**Machine Learning Assignment 3**

**1. Explain the term machine learning, and how does it work? Explain two machine learning applications in the business world. What are some of the ethical concerns that machine learning applications could raise?**

**Machine Learning**: Machine learning is the process of making systems that learn and improve by themselves, by being specifically programmed. The ultimate goal of machine learning is to design [algorithms](https://www.simplilearn.com/10-algorithms-machine-learning-engineers-need-to-know-article) that automatically help a system gather data and use that data to learn more. Systems are expected to look for patterns in the data collected and use them to make vital decisions for themselves. In general, machine learning is getting systems to think and act like humans, show human-like intelligence, and give them a brain. In the real world, there are existing machine learning models capable of tasks like :

* Separating spam from actual emails, as seen in Gmail
* Correcting grammar and spelling mistakes, as seen in autocorrect

**Machine Learning Applications**: Thanks to machine learning, the world has also seen design systems capable of exhibiting uncanny human-like thinking, which performs tasks like:

* Object and image recognition
* Detecting fake news
* Understanding written or spoken words
* Bots on websites that interact with humans, like humans
* Self-driven cars

**Machine Learning Steps :** The task of imparting intelligence to machines seems daunting and impossible. But it is actually really easy. It can be broken down into 7 major steps:

1. **Collecting Data:** As we know, machines initially learn from the [data](https://www.simplilearn.com/what-is-data-article) that we give them. It is of the utmost importance to collect reliable data from reliable source so that our machine learning model can find the correct patterns. The quality of the data that we feed to the machine will determine how accurate our model is. If we have incorrect or outdated data, we will have wrong outcomes or predictions which are not relevant.
2. **Preparing the Data**: After we have your data, we have to prepare it. We can do this by :

* Putting together all the data we have and randomizing it. This helps make sure that data is evenly distributed, and the ordering does not affect the learning process.
* Cleaning the data to remove unwanted data, missing values, rows, and columns, duplicate values, data type conversion, etc. We might even have to restructure the dataset and change the rows and columns or index of rows and columns.
* [Visualize the data](https://www.simplilearn.com/data-visualization-article) to understand how it is structured and understand the relationship between various variables and classes present.
* Splitting the cleaned data into two sets - a training set and a testing set. The training set is the set our model learns from. A testing set is used to check the accuracy of our model after training.

### 3. Choosing a Model: A machine learning model determines the output you get 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, you also have to see if our model is suited for numerical or categorical data and choose accordingly.

### 4. Training the Model: Training is the most important step in machine learning. In training, we pass the prepared data to our machine learning model to find patterns and make predictions. It results in the model learning from the data so that it can accomplish the task set. Over time, with training, the model gets better at predicting.

### 5. Evaluating the Model: After training our model, we have to check to see how it’s performing. This is done by testing the performance of the model on previously unseen data. The unseen data used is the testing set that we split our data into earlier. If testing was done on the same data which is used for training, we 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 us disproportionately high accuracy.

When used on testing data, you get an accurate measure of how your model will perform and its speed.

### 6. Parameter Tuning: Once you have created and evaluated your model, see if its accuracy can be improved in any way. This is done by tuning the parameters present in your model. Parameters are the variables in the model that the programmer generally decides. At a particular value of your parameter, the accuracy will be the maximum. Parameter tuning refers to finding these values.

### 7. Making Predictions: In the end, you can use your model on unseen data to make predictions accurately.

**2. Describe the process of human learning:**

**i. Under the supervision of experts**

**ii. With the assistance of experts in an indirect manner**

**iii. Self-education**

**3. Provide a few examples of various types of machine learning.**

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: [supervised](https://www.techtarget.com/searchenterpriseai/definition/supervised-learning) learning, [unsupervised](https://www.techtarget.com/whatis/definition/unsupervised-learning) learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict.

* Supervised learning: In this type of machine learning, [data scientists](https://www.techtarget.com/searchenterpriseai/definition/data-scientist) supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.
* Unsupervised learning: This type of machine learning involves algorithms that train on unlabeled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.
* Semi-supervised learning: This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled [training data](https://www.techtarget.com/searchenterpriseai/feature/Using-small-data-sets-for-machine-learning-models-sees-growth), but the model is free to explore the data on its own and develop its own understanding of the data set.
* Reinforcement learning: Data scientists typically use [reinforcement learning](https://www.techtarget.com/searchenterpriseai/definition/reinforcement-learning) to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.

1. **Examine the various forms of machine learning.**

**Supervised machine learning** requires the [data scientist](https://www.techtarget.com/searchbusinessanalytics/feature/Key-differences-of-a-data-scientist-vs-data-engineer) to train the algorithm with both labeled inputs and desired outputs. Supervised learning algorithms are good for the following tasks:

* Binary classification: Dividing data into two categories.
* Multi-class classification: Choosing between more than two types of answers.
* Regression modeling: Predicting continuous values.
* Ensembling: Combining the predictions of multiple machine learning models to produce an accurate prediction.

**Unsupervised machine learning** algorithms do not require data to be labeled. They sift through unlabeled data to look for patterns that can be used to group data points into subsets. Most types of deep learning, including [neural networks](https://www.techtarget.com/searchenterpriseai/definition/neural-network), are unsupervised algorithms. Unsupervised learning algorithms are good for the following tasks:

* Clustering: Splitting the dataset into groups based on similarity.
* Anomaly detection: Identifying unusual data points in a data set.
* Association mining: Identifying sets of items in a data set that frequently occur together.
* Dimensionality reduction: Reducing the number of variables in a data set.

**Semi-supervised learning** works by data scientists feeding a small amount of [labeled training data](https://www.techtarget.com/searchenterpriseai/feature/Labeled-data-brings-machine-learning-applications-to-life) to an algorithm. From this, the algorithm learns the dimensions of the data set, which it can then apply to new, unlabeled data. The performance of algorithms typically improves when they train on labeled data sets. But labeling data can be time consuming and expensive. Semi-supervised learning strikes a middle ground between the performance of supervised learning and the efficiency of unsupervised learning. Some areas where semi-supervised learning is used include:

* **Machine translation:** Teaching algorithms to translate language based on less than a full dictionary of words.
* **Fraud detection:** Identifying cases of fraud when you only have a few positive examples.
* **Labelling data:** Algorithms trained on small data sets can learn to [apply data labels](https://www.techtarget.com/whatis/definition/data-labeling) to larger sets automatically.

**Reinforcement learning** works by [programming an algorithm](https://www.techtarget.com/searchenterpriseai/feature/5-types-of-machine-learning-algorithms-you-should-know) with a distinct goal and a prescribed set of rules for accomplishing that goal. Data scientists also program the algorithm to seek positive rewards -- which it receives when it performs an action that is beneficial toward the ultimate goal -- and avoid punishments -- which it receives when it performs an action that gets it farther away from its ultimate goal. Reinforcement learning is often used in areas such as:

* **Robotics:** Robots can learn to perform tasks the physical world using this technique.
* **Video gameplay:** Reinforcement learning has been used to teach bots to play a number of video games.
* **Resource management:** Given finite resources and a defined goal, reinforcement learning can help enterprises plan out how to allocate resources.

**4.Can you explain what a well-posed learning problem is? Explain the main characteristics that must be present to identify a learning problem properly.**

Well Posed Learning Problem – A computer program is said to learn from experience E in context to some task T and some performance measure P, if its performance on T, as was measured by P, upgrades with experience E.

Any problem can be segregated as well-posed learning problem if it has three traits –

Task

Performance Measure

Experience

Certain examples that efficiently defines the well-posed learning problem are –

1. To better filter emails as spam or not

Task – Classifying emails as spam or not

Performance Measure – The fraction of emails accurately classified as spam or not spam

Experience – Observing you label emails as spam or not spam

2. A checkers learning problem

Task – Playing checkers game

Performance Measure – percent of games won against opposer

Experience – playing implementation games against itself

3. Handwriting Recognition Problem

Task – Acknowledging handwritten words within portrayal

Performance Measure – percent of words accurately classified

Experience – a directory of handwritten words with given classifications

4. A Robot Driving Problem

Task – driving on public four-lane highways using sight scanners

Performance Measure – average distance progressed before a fallacy

Experience – order of images and steering instructions noted down while observing a human driver

5. Fruit Prediction Problem

Task – forecasting different fruits for recognition

Performance Measure – able to predict maximum variety of fruits

Experience – training machine with the largest datasets of fruits images

6. Face Recognition Problem

Task – predicting different types of faces

Performance Measure – able to predict maximum types of faces

Experience – training machine with maximum amount of datasets of different

face images

7. Automatic Translation of documents

Task – translating one type of language used in a document to other language

Performance Measure – able to convert one language to other efficiently

Experience – training machine with a large dataset of different types of languages.

**6. Is machine learning capable of solving all problems? Give a detailed explanation of your answer.**

**7. What are the various methods and technologies for solving machine learning problems? Any two of them should be defined in detail.**

## List of Common Machine Learning Algorithms: Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1. **Linear Regression**
2. **Logistic Regression**
3. **Decision Tree**
4. **SVM**
5. **Naive Bayes**
6. **kNN**
7. **K-Means**
8. **Random Forest**
9. **Dimensionality Reduction Algorithms**
10. **Gradient Boosting algorithms**
    1. **GBM**
    2. **XGBoost**
    3. **LightGBM**
    4. **CatBoost**

**8. Can you explain the various forms of supervised learning? Explain each one with an example application.**

**Supervised Learning :**Supervised learning is when the model is getting trained on a labelled dataset. A **labelled** dataset is one that has both input and output parameters.

**Types of Supervised Learning:**

1. **Classification:**It is a Supervised Learning task where output is having defined labels(discrete value). For example in above Figure A, Output – Purchased has defined labels i.e. 0 or 1; 1 means the customer will purchase and 0 means that customer won’t purchase. The goal here is to predict discrete values belonging to a particular class and evaluate them on the basis of accuracy.   
   It can be either binary or multi-class classification. In **binary** classification, the model predicts either 0 or 1; yes or no but in the case of **multi-class** classification, the model predicts more than one class.

**Example:** Gmail classifies mails in more than one class like social, promotions, updates, forums.

1. **Regression:**It is a Supervised Learning task where output is having continuous value.   
   Example in above Figure B, Output – Wind Speed is not having any discrete value but is continuous in the particular range. The goal here is to predict a value as much closer to the actual output value as our model can and then evaluation is done by calculating the error value. The smaller the error the greater the accuracy of our regression model.

**Example of Supervised Learning Algorithms:**

* Linear Regression
* Nearest Neighbor
* Gaussian Naive Bayes
* Decision Trees
* Support Vector Machine (SVM)
* Random Forest

**9. What is the difference between supervised and unsupervised learning? With a sample application in each region, explain the differences.**

The main differences of supervised vs unsupervised learning include:

* The need for labelled data in supervised machine learning.
* The problem the model is deployed to solve. Supervised machine learning is generally used to classify data or make predictions, whereas unsupervised learning is generally used to understand relationships within datasets.
* Supervised machine learning is much more resource-intensive because of the need for labelled data.
* In unsupervised machine learning it can be more difficult to reach adequate levels of explainability because of less human oversight.

**10. Describe the machine learning process in depth.**

**a. Make brief notes on any two of the following:**

**(i)MATLAB is one of the most widely used programming languages.**

**ii. Deep learning applications in healthcare**

**DEEP LEARNING IN HEALTHCARE**

Deep learning is assisting medical professionals and researchers to discover the hidden opportunities in data and to serve the healthcare industry better. Deep learning in healthcare provides doctors the analysis of any disease accurately and helps them treat them better, thus resulting in better medical decisions.

1. Drug discovery: Deep learning in healthcare helps in discovery of medicines and their development. The technology analyzes the patient's medical history and provides the best treatment for them. Moreover, this technology is gaining insights from patient symptoms and tests.
2. Medical imaging: Medical imaging techniques such as MRI scans, CT scans, ECG, are used to diagnose dreadful diseases such as heart disease, cancer, brain tumor. Hence, deep learning helps doctors to analyze the disease better and provide patients with the best treatment.
3. Insurance fraud: Deep learning is used to analyze the medical insurance fraud claims. With predictive analytics, it can predict fraud claims that are likely to happen in the future. Moreover, deep learning helps insurance industry to send out discounts and offers to their target patients.
4. Alzheimer's disease: Alzheimer is one of the significant challenges that medical industry faces. Deep learning technique is used to detect Alzheimer’s disease at an early stage.
5. Genome: Deep learning technique is used to understand a genome and help patients get an idea about disease that might affect them. Deep learning has a promising future in genomics, and also insurance industry. [Entilic says](https://www.enlitic.com/) that they use deep learning technique to make doctors faster and more accurate. [Cellscope](https://www.cellscope.com/) uses deep learning technique and helps parents to monitor the health of their children through a smart device in real time, thus minimizing frequent visits to the doctor. [Deep learning in healthcare](https://www.allerin.com/blog/3-ways-deep-learning-is-reinventing-the-healthcare-industry) can provide doctors and patients with astonishing applications, which will help doctors to make better medical treatments.

**iii. Study of the market basket**

In market basket analysis (also called association analysis or frequent itemset mining), you analyze **purchases that commonly happen together**. For example, people who buy bread and peanut butter also buy jelly. Or people who buy shampoo might also buy conditioner. What relationships there are between items is the target of the analysis. Knowing what your customers tend to buy together can help with marketing efforts and store/website layout.

Market basket analysis isn’t limited to shopping carts. Other areas where the technique is used include analysis of [fraudulent insurance claims](https://support.sas.com/resources/papers/proceedings14/1837-2014.pdf) or credit card purchases.

Market basket analysis can also be used to cross-sell products. [Amazon](https://www.amazon.com/gp/product/B00LAU8QCS/ref=s9_simh_gw_g193_i6_r?ie=UTF8&fpl=fresh&pf_rd_m=ATVPDKIKX0DER&pf_rd_s=desktop-1&pf_rd_r=0VJDRXV0B1ASKDM3Q2YG&pf_rd_t=36701&pf_rd_p=2079475242&pf_rd_i=desktop)famously uses an algorithm to suggest items that you might be interested in, based on your browsing history or what other people have purchased.

**iv. Linear regression (simple)**

**Linear Regression** is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used. It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a \*X + b.

The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above.

In this equation:

* Y – Dependent Variable
* a – Slope
* X – Independent variable
* b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

**11. Make a comparison between:-**

**1.Generalization and abstraction**

Abstraction is a technique to reduce the complexity of a problem by filtering out irrelevant properties while preserving all the important ones necessary to still be able solve a given problem. Generalization is a technique to apply knowledge previously acquired to unseen circumstances or extend that knowledge beyond the scope of the original problem. Humans show great capability in abstracting and generalizing knowledge in everyday life. RL needs abstraction and generalization as well to deal successfully with contemporary technological challenges, given the huge state and action spaces that characterize real world problems.

2. Learning that is guided and unsupervised

The main difference between supervised vs unsupervised learning is the need for labelled training data. Supervised machine learning relies on labelled input and output training data, whereas unsupervised learning processes unlabelled or raw data. In supervised machine learning the model learns the relationship between the labelled input and output data. Models are finetuned until they can accurately predict the outcomes of unseen data. However, labelled training data will often be resource intensive to create. Unsupervised machine learning on the other hand learns from unlabelled raw training data. An unsupervised model will learn relationships and patterns within this unlabelled dataset, so is often used to discover inherent trends in a given dataset.

So overall, supervised and unsupervised machine learning are different in the approach to training and the data the model learns from. But as a result, they also differ in their final application and specific strengths. Supervised machine learning models are generally used to predict outcomes for unseen data. This could be predicting fluctuations in house prices or understanding the sentiment of a message.

Models are also used to classify unseen data against learned patterns. On the other hand, unsupervised machine learning techniques are generally used to understand patterns and trends within unlabelled data. This could be clustering data due to similarities or differences, or identifying underlying patterns within datasets. Unsupervised machine learning can be used to cluster customer data in marketing campaigns, or to detect anomalies and outliers.

The main differences of supervised vs unsupervised learning include:

* The need for labelled data in supervised machine learning.
* The problem the model is deployed to solve. Supervised machine learning is generally used to classify data or make predictions, whereas unsupervised learning is generally used to understand relationships within datasets.
* Supervised machine learning is much more resource-intensive because of the need for labelled data.
* In unsupervised machine learning it can be more difficult to reach adequate levels of explainability because of less human oversight.

**3. Regression and classification**

Classification and Regression are two major prediction problems that are usually dealt with in Data mining and machine learning.

**Classification** is the process of finding or discovering a model or function which helps in separating the data into multiple categorical classes i.e. discrete values. In classification, data is categorized under different labels according to some parameters given in input and then the labels are predicted for the data.   
The derived mapping function could be demonstrated in the form of “IF-THEN” rules. The classification process deal with the problems where the data can be divided into binary or multiple discrete labels.

**Regression** is the process of finding a model or function for distinguishing the data into continuous real values instead of using classes or discrete values. It can also identify the distribution movement depending on the historical data. Because a regression predictive model predicts a quantity, therefore, the skill of the model must be reported as an error in those predictions

Comparison between Classification and Regression:

| Parameter | CLASSIFICATION | REGRESSION |
| --- | --- | --- |
| Basic | The mapping function is used for mapping values to predefined classes. | Mapping Function is used for the mapping of values to continuous output. |
| Involves prediction of | Discrete values | Continuous values |
| Nature of the predicted data | Unordered | Ordered |
| Method of calculation | by measuring accuracy | by measurement of root mean square error |
| Example Algorithms | Decision tree, logistic regression, etc. | Regression tree (Random forest), Linear regression, etc. |