1. **Explain the difference between supervised machine learning and unsupervised machine learning.**

In supervised machine learning we “teach” the program using already defined examples. Using these examples, the program can then draw accurate conclusions when it is given new data.

In unsupervised machine learning we don’t “teach” the machine anything. We give it a bunch of data and then it must figure out patterns and relations by itself. This could potentially lead to patterns and conclusions no human could have thought of.

1. **Why do we need machine learning?**

The real word tends to be very complicated. Problems can have millions of dimensions to consider and very complex predictor functions. When we can’t rely on numerical means, we need something else. Machine learning is one possible option.

1. **Name some different approaches to machine learning**Clustering, Artificial neural networks, Support vector machines, representation learning, Bayesian networks.
2. **What is classification?**Classification is when you try to identify in which group a new observation belongs based on a training set of data containing observations which group is known. Classification is a form of supervised learning.
3. **Explain what general principles can be used when choosing which data to use for training.**Make sure that the data you choose:

● is large enough so that it is statistically meaningful and can therefore produce a good result

● is a good representation of all possible data and not a selection of very unique cases

(Also make sure to never train on your test-data as this will completely invalidate your tests.)

1. **Explain how underfitting and overfitting happens.**Underfitting happens when you have a too small data set, and therefore no actual correlations could be drawn between input and (wanted) output.

Overfitting is when the training data does not represent the entire data set and will therefore not match the unseen data. When your model performs well on the training data but not on the actual evaluation data or when predicting. The model is adapted to the training data too hard and not generalized to predict unseen data

1. **Give at least three examples of what machine learning can be used for.**Recognize human faces, decipher hand-writing, detecting credit fraud, spamfilters, translation, detecting things in images, playing games such as chess, identifying ways to increase efficiency and save money, making routes more efficient and predicting potential problems.
2. **What does it mean that machine learning systems don’t use single value logic?**Machine learning will not give you a single exact answer but work with probabilities, it will not give you “this is the correct answer”, but rather “this is a probable answer”.
3. **What are some problems with machine learning?**We don’t know how the things we have created work in their entirety, they are a black-box, which can lead to problems. For example we can’t always determine whether the system made a correct prediction or not if the problem is subjective.
4. **What is a classifier?**A classifier is implemented by a classification algorithm. It maps input data to a category, it analyzes the data and signs a label to it as output. Gains domain knowledge with supervised learning. It is effectively putting the input data into a category/class. For example: The classes of deciphering handwritten digits would be 10 classes, 1 for each digit. (base 10).
5. **Explain how supervised learning is used to implement a classifier.**

Automatically derive a function from labeled training data. Labeled training data is input data points with the label being the correct answer for the prediction. Hopefully this allows the algorithm to then predict the class label of future unseen instance where we want the unknown label. We supervise the algorithm by feeding it data and telling it what the label is. **Collect training data -> Train classifier -> Make predictions**

1. **Explain briefly what Support Vector Machines (SVM) are used for and how they work.**The SVM model represents the examples of multi-dimensional datapoints in space. It is used for classification and regression analysis. SVM learns by supervised learning to divide the training data into subspaces (classes). For example, in the 2d plane (with 2d data points). We try to divide the plane in two by drawing a line that effectively divides the datapoints so that each class ends up on different sides of the line. Then when we want to classify a new datapoint, we see which subspace it goes into.
2. **Explain conceptually how unsupervised learning can be useful.**

When we do not have labels to our data and want to establish some structure in the data set, we can represent the data in a space and cluster the datapoints. This way the system can identify commonalities in the data. When new datapoints arrive, we can see which cluster the datapoint could belong to. Basically, classifying datapoints with the cluster as a label. Getting distinct clusters can help inferring a conclusion about what the cluster represents and thus the messy dataset has transformed into useful knowledge.

1. **Explain classification and when it can be used.**

Classification is when a system takes input data which already have labels on them

and uses it for training. The system then creates classification rules which can be used in order to assign labels to new input data. Classification can be used with Supervised Learning, for example if a bank wants to calculate if a new customer is low-risk or high-risk.

1. **Explain clustering and when it can be used.**

Clustering is when a system tries find clusters or groupings in the input by searching

for similarities within the input and group those into the same cluster. Clustering can

be used with Unsupervised Learning, for example if a company wants to know what

its demographic is by searching through its customer data to see which type of

customer most frequently occurs. It can also be used in image compression where

pixels with similar colour is grouped together in order to reduce the total number of

pixels.

1. **Describe a way to prevent overfitting.**

One way to overcome overfitting is to set aside parts of the training data which will

not be used for training the system and will not be visible to it during the learning

phases. This can be done by dividing the data into three parts, the first part will be

used to train the system, the second part will be used to test the system and sets the initial performance of it, the last part will be used to validate the system but no system adjustment will be done based on this data. Initially the performance of both the training and validation data will decrease but at some point in time the performance of the validation data will increase while it will still decrease for the training data. This is when overfitting starts and the training of the system will stop.

1. **What is the difference between Accuracy, Precision and Recall?**

The *accuracy* is the ratio of the number of correctly recognised patterns to the number of total classified patterns. The *precision* is the ratio of correct recognition of a specified pattern of the total number of recognitions of that pattern. Precision doesn’t consider when the system missed recognising the specified pattern but *recall* take that into consideration. Recall is the ratio of correct recognition of a specified pattern of the total number of that specified pattern.

1. **What is machine learning?**

Machine learning is the part of AI that helps machines learn to understand completely on their own by going through large amounts of data. Unlike pure programming where you're talking about the computer exactly what it should do, algorithms are used that allow the computer to interpret and learn from the data it processes to then predict patterns.

1. **Explain supervised learning.**

Supervised learning means that you have a data set where you already know the correct answer and want to create / learn a model that can then be used to make predictions from data where the correct answer is not known.

1. **Explain unsupervised learning.**

Unsupervised learning identifies patterns in data and tries to find similarities to divide data into categories. If the data is missing labels, it will not be able to train with supervised algorithms. Instead, using unattended learning, it is up to the algorithm itself to find patterns in the data to categorie it.

1. **What are the different use cases where machine learning algorithms can be used?**  
   Some use cases where ML can be used:

• Fraud Detection

• Face detection

• Natural language processing

• Spam detection

1. **Explain semi-supervised learning.**

Semi-supervised learning is a mixture of supervised and unsupervised learning. Semi-supervised learning uses both labeled and unlabeled data to learn and understand the complete dataset. Since labeled data is more expensive than unlabeled data, a combination of slightly labeled data and a large amount of unlabeled data is used instead.

1. **Explain linear regression, what are the pros and cons?**Using linear regression one can create a line through some data to predict future data, or hypothetical data. This is good because it would work in many cases where the data is quite linear. The drawback on the other hand is that a lot of data is not linear and can not be very well represented by a line at all.
2. **Explain reinforcement learning.**

A reinforcement learning system uses positive and negative reinforcement to learn. Positive reinforcement when it succeeds and negative reinforcement when it fails. An example is a robot that should pick up an object, if it fails to pick it up it is given negative reinforcement, if it succeeds it is given positive reinforcement. This reinforcement could be in form of a value it adds or subtracts.

1. **Explain the differences in how testing data and training data is used?**

The training data is used to create the model by training the machine using machine learning. Testing data is used for validation of the created model. Often a hold-out technique is used so that the testing data is never used for training. To do this, you reserve a portion of all the data you have specifically for testing/validation.

1. **Discuss how cross-validation is done and when?**

Cross-validation is used when the complete dataset is so small that you cannot afford to separate the data into separate pieces for training versus validation. Instead the total dataset is divided into k separate chunks/folds. During training, the training is performed on the remaining k-1 folds and the k-fold is used for validation. This continues with training done on k-2 and k-1 used for validation and so on until all sets have been validated against. The average of all validations is the performance of the model.

1. **Explain what an Epoch is?**

An Epoch is a complete run of training on all the training data. Often needs to train multiple epochs to bring down the error rate.

1. **Explain what a loss function is?**

A loss function is used to for the calculation of “error” of conclusion. The loss function such as mean squared error are used to penalize deviations from the target. Loss function are used to make the results of machine learning converge towards the most correct model.

1. **What is the structure of a ML program like? Describe each part briefly.**

The input, tool (model) and output. The input is about what data the program is given.The tool (or model) is about which algorithm to use to form the given input to the desired output with as high accuracy as possible (however not 100%). The output is what happens to the data once it has been processed by the program and is defined based on what the input data set is.

1. **Explain the difference between classifiers and clustering?**

Classifiers involves supervised learning. It takes input, interprets it and puts a label on it, based on predictions. Spam detection is an example.

Clustering is for unsupervised learning, where examples are grouped together into subsets by the machine learning algorithm by looking at the patterns and spotting similarities. K-means clustering is an example, where you classify into K numbers of groups. It's done by minimizing the distance from the cluster center to the sum of the individual elements.

1. **What is Dimensionality reduction(Decomposition)?**

Dimensionality reduction is reducing the number of random variables, taking into

account which ones are the crucial ones for the system. A typical example of usage of dimensionality reduction method is when we have high-dimensional datasets(more

than 10 dimensions), and we want to apply a K-nearest neighbors algorithm. In order to avoid the curse of dimensionality(various phenomena happening when one

manipulates data sets in higher dimensions), we use the dimensionality reduction.

1. **What is the difference between data mining and machine learning?**

Data mining is carried out by a *person*, in a specific situation, on a particular data set, with a goal in mind. Data Mining is about using Statistics as well as other programming methods to find patterns hidden in the data which can be used to *explain* some phenomenon.

Machine Learning uses Data Mining techniques and other learning algorithms to build models of what is happening behind some data so that it can *predict* future outcomes on the basis of self-build models.

1. **What is “genetic algorithms”?**

A genetic algorithm is mimicking natural selection by introducing random

mutations to new generations and allowing the most successful ones to become the base for new generations.

1. **What assumptions are made in the Semi-supervised machine learning?**

Continuity assumption - Points near each other are more likely to share labels

Cluster assumption - The data tend to form clusters. Points in the same cluster are more likely to share a label.

Manifold assumption - If the data lies in 3d but could be adapted to 2d.

1. **What are K-means and what type of machine learning is it?**

Unsupervised learning. When we have data that is not labeled and no output variables.

We need to find patterns and relationships. It uses centroids to form clusters.

1. **Explain how a regression algorithm works and which supervised category it falls into (supervised, semi-supervised or unsupervised)**

Regression is used to find a correlation between values in, for example, housing prices. The algorithm tries to find the most suitable formula for representing the house price - for example using a polynomial function, price = c1 + area \* c2 + area^2 \* c3, where the algorithm tries to figure out c1, c2 and c3 for some given house prices. By testing mathematical model best suits the data that is given, the algorithm learns how to predict a new houses price, depending on the calculated values for c1, c2 and c3 and the area of the

new house. This algorithm is supervised, since the prices have to made sure to be accurate, it cannot be assumed that the algorithm will choose the right formula, which may end up giving an inaccurate prediction.

1. **Explain what centroid clustering is and how it works.**

Centroid clustering is when datasets/datapoints are clusters relatively close to each other. This makes it possible to group the points by finding the middle point of all clusters and “draw lines” that separates the different groups. We can then use the lines to determine which category an input data belongs to.

1. **Machine learning consists of three parts, name those parts.**

Input: which have the datasets and information that the program is trained with.

Tool: Is the algorithm the program picks to interpret the input data.

Output: Classifying the data, depending on the tools probability.

1. **What are the advantages and disadvantages of using machine learning instead of manually formed rules?**

The main advantage is that it is a lot easier time efficient to use a machine learning

algorithm. Because the algorithm itself is generic you only have to fill in data. Maybe you have to spend some time on making the data ready to be used by the algorithm but all in all it is a lot easier and faster. The disadvantages are that the quality of the results very much depends on the quality of the filled in data. If the input data set has low quality, is too small or not representative the results are appropriate. Another big problem is that the decision making process is pretty much a black box. This can lead to the problem that the users simply don’t know why a decision is made and are also not able to detect wrong decisions.

1. **Describe and explain which problems can occur in an organization by using machine learning (especially after a long time)?**

The main problem that can occur after using machine learning for a long time is something I call the “black box” effect. This means that the users don’t know what exactly happens inside the decision making system. This makes it impossible to judge if the solution is correct or not. Especially after using such a system for a long time a organization maybe loses the ability to make these decisions by human experts because they don’t have any human expert anymore or the human experts lost their knowledge because they did not need to use it for a long time.

1. **Explain the k-nearest neighbor algorithm.**

At its most basic level, it is essentially classification by finding the most similar data points in the training data, and making an educated guess based on their classifications.

1. Compute a distance value between the item to be classified and every item in the

training data-set

2. Pick the k closest data points (the items with the k lowest distances)

3. Conduct a “majority vote” among those data points — the dominating classification in that pool is decided as the final classification.

1. **Discuss the drawbacks of machine learning? the technique is old, why wasn't it used much before?**

Machine learning is very dependent on it´s dataset, it can only base it output on it´s dataset and thus we need to provide the program with a lot (probably thousands) of examples of data that belongs to a category and do not belong to it. The program can get to much data in the dataset(overfitting) where it becomes hard for the program to classify inputs, the program can also have to little data in the dataset (underfitting). Both cases result in that it´s hard or almost impossible to categories the data.

The technique is old and has not been used because the computers were not strong enough.

1. **How does ML actually learn?**

The systems needs data in order to make a prediction or decision, als known as training data. Therefore the quantity and quality of the training data is vital. Feeding the system with random data isn't useful, since it would be impossible to draw any logical conclusion or see any patterns.

For example if you feed the system with many variety of images of dogs and cat, the system is able to process the data and learn from the data to distinguish a dog from a cat.

1. **What is the difference between feature and label in supervised learning?**

Features is (preferably) independent variables that represents categories/attributes in the data set. Features are the parameters that our system will analyze and sign a label to as output. When training using supervised learning, the label is given as a correct answer and the system correlates the features to the labels.

1. **What’s the difference between traditional programming and machine learning?**

The difference between traditional programming and machine learning is, that for traditional you only have the input data and the program that delivers the output. For machine learning you give both input and output data into the machine learning training approach and receive a model as a solution.

1. **What is a predictor function?**A predictor function (or hypothesis) is a function that is used in supervised learning. The “learning” consists of trying to optimize this function to a set of data.