Machine learning can be divided into three main types, which?
Supervised, unsupervised and reinforcement learning

What is a feature in the context of machine learning?
A feature describes a certain aspect of an object or phenomenon that we would like to classify. It is usually expressed as a numerical measurement.

Mathematically a classifier can be described as a function f(x;w1, w2,...,wn) -> Ω

a) what is x, w1, w1,...,wn and Ω?

x is a vector of input features, w1,w2,...,wn are weights or parameters of the claffifier that are adjusted in the training phase and Ω is a set of discrete class labels

b) What is the difference between regression and classification?

In classification we want to learn to predict a discrete class label. In regression we want to learn to predict a continuous variable, for example a temperature, probability, stock price etc.

c) How is learning of a classifier for a pattern recognition problem performed?

The parameters w1,w2,...,wn are adjusted by an optimization procedure so that the system obtains the optimal skill to classify training data.

It is important that a learning method is able to generalize. What does this mean?
A learning method generalizes well if it performs well on previously unseen data, for example that it can generalize what it has learned on training data to new data.

In supervised learning, we are usually training a classifier to minimize the error on training data. But what we really want is a low generalization error. How can we estimate the generalization error of a classifier?
The generalization error can be estimated by testing the classifier on a test data set that was not used in the training.

Why is it important to divide the data set into a training set and a test set when training a learning system?
To avoid over-fitting.

How can you notice if a supervised machine learning algorithm has overtrained?
The algorithm has overtraining if it performs much better on the training data than on a test data set which it has not seen before.

You have 900 labeled training samples and want to evaluate how well different algorithms perform for this data. Explain/sketch how you would do this with 3-fold cross-validation.
Divide the data set into equal parts, P1, P2, P3 with 300 samples each. Then train and evaluate 3 times as follows.
Train using P1 and P2 and evaluate using P3
Train using P1 and P3 and evaluate using P2.
Train using P2 and P3 and evaluate using P1.
The total performance is the average classification accuracy of the 3 runs.

Mention one advantage and one disadvantage of the k-nearest neighbor classifier.
Advantages are that k-nearest neighbors require no training and it is easy to implement. Disadvantages are that one must store all training examples and it takes a long time to classify if the training data set is large, as the distance to all training samples must be calculated.

What is the purpose of the “bias weight” in a perceptron?
The “bias weight” makes it possible to move the discriminating hyperplane away from the origin.

What happens if the input to the bias weight in a perceptron is set to the constant value 2?
The input value to the bias weight is typically 1, but can be any value except zero.

In gradient descent optimization, for example when training a neural network, what may be the effect of:
A too long step length? The effect of a too long step length may be an oscillating behaviour that may have the effect that we never converge to a local minimum.
A too short step length? The effect of a too small step length is that we move very slowly towards the local minimum, i.e. the training takes a long time.

What is the difference between batch learning and online learning?
In batch learning, all the training samples are used for updating the classifier. In online learning, the classifier is updated using one training sample at a time.

In supervised learning, why is it a problem to train a classifier by minimizing the number of false classifications using gradient descent? That is to minimize ∑Nk=1I(zk6=yk) where zk is the output from the classifier, yk the correct label, I(zk6=yk) is equal to 1 if zk6 = yk and 0 otherwise, and N is the number of training examples.
The cost function is piecewise flat and not differentiable at all points, so that we do not get any information from the gradient which is required in gradient descent.

What is the maximum margin principle that is used for example in SVM?
The maximum margin principle says that the decision boundary between two classes should be placed to that the distance between the boundary and the training data should be as large as possible, i.e. the error-margin should be maximized.

Are SVM particularly useful when the training data set is rather small or when it is very large?
SVM have good generalization properties and are usually better than e.g. back-propagatioon networks on small datasets. On large datasets however it gets computationally heavy since the kernel matrix grows quadratically with the number of samples.

What is the meaning of Covers theorem in words?
The probability that classes are linearly separable increases when the features are nonlinearly mapped to a higher dimension feature space.

Describe briefly how a neural network could be trained to predict the temperature for the next day.
One may use a neural network with 5 input nodes, one for the temperature of each of the 5 past days. Given historical data over 6-day periods, one can use the first 5 days as training data and the 6th day as the correct answer. The neural network can in this way be trained to predict tomorrow's temperature based on the past 5 days.

Is it appropriate to train the network with y = +/- 1 if we for example use tanh() as the activation function?
Not really, tanh(x) never becomes +/- 1, just very close. If you try to use +1 or -1 you may force the weights to become very large values and risk the robustness of the network. Try using a slightly smaller value. In practise it often does work though.

Would we gain anything by using a nonlinear activation function in the output layer, instead of a linear function?
The activation function in the output layer can not provide more advanced class boundaries or such. However it can bound the output values which is a good protection against outliers, e.g. verydevaiting training samples will not affect the solution as much with some kind of sigmoid activation function.

How can the problem of overfitting be avoided when training a neural network?
Use a large number of samples in each label per bin. We also want to use a relatively small learning rate so that the algorithm doesn't adjust to the noise in the data set.

What is a decision tree/CART?
One of the many models are selected, and the choice of model is a function of the input variables. In the decision tree, a sequence of binary decisions(ordered as a binary tree) yields the chosen model. The models are in the leafs of the tree and are often very simple e.g. “+1).

What is the difference between a classification tree and a regression tree?
The leafs of the regression tree hold real values, not class labels.

What is a decision stump?
A decision stump is a decision tree with a single node. It partitions the input space into two regions. The linear decision surface is perpendicular to one of the features axis.

Which parameters are we optimizing in a decision stump?
The polarity and the threshold.

Why is this cost function always <= after optimization?
If the performance of the classifier is not better than 0.5, we can always change the polarity(flip the signs) so that we get 1-error in performance.

What is bagging?
Bagging, short for bootstrap aggregation, is a committee method where the individual models are trained on separate bootstrap data sets that are created from the original dataset using random sampling with replacement.

What is boosting?
Boosting is also a committee method but the base models are trained in sequence, and each base model is trained using a weighted form of the data set in which the weighting coefficient associated with each data sample depends on the performance of the previous classifiers.

What is a weak classifier?
A weak classificer is a simple classifier that may perform only slightly better than a random classifier. That is for two classes, a weak classifier may give a classification accuracy of only slightly above 50%. The opposite of a weak classifier is a strong classifier which aims to have a very high classification accuracy.

Describe one weak classifier.
Small decision trees or even decision stumps are typical weak classifiers.

Describe the difference in feedback between supervised and reinforcement learning.
In supervised learning, the feedback is the correct action or class for every input, situation or state. In reinforcement learning a scalar feedback(reward or punishment) is obtained for taking certain actions or reaching certain states, but the correct or optimal actions are not given to the learning system.

Describe the “temporal credit assignment problem”
The problem refers to the problem of assessing individual actions in a sequence of actions when the reward/punishment comes delayed. For example, which were the winning moves that led to the victory in a game.

What is the difference between the value function V and the Q function in reinforcement learning?
The V-function describes the value(expected reward) that will be obtained in the future when being in a certain state, when following a certain policy (which action to take in every state). Each policy has a different value function. The Q function describes the value for each action in each state, given that the optimal policy is followed after the action has been taken.

Explain the principle of greedy exploration.
The discount factor controls the trade off between optimizing for immediate rewards and long term rewards.

What is meant by hard clustering and soft/fuzzy clustering?
Hard clustingering means that each training sample is assigned to only one cluster, whereas soft/fuzzy clustering means that a training sample can belong to several clusters to certain degrees.

What is the optimization algorithm called that is used in k-Means clustering and mixture of gaussian clustering?
Expectation maximization, commonly known as the EM-algorithm.

What is the difference between k-means clustering and mixture of gaussian clustering?
In k-means clustering, only the centers of the clusters are modelled, whereas in mixture of gaussian clustering, also the cluster shape is modelled using the covariance matrix, which can model circular and elliptical cluster shapes.

What is determined by the parameter k in the k-nearest neighbors and k-means algorithms respectivly?
k-NN: k is the number of the stored data vectors that vote for the classification decision.
k-means: k is the number of clusters.

What is defined by a kernel function?
A kernel function defines the inner product(x1,x2)=Φ(x1)TΦ(x2) in the new feature space. Thus κ(x1,x2) specifies the feature space by defining how distances and angles are measured, instead of explicitly stating the mapping function Φ(x).

Explain the principal differences between LDA and PCA, if any
While PCA maximizes the variance globally LDA maximizes the ratio of the between class and within class variances. This also means that LDA needs some sort of pre classified training data.

What is being optimized in LDA?(Linear Discriminant analysis)
The quotient between the means of the clusters and the variance of the projections of the clusters.

Mention one method that can speed up the gradient descent method.
Add a momentum term.

When can a sigmoid function in the output layer improve the accuracy?
It can improve it when the distribution of the two classes are different. Since the boundary will be different between the two classes centroids.

What is the benefit of a convolutional layer instead of a fully connected layer in a multilayer perceptron?
There are fewer parameters to train in the convolutional layer, which reduces the risk of overfitting and therefor give better generalization properties.

Give two examples of the general algorithm Expectation Maximization.
K-means algorithm and mixture of gaussians

What is the difference between gradient descent and expectation maximization?
The expectation maximization is a model using latent variables that are discrete and the algorithm iterates over estimating the model parameters and the latent variables.

Why does convergence get slower in deep networks when using the sigmoid function?
The sigmoid function leads to vanishing gradients. Since the gradients from all layers are multiplied in the chain rule, deep networks suffer exponentially from this. This can be avoided by using the relu function instead of the sigmoid.

What is defined by the kernel function?
The kernel function k(xi,xj) defines a scalar product between two vectors xi and xj that have been mapped to a (usually high dimensional) feature space via a function ϕ()

What is required from the optimization method in order to be able to solve it with a kernel method?
What is required is that we need to be able to formulate the problem in terms of scalar products between the samples, so that we can swap in a nonlinear function, the so-called kernel trick. A first step to do this is to show that the optimal solution can be written as a linear combination of the training samples.

How does the brute force optimization method work?
You test values of w in the error(w) function, and simply choose the w that gives the least error.

When should you use genetic methods/algorithms?
Genetic algorithms should be used when we are dealing with discrete variables which are not differentiable, and therefore cannot use gradient descent.

How can one implement a reward function r(s,a) that promotes the shortest path problem?
By initializing the negative reward in each state except for the end state.