

|  |
| --- |
| Now let us start with the first session of the module: Introduction to Machine Learning (ML)  You will learn about the basic concepts of artificial intelligence. This session will address the following questions: How does machine learning work? What are the main steps? What can machine learning do? What are advantages and disadvantages of machine learning?  All of you have a background in data science, so normally this module should just be a refresher, and some may even find it too easy and somewhat redundant. However, considering the various levels of experiences within the group, it is important that we introduce the basic principles of machine learning again to everyone.  This is why this module is designed to give a first introduction into the topic. The good news is that because the module is online and self-paced, you can advance at your own speed according to your background and prior knowledge.  Ready?  Let's go! |
| **Basic concepts of AI and ML**  Artificial Intelligence is defined as the branch of computer science that is concerned with the automation of intelligence behaviour. It is a field of study that seeks to explain and emulate intelligent behavior in terms of computational processes. (Lugar et al ,1993) |
| Artificial Intelligence (AI) is the simulation of human intelligence by machines, it is the ability to:   * Perceive the inter-relationship of facts * Learn and understand from experience * Acquire and retain knowledge * Respond quickly and successfully to a new situation |
| The central principles of AI include:   * Reasoning, knowledge, planning, learning and communication.   **AI is about bringing together computers and humans in ways that enhance human life** |
| **What is Learning?**   * Learning is a mechanism by which an agent improves its performance by observing the world. * Learning is the use of experience to gain expertise. * If the agent that is learning is a computer, we call it machine learning |
| Some examples of learning - classifying spam email:   * Let’s say you decided to write your own spam email detector program. * When your program gets a new email, it compares it to all emails that have been labelled in the past and if it matches one of the spam emails seen in the past it is classified it as spam if it does not match any past spam email it classifies it as not-spam. * Whenever the program fails to detect a spam email because it does not match one of the past emails you inform it that this email was spam. The program then filters out future emails that match the newly added spam email.   Source: Understanding Machine Learning: From Theory to Algorithms: <https://dl.acm.org/doi/book/10.5555/2621980> |
| **The learning by memorization has one big disadvantage:** It is not very useful when a radically different situation is encountered such as when you see a spam email you have never seen before.  The ability to use past experience to make inference about novel experiences is known as **generalization.**  A good learning system must have the ability to generalize to broader population than it has seen.  Source: Understanding Machine Learning: From Theory to Algorithms: <https://dl.acm.org/doi/book/10.5555/2621980> |
| **Why and when do we need Machine Learning?**  **1. Tasks that are too hard to program**   * Image recognition: this can be done easily by humans and animals, but it is hard to write as a rule-based program * Driving vehicles * Speech Recognition   How image recognition works.  Source: <https://www.fritz.ai/image-recognition/> |
| **2. Tasks that can’t be handled by humans:**   * Analyzing genomic data of billions of humans * Determining which ads/products to show to which users when you have hundreds of millions to billions of users.   DNA Double Helix with Data | The DNA double helix rests on a… | Flickr  <https://www.flickr.com/photos/genomegov/27862777945> |
| **3. Tasks that require to continuous adaptability to their input:**   * Detecting spam emails: Spammers change their techniques everyday writing a program to cope with this requires a team on standby. * Consumer interests change over time; thus, using machine learning to recommend products to the customer is necessary.   + A person that loved action movies last year may now be interested in watching comedy.   + A user that loved buying video games online at age 16 may want to buy books at age 30. |
| **What can Machine Learning do?**  Here are some examples:   1. **Recognize patterns:**  * Facial identities or facial expressions * Handwritten or spoken words * Medical images  1. **Generate patterns**  * Generating images or motion |
| 1. **Recognizing anomalies**  * Banks check loan applications before making a decision. If the system detects that some of the documents are fraudulent, for example, that your tax number doesn’t exist in the system, it will notify the bank employer. * Help doctors with diagnosis detecting unusual patterns in MRI and test results  1. **Predictions**  * Future stock prices or currency exchange rates |
| **Machine Learning: Advantages/Disadvantages**  **Advantages of Machine Learning**   * Efficient Management of Data: Machine Learning is beneficial because it primarily facilitates efficient data management in computers. By managing vast amounts of data and closely observing it, computers make use of ML and interpret data. This management is very helpful in the field of technology and has made the process of data management rapid. With the help of numerous methods of ML, computers identify relevant patterns that help humans to carry out otherwise tedious tasks. * Valuable Use in Technology: With so many applications of Machine Learning, the world of technology revolves around machine learning. Its valuable use in the field is highly commendable. From Natural Language Processing to Artificial Intelligence, machine learning is omnipresent as it empowers technology to leap forward. With so many benefits of machine learning, the concept of artificial intelligence is advanced more than ever. |
| * Automated Operations: Machine Learning is defined as the process wherein machines such as computers are made to learn human skills without needing human assistance. This merit of machine learning- its ability to carry out automatic operation- has made it a reputed and regarded concept in the realm of technology. Not only does it work independently, but it also keeps humans away from the process, thus saving time and energy from most of the processes involved in the concept. |
| **Disadvantages of Machine Learning**   * Manual Algorithm Selection: While much of the machine learning process is automated and works with the help of computers, the process of algorithm selection in machine learning is still manual and requires human assistance. Is perhaps one of the biggest limitations of machine learning. This means that while computers design the algorithms, humans designate the algorithms that are supposed to be included in the interpretation of data. This can be a tedious task as it requires humans to run data through various algorithms and identify the one that works the best. * Delayed Resolution of errors: While management of data is a forte of machine learning, the issue of errors remains a highlighting drawback. Furthermore, the delayed resolution of errors is what keeps the concept of machine learning from becoming perfect! This means that even though the errors occur in the process of Machine Learning, the resolution of errors is a much-delayed process due to the strong reliance of machines on algorithms. This particular trait can be a major disadvantage as it can hamper the process of data management. |
| * Requirement of Extensive Resources: Lastly, the requirement of extensive resources in machine learning is a concern for many. Why? Because the interpretation of data can be a time-taking process that requires many other equipment's attached to your computer. * Some ML models are black box models meaning that the models can get good results, but no one is able to explain why the models get good results. Neural networks are the most well-known black box ML models. |
| **Some Types of Machine Learning**  Let’s say the company *example.com* hires you to create a spam filtering system.  Case 1: The company provided 10000 emails that were labelled as spam/not-spam by email analysis experts and asked you to create a system that generalizes beyond the 10000 emails.   * When you use features of your data and labels to teach a machine learning model it is known as **supervised learning.**   Case 2: The company provided 10000 emails that were not labelled and asks you to identify unusual emails.   * When you use features of your data **only** to detect patterns in your data it is known as **unsupervised learning**   Case 3: The company provided 10000 emails but only 3000 of those were labelled as spam/not-spam. But the company asked you to use all 10000 emails to train a model by devising a way to label the unlabeled 7000 examples.   * When you use labelled and unlabeled examples together for training a machine learning model it is called **semi-supervised learning**. Semi-supervised learning requires some approach that generates labels for unlabeled examples.   Case 4: Now, let’s say the company asked you to create an email summarization system that returns a summary of the email with gist of the email unchanged but with as few words as possible. The company gave you the 10000 emails it has.   * You decide to train the system by having it output a summary of the examples you gave it and penalizing it for producing a summary that contains more than 50 words. Every word more than 50 words constitutes a penalty. To make sure you have a sound text output you also penalize your model whenever it produces a grammatically incorrect output. * When you use rewards and penalties to train an ML model to behave in a desired manner it is called reinforcement learning. In Reinforcement learning, the model is trained by interacting with the environment/data producing an output and receiving a reward or penalty for the output it received.   Main families of ML methodsSource: <https://labelyourdata.com/articles/machine-learning-and-training-data> |
| **A simple example: Supervised Learning**  Imagine you have just arrived in Zodirian island to harvest and export papayas to Rwanda.  You want to predict whether a papaya you see is sweet or not. Based on your experience with oranges you decide to determine papaya sweetness based on the colour of the papaya and the softness of the papaya. We will refer to the colour and softness as the selected features. You want to learn a rule that allows you to predict the sweetness of a papaya given the papaya’s colour and softness. To learn something, we need a learner. In this case, learner is an **algorithm that takes training data to output a prediction rule (classifier)**.    Source: Understanding Machine Learning: From Theory to Algorithms: <https://dl.acm.org/doi/book/10.5555/2621980>  Image source: <https://freesvg.org/1546870776> |
| **Inputs of a Learner: Training Data**   * The input of a learner is a training data. * **Domain Set**: The set of all objects that we would like to make predictions for   + In the papaya classification example, the domain set is the set of all papayas in the island.   + Sometimes objects in domain set maybe represented by their features (colour and softness) in our case. * **Label Set**: The set of all applicable labels to the objects in the domain set   + In the papaya classification example, the labels are sweet/not-sweet. |
| * **Training data**: is a set of labelled domain points.   **{**  **(papaya1\_color, papaya1\_softness, sweet),**  **(papaya2\_color, papaya2\_softness, sweet),**  **(papaya3\_color, papaya3\_softness, not-sweet),**  **…,**  **(papaya50\_color, papaya50\_softness, not-sweet),**  **…}**  **The training data above is a set of (feature1, feature2, label). In an unsupervised ML task, you will not have a label in the training data.** |
| **Distribution of Labels**  To understand the content of the next few concepts, we need to understand the concept of distribution.  Distribution is a function that tells how likely a certain characteristic is in some group or population. For example, the following distribution indicates the likelihood of scores in some exam administered at some school. |
| The above figure shows based on data collected over many years how likely a student is to get a score of 1, 2, 3, ...,10, out of 10. The figure shows that there is less than 5% chance that a student scores a 1 out of 10 and greater than 25% chance that a student scores 5.  So, the distribution of student scores tells us the likelihood of scores.  Every observable phenomenon has a distribution of the events it contains. For instance, the distribution of number of lightning strikes per hour during a rainstorm, the distribution of number of car accidents in a day in Kigali. Let’s take the second example, the distribution of number of car accidents in Kigali tells us that how likely it is to have no accidents, one accident, two accidents, … in the city of Kigali. |
| Every observable quantity has a distribution over other quantity. For example, the distribution of weight over people that are taller than 170 cm. If we know the distribution of weight for people taller than 170 cm, we can say many things about the group like 40% of people taller than 170cm weigh more than 70kg, only 10% of people taller than 170cm weigh less than 50kg, and so on… |
| **Training Data Generation**   * We have seen that any observable phenomenon or quantity has a distribution. Let’s see how we can use this with our example. * Thus, there is a distribution of sweetness of papayas over all attributes of papaya. We know this because sweetness is an observable attribute of papayas.   + What does this mean? Sweet papayas occur in certain frequencies for a given papaya attribute. Look at the following simple statements about the distribution of sweetness.     - 30% of yellow papayas are sweet     - 60% of hard papayas are not sweet.     - 45% of mushy and green papayas are sweet.   + All of the above example tells us the likelihood of being sweet/not-sweet given attributes (color and/or softness) of our data |
| * No one knows what the actual distribution of sweetness over the papaya attributes is * If we knew the distribution, we wouldn’t need machine learning.   + We would simply use the distribution to determine the best classifier. * Since we have no idea what the distribution of the labels over our domain set is, we use the next best thing: samples collected from our domain set.   + In the papaya classification example, we go to the Zodirian island and collect papayas.   Simply collecting papayas is not enough to create a training set. Why? |
| **Training Data Generation: Labelling Function**   * A training set needs to have points in the domain set and their respective labels (in supervised ML task) * To get labels, we need a labelling function. * We use the term labelling function to encapsulate any human or system that can provide labels for a given ML task.   + In many instances, the labels are provided by humans. In this case, the human can be considered a labelling function that takes the features as input and give you a label for the sample as output.   + In some instances, labels can be computed in an automated manner. Let’s say you want to train a model that predicts crop field area from satellite images. To obtain labels for your training dataset you may simply use an API to query the local district office’s database for the size of the crop fields at the coordinates of interest. Once the model is trained on area where field area is available the ML model can be applied in an area where the local district office doesn’t have such data. |
| * + In this case, the code you use to query the database to obtain labels can be considered a labelling function. * Where do we get the labels from?   + A labelling function is a function that takes **features** of your data to output the reference label or the ground truth.   + In the case of papaya sweetness, it can be any function (or human) that can use a papaya in some way to **correctly** tell you whether a papaya is sweet or not. |
| * + Remember, selected features are the features we chose to use to do our ML task.   + If we know a labelling function that can take the **selected features** and label all instances in the domain set correctly, we don’t need machine learning.     - We already have a solution     - In general, the labelling function must use features other than the set of **selected features**.     - There are some exceptions where the features for prediction/training and labelling are the same.   + To get the labels we use a labelling function that can use features other than the selected features.     - In the papaya classification example, we use a human as labelling function.       * The human tastes the papayas to label the papayas we collected as sweet/not sweet.       * Notice how the feature used by the “labelling function” is different from the features we would like to use with our classifier (taste for labelling vs. colour and softness for classification) * This section discussed the general approach to generating training data the following section will further discuss many practical aspects of training data generation such as sources of training data.   Source: Understanding Machine Learning: From Theory to Algorithms: https://dl.acm.org/doi/book/10.5555/2621980 |
| **Story so far**   * We have a domain set over which we would like to make predictions. * We have a label set which contains the set of all possible labels. * A prediction rule/classifier that takes the selected features to output a predicted label. * A labelling function that takes some features not in the selected features to output the true labels |
| **How can we determine if the prediction rule/classifier is good?** |
| **Measure of Success**   * The **error of a classifier** (prediction rule) is the probability that the **classifier** outputs the **wrong label** for a random data point sampled from the domain set.   + This error is also known as the generalization error, or true error. * What does it mean to output a wrong label?   + A classifier is said to output a wrong label if the label it produced does not match the label produced by the labelling function for the same input. * **The goal of machine learning is to minimize generalization error for the task at hand.** * A model with good generalization is a model with the smallest generalization error.   Source: Understanding Machine Learning: From Theory to Algorithms: https://dl.acm.org/doi/book/10.5555/2621980 |
| **Learner Input and Output: Recapped**   * A learner is an **algorithm** that takes a training set generated from an **unknown distribution** as input and outputs a classifier. * The classifier takes features of an element in the domain set. * The classifier outputs labels * The learner does not have access to the distribution of labels or the labelling function.   + Without knowing these two it is not possible to determine the generalization error. Why?   + So how will the learner evaluate the quality of the classifier it outputs? |
| **Training Error**   * Since a learner can’t assess the generalization error of a classifier it uses the training error to assess the quality of the learned classifier **during training** * Training error is defined as the fraction of training examples that a learned classifier produces wrong output on.   + If the labels produced by a classifier on 20 of 100 training examples is wrong the training error will be 0.2 * Training error is useful as it does not require access to the distribution. * However, training error can be misleading and lead to overfitting.   Source: Understanding Machine Learning: From Theory to Algorithms: https://dl.acm.org/doi/book/10.5555/2621980 |
| **Overfitting**   * Look at the image of your favorite island at the bottom. * Different papayas with different features are shown on the image. * The labels assigned by the labelling function are shown on the papayas.   + Not sweet ❌   + Sweet **✔️**     Image source: <https://freesvg.org/1546870776>  Source: Understanding Machine Learning: From Theory to Algorithms: <https://dl.acm.org/doi/book/10.5555/2621980> |
| * Now assume you only collected data from the region encircled by the rectangle. * This data can be considered noisy because it misrepresents aspects of the true distribution. * A good ML model is expected to learn most of the things that are generally true and disregard noisy aspects of the data. * From the data shown in this region some bad learner might output a classifier that predicts sweet if the papaya is yellow and not-sweet otherwise. This can happen because the collected training dataset that contains only yellow sweet papayas. * A good learner is expected to base its classification on as many of the selected features as possible and give rules that are generally true. * A bad learner will simply give rules that are true for the training dataset (maybe even perfectly true) that are not generally true. * Clearly the rule the classifier learned is not true outside the rectangle and the classifier will not generalize well.     Image source: https://freesvg.org/1546870776 |
| * Overfitting is when the classifier learns so much about the training data that it learns facts about the training data that are not generally true. * In formal terms, overfitting is when a classifier learns the noise in the training data. * A classifier that overfits the training data is not good at generalizing to unseen examples. * The primary goal of learning is generalization; thus, a classifier that overfits the training data is not a good classifier. * Overfitting can result from poorly collected training set or using a classifier that is too big.   Source: Shai Shalev-Shwartz and Shai Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press 40 W. 20 St. New York, NY, United States, or <https://dl.acm.org/doi/book/10.5555/2621980> |
| **Underfitting**  An ML model can also underfit a dataset. A model is said to be **underfitted** when it is not able to capture the underlying trend of the data. It means the model shows poor performance even with the training dataset. In most cases, underfitting issues occur when the model is not perfectly suitable for the problem that we are trying to solve. The graphic below shows how models can overfit or underfit a given dataset.    Source : <https://labelyourdata.com/articles/machine-learning-and-training-data> |
| **Measuring Generalization Error**   * To determine the actual usefulness of a model one must measure the generalization error. * It is impossible to measure the actual generalization error of a classifier without knowing the distribution of your labels over your domain set.   + But no one knows the distribution. * It is possible to approximate the generalization error by measuring the error on randomly sampled data that was not part of the training data |
| * A randomly sampled data that is used for measuring the generalization error of a classifier is known as test data. * The error of a classifier on unseen data is known as test error. * In practice, the goal of machine learning is minimizing test error. * Determining the actual generalization error requires access to the distribution which no one knows |
| **Recap**  So far what we have studied are the overarching goals of Machine Learning.  We defined learning as “gaining expertise from experience.” Everything we have discussed so far is related with this statement.  **Learner**: Uses experience to produce an expert  **Prediction rule (classifier)**: Is the expert  **Training data**: Is the experience that teaches the expert  **Test error:** Determines the level of expertise of the expert. The lower the test error the better the expert is.  Putting this in mind we will now look at the key steps in Machine Learning. Then, the following section will also go into further detail on many practical aspects of machine learning. |
| **Key Steps in Machine Learning**   1. **Identifying the domain set, the label set, and the feature selection**  * Identifying the domain set and label allows you to determine the input and output of the classifier you will train * The features you will use also determine the structure of your classifier.   + You determine the features to use for an ML problem using domain knowledge, and experience with similar problems |
| 1. **Data collection**  * Data collection is the process of obtaining training data that will be used as input to the learning algorithm and the test data that will be used to measure the generalization error. * When collecting data, we must always ensure that the collected data has similar characteristics to the domain set. This is called representativeness.   + If you are training a model to determine credit eligibility of a Rwandan but you only use people from Kigali as your training data, your model might fail frequently when dealing with non-Kigalians.   + Similarly, if your data collected from city of Butare contains only credit eligible individuals but your data from Musanze contains credit ineligible individuals your classifier might learn to simply reject everyone from Musanze and accept everyone from Butare. This is known as a bias. * You must always use up-to-date and correct data to prevent wrong outcomes or predictions. * Good data is relevant, contains very few missing and repeated values, and has a good representation of the various subcategories/classes present. |
| 1. **Preparing the Data**  * After collecting data, they have to be prepared. Here are some of the steps that can be used to prepare data that will be discussed in greater detail in 4.2.   + Data is put together and randomized. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.   + Data cleaning   + Data visualization   + Splitting the cleaned data into two sets   + Balancing Label frequency   + Standardization   + One hot encoding   + Augmentation   Source: https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding |
| 1. **Choosing a Model**  * A machine learning model determines the output to be obtained after running a machine learning algorithm on the collected data. It is important to choose a model which is relevant to the task at hand. Over the years, scientists and engineers developed various models suited for different tasks like speech recognition, image recognition, prediction, etc. Apart from this, it is better to check if the selected model is suited for numerical or categorical data and choose accordingly.  1. **Training the Model**  * Training is the most important step in machine learning. In training, you pass the prepared data to your machine learning model to find patterns and make predictions. The learning algorithm (learner) modifies the classifier (prediction rule) based on the training data. * Generally, the objective of the learning algorithm in training is to minimize training errors. It is important to take great care to avoid overfitting to the training data. |
| 1. **Hyperparameter Tuning**  * Once you have created /selected the model, see if its accuracy can be improved in any way. This is done by tuning the hyperparameters present in the model. The following section will discuss what hyperparameters are in detail. |
| 1. **Evaluating the Model**  * After training the model, it is necessary to check how it’s performing. This is done by testing the performance of the model on previously unseen data.. If testing was done on the same data which is used for training, you will not get an accurate measure, as the model is already used to the data, and finds the same patterns in it, as it previously did. This will give disproportionately high accuracy.  1. **Deployment**  * In the end, you can use your model in a production environment to make predictions on unseen data accurately. * Deploying a model may require further setup depending on the computing environment that the classifier will be deployed to. Deploying to a distributed computing environment with hundreds of nodes is very different from deploying to a single server. Similarly, deploying to a single server is very different from deploying to microcontroller unit. |

**Exercise materials and tasks**

**Quiz questions**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Please answer the following questions to test your understanding of artificial intelligence:   1. Machine Learning is the use of experience to gain expertise. The level of expertise of a learned model is evaluated by:    1. The fraction of examples seen in the past the ML model can exactly remember and use to compare future examples.    2. **The probability that the ML model will make a mistake on a randomly chosen unseen example.**    3. The probability that the ML model will make a mistake on a randomly chosen member of validation set.    4. The fraction of training examples that a learned classifier produces wrong output on 2. For which of the following tasks would it make sense to use machine learning?    1. To predict the output of a chemical reaction whose input elements and output compounds are perfectly known.    2. **To do a task such as face recognition that humans can do easily.**    3. To determine a prediction rule for a task where the distribution of the labels over the selected features is perfectly known.    4. **When there is a classification task that can be achieved by using certain features but for some reason, we would like different set of features.** 3. Which of the following cannot cause overfitting?    1. **Using a model so simple that it can’t perform well on the training data**    2. Training a large and complex model to the point that it learns everything about the data.    3. Using a poorly sampled data that is not representative of the domain set. 4. You learned that ML has several advantages and disadvantages. For the given statements, sort them into advantages and disadvantages: *efficient management of data, black box models, interpretation of data can be a time-taking process, valuable use in technology, automatic operations without human assistance, algorithm selection is manual, delayed resolution of errors*  |  |  | | --- | --- | | Advantages | Disadvantages | |  |  |   Correct answer:   |  |  | | --- | --- | | Advantages | Disadvantages | | efficient management of data, valuable use in technology, automatic operations without human assistance | **black box models, interpretation of data can be a time-taking process, algorithm selection is manual, delayed resolution of errors** |  1. Which of the following is true about the learning algorithm and classifier? 2. The learning algorithm (learner) and the classifier are the same thing. 3. Once a classifier is trained the learning algorithm and the classifier are used to classify data 4. **The learning algorithm is an algorithm that modifies a given classifier using the training data.** 5. **A classifier uses the selected features to output a predicted class.** |