comes with labels. Unlabeled data is data that comes with no labels



Supervised learning

- Supervised learning: The branch of machine learning that works with labeled data


Dog
Dog
Cat
Cat
Dog: Bigger ears, wags tail
Cats: Whiskers, fur
I think you're a dog.
What am I?
Supervised learning model
Supervised learning model
Supervised learning model
Remember
Formulate
Predict

Types of data

- Numerical data is any type of data that uses numbers such as 4, 2.35, or –199. Examples of numerical data are prices, sizes, or weights

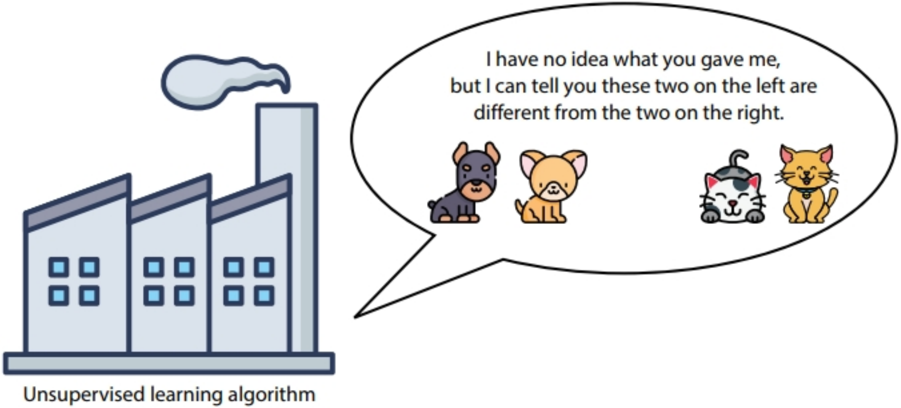
- Categorical data is any type of data that uses categories, or states, such as male/female or cat/dog/bird. For this type of data, we have a finite set of categories to associate to each of the data points.

Types of supervised learning

- Regression models are the types of models that predict numerical data. The output of a regression model is a number, such as the weight of the animal
  - Model 1: housing prices model (regression). In this model, each data point is a house. The label of each house is its price. Our goal is that when a new house (data point) comes on the market, we would like to predict its label, namely, its price.
- Classification models are the types of models that predict categorical data. The output of a classification model is a category, or a state, such as the type of animal (cat or dog)
  - Model 2: email spam–detection model (classification)

Unsupervised learning

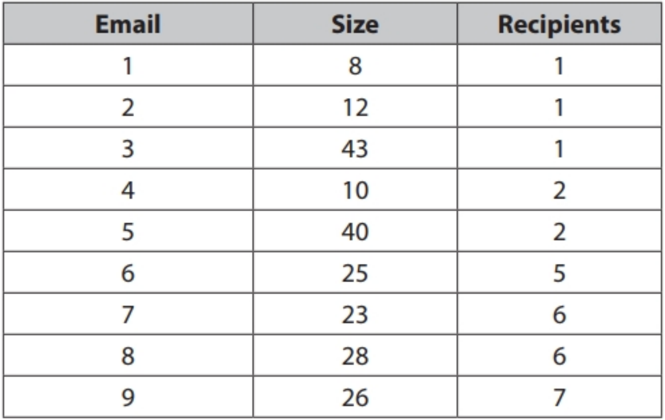
- Unsupervised learning: The branch of machine learning that works with unlabeled data


I have no idea what you gave me, but I can tell you these two on the left are different from the two on the right.
Unsupervised learning algorithm

Branches of unsupervised learning

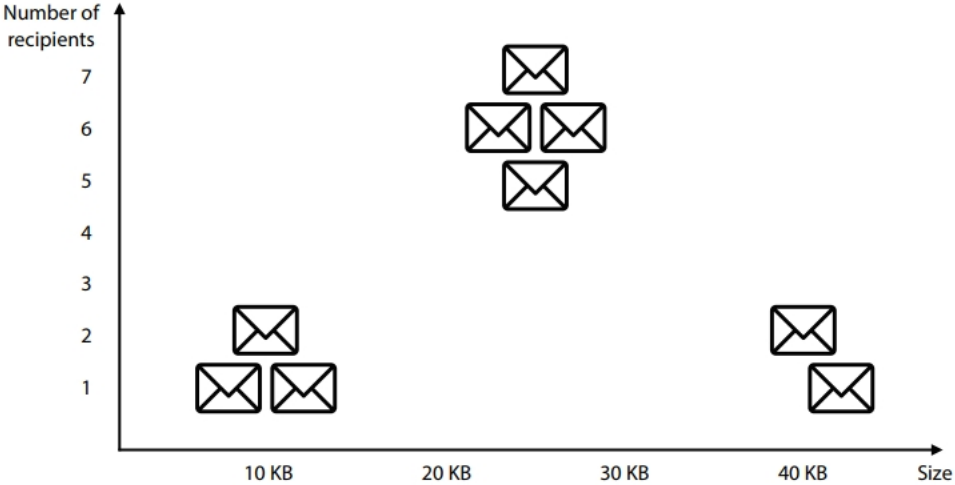
- **Clustering** algorithms: The algorithms that group data into clusters based on similarity
- **Dimensionality reduction** algorithms: The algorithms that simplify our data and faithfully describe it with fewer features
- **Generative** algorithms: The algorithms that can generate new data points that resemble the existing data

Example: clustering emails

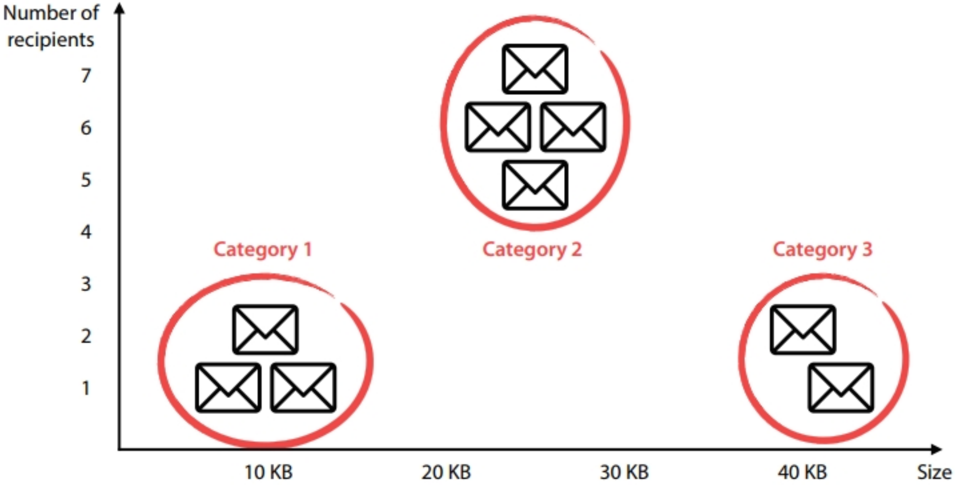


Email	Size	Recipients
1	8	1
2	12	1
3	43	1
4	10	2
5	40	2
6	25	5
7	23	6
8	28	6
9	26	7

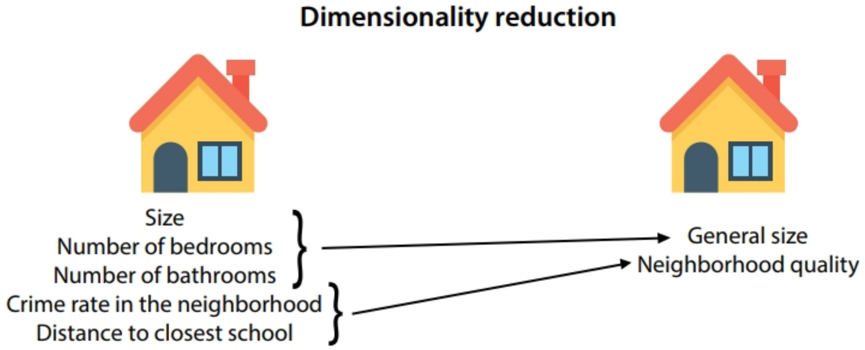
Example: clustering emails



Example: clustering emails



Example: dimensionality reduction



- The fancy word for the number of columns in a dataset is dimension
    - All we're doing is reducing the number of columns

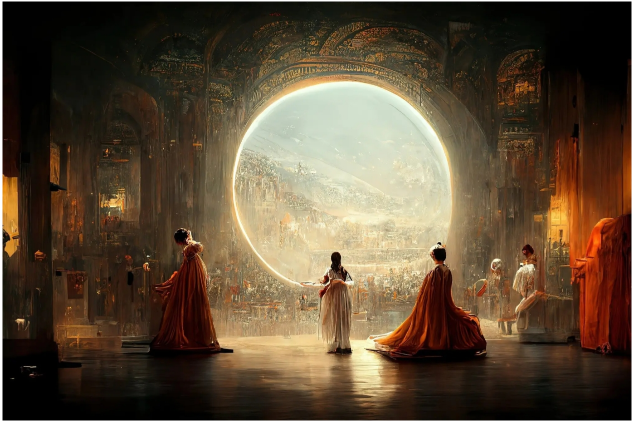
Semi-supervised learning

- Semi-supervised machine learning is implemented for datasets where the output is given for only a few instances of the inputs
    - Step 1: supervised learning on labeled data
    - Step 2: unsupervised learning on the remaining unlabeled data
- The advantage is that a lot of effort and computational cost are saved because collecting and labeling large datasets can be very expensive
- The patterns and the similarities among the data instances are discovered, which brings more insight into the dataset structure

Generative machine learning

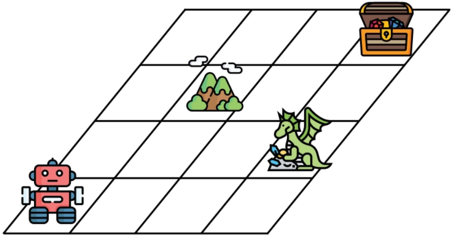
- If you have seen ultra-realistic faces, images, or videos created by computers, then you have seen generative machine learning in action
- given a dataset, can output new data points that look like samples from that original dataset
  - Generative adversarial networks (GANs)

- Jason Allen’s A.I.-generated work, "Théâtre D’opéra Spatial," took first place in the digital category at the Colorado State Fair



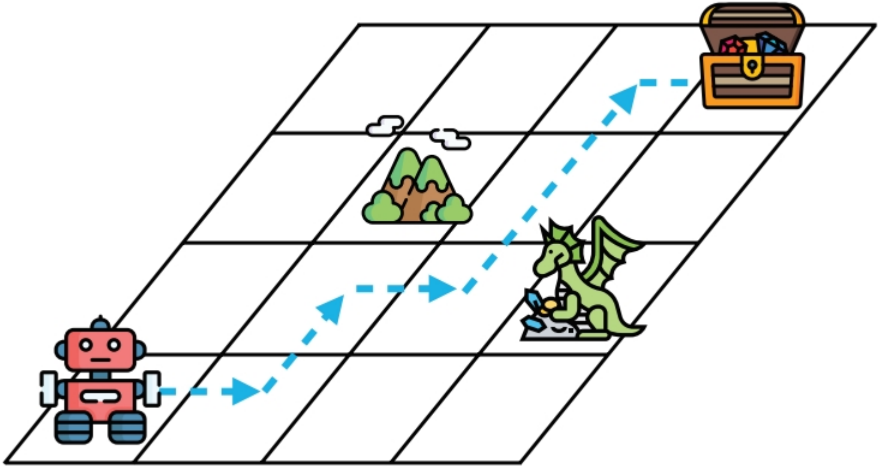
Reinforcement learning

- Reinforcement learning is a different type of machine learning in which **no data** is given, and we must get the computer to perform a task. Instead of data
  - The model receives an environment and an agent who is supposed to navigate in this environment
  - The agent has a goal or a set of goals
  - The environment has rewards and punishments that guide the agent to make the right decisions to reach its goal



Grid world

- we see a grid world with a robot at the bottom-left corner. That is our agent. The goal is to get to the treasure chest in the top right of the grid. In the grid, we can also see a mountain, which means we cannot go through that square, because the robot cannot climb mountains. We also see a dragon, which will attack the robot, should the robot dare to land in its square, which means that part of our goal is to not land over there
- To give the robot information about how to proceed, we keep track of a score
  - The score starts at zero. If the robot gets to the treasure chest, then we gain 100 points. If the robot reaches the dragon, we lose 50 points. And to make sure our robot moves quickly, we can say that for every step the robot makes, we lose 1 point, because the robot loses energy as it walks.



- Reinforcement learning has numerous cutting-edge applications, including the following:
  - Games: recent advances in teaching computers how to win at games, such as Go or chess, use reinforcement learning. Also, agents have been taught to win at Atari games such as Breakout or Super Mario.
  - Robotics: reinforcement learning is used extensively to help robots carry out tasks such as picking up boxes, cleaning a room, or even dancing!
  - Self-driving cars: reinforcement learning techniques are used to help the car carry out many tasks such as path planning or behaving in particular environments.

Transfer task

- Task 1: For each of the following scenarios, state if it is an example of supervised or unsupervised learning. Explain your answers. In cases of ambiguity, pick one and explain why you picked it

1. A recommendation system on a social network that recommends potential friends to a user
2. A system in a news site that divides the news into topics
3. The Google autocomplete feature for sentences
4. A recommendation system on an online retailer that recommends to users what to buy based on their past purchasing history
5. A system in a credit card company that captures fraudulent transactions

Transfer task

- Task 2: For each of the following applications of machine learning, would you use regression or classification to solve it? Explain your answers. In cases of ambiguity, pick one and explain why you picked it

1. An online store predicting how much money a user will spend on their site
2. A voice assistant decoding voice and turning it into text
3. Selling or buying stock from a particular company
4. YouTube recommending a video to a user

References

- https://end-to-end-machine-learning.teachable.com/courses
- https://www.udacity.com/course/intro-to-machine-learning-with-tensorflow-nanodegree--nd230
- https://www.coursera.org/specializations/machine-learning-introduction

Questions?