第一讲
机器学习概念
Lecture 1
Concepts of Machine Learning

明玉瑞 Yurui Ming

yrming@gmail.com

声明
Disclaimer

什么是机器学习
What's Machine Learning (ML)

- 显式编程（Explicit Programming）

  指定算法的每一步骤

  Specify each step of the algorithm

  - 以查找二分查找算法为例

    Take binary search algorithms for example

```
BinarySearch(list[], min, max, key)
while min ≤ max do
    mid = (max+min) / 2
    if list[mid] >key then
        max = mid-1
    else if list[mid] <key then
        min = mid+1
    else
        return  mid
    end if
end while
return  false
```


Search the number 7 in the array
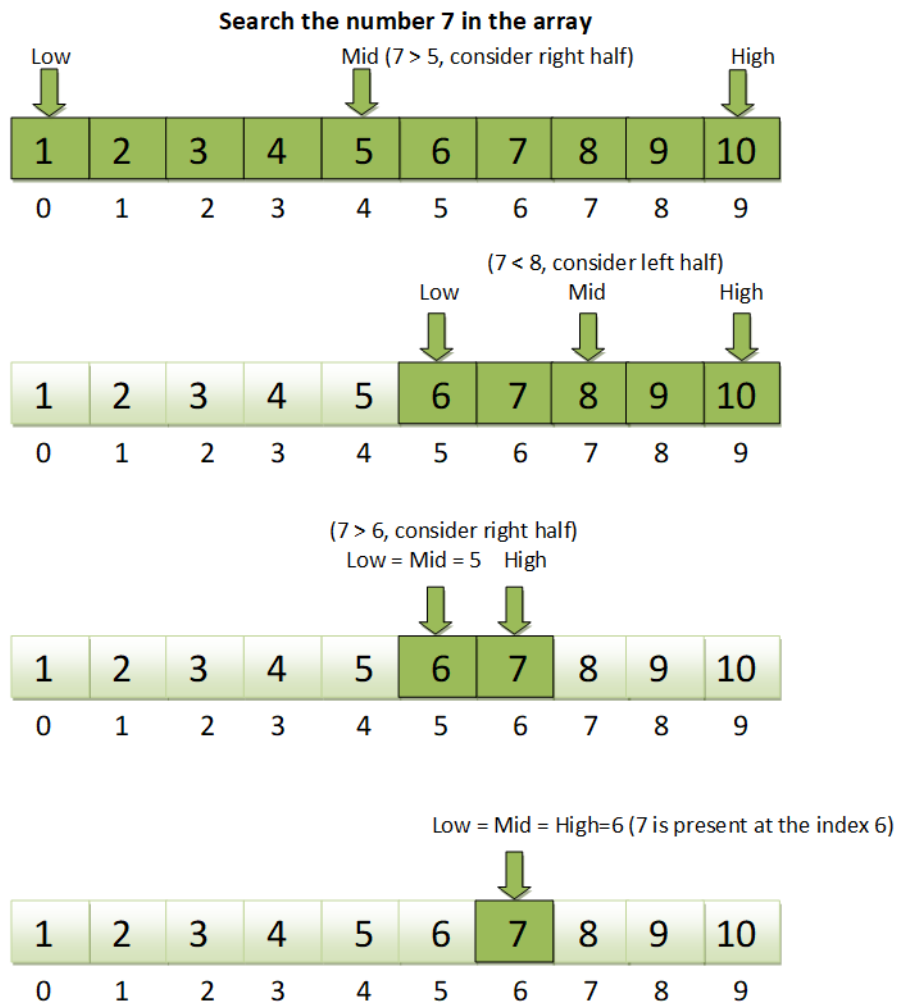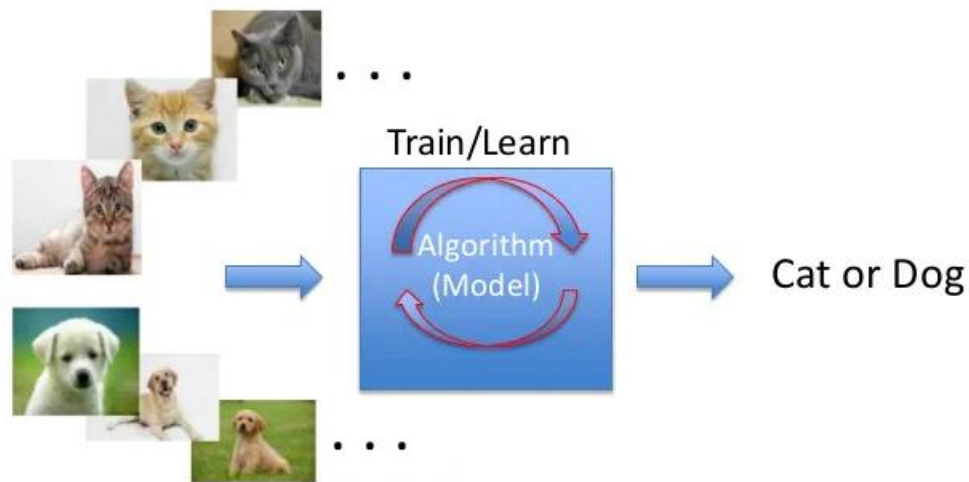
什么是机器学习
What's Machine Learning (ML)

▶ 隐式编程（Implicit Programming）

由模型或算法自动学习解决问题的步骤

The procedures to solve the problem are learned by the model or the generic algorithm via data-driving.

  ▶ 以图像分类为例子

    Exemplified by image classification

什么是机器学习
What's Machine Learning (ML)

- 人工智能（Artificial Intelligence或AI），机器学习（Machine Learning或ML）与深度学习（Deep Learning或DL）三个概念之间的定义，区别与联系

  - 人工智能：理解生物智能并尝试据此构建机器智能程序，使之可以创造性地解决一些问题；其任务之一是理解智能。

    AI: To understand biological intelligence and based on which to design computer algorithms to solve problems creatively; one of the major task of AI is to understand intelligence

  - 机器学习：被认为是人工智能的一个子集，强调无需显式编程，从数据或经验中学习提高，达到解决同类问题的目的。

    ML: a subset of AI, which emphasizes on no-explicit programming and directly learning and improving from data or experience, and the subsequent generalization to homogeneous problems
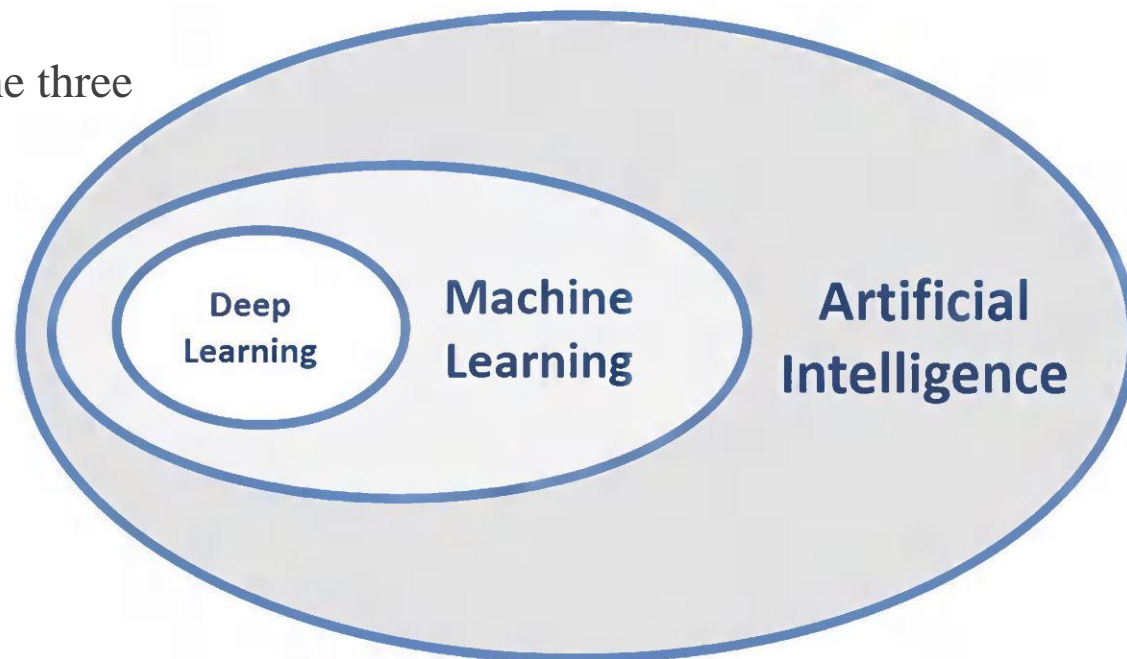
什么是机器学习
What's Machine Learning (ML)

▶ 深度学习：被认为机器学习的一个子集，依靠神经网络模型而不是其它如贝叶斯方法来解决从数据中学习的问题。

DL: a subset of ML, which relies on the theories and practice of neural networks instead of other methods such as Bayesian methods to learn from data to solve problems posted in ML.

▶ 三者之间的关系

The relations between the three

机器学习分类
Categories of Machine Learning (ML)

▶ 机器学习主要包括三种类型：

ML is mainly categorized into three paradigms:

  ▶ 监督学习（Supervised Learning）

  ▶ 无监督学习（Unsupervised Learning）

  ▶ 强化学习（Reinforcement Learning）

机器学习分类
Categories of Machine Learning (ML)

- 监督学习（Supervised Learning）

  - 监督学习是通过提供输入数据和标签引导计算机学习来完成的。例如，可以向机器提供带有相应标签（例如小猫）的图片，引导计算机学习小猫的内在特征，以便进行后续的泛化。之后，如果我们提供了另一只没有标签的小猫的图像，则机器将能够预测该图像就是小猫的图像。

    Supervised Learning is a learning paradigm based on example input data with corresponding ground-truth labels to train the models into optimal states for inference.
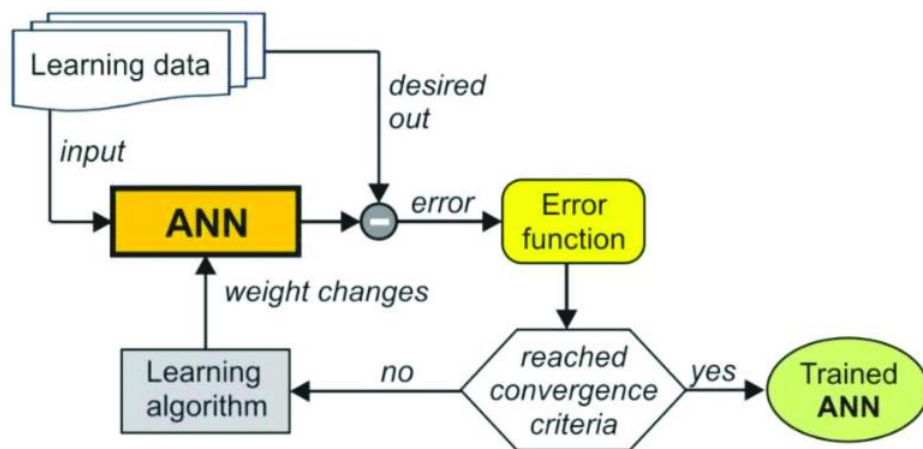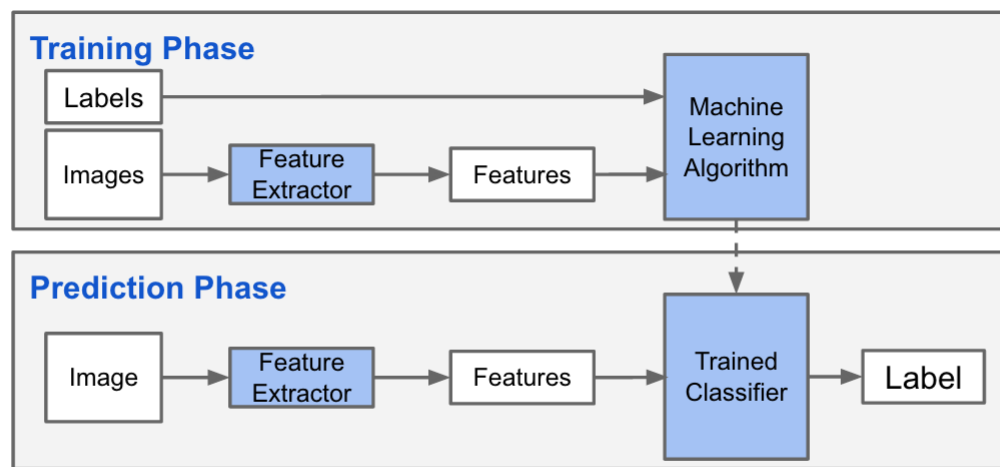
机器学习分类
Categories of Machine Learning (ML)

▶ 监督学习（Supervised Learning）

　　▶ 监督学习通常包括两个阶段：训练和测试阶段。

　　The process of applying SL consists of two stages: training and testing.

　　▶ 基于监督学习的例子包括支持向量机，神经网络等

　　Algorithms and techniques used for supervised learning include support vector machine (SVM), neural networks (NN), etc.
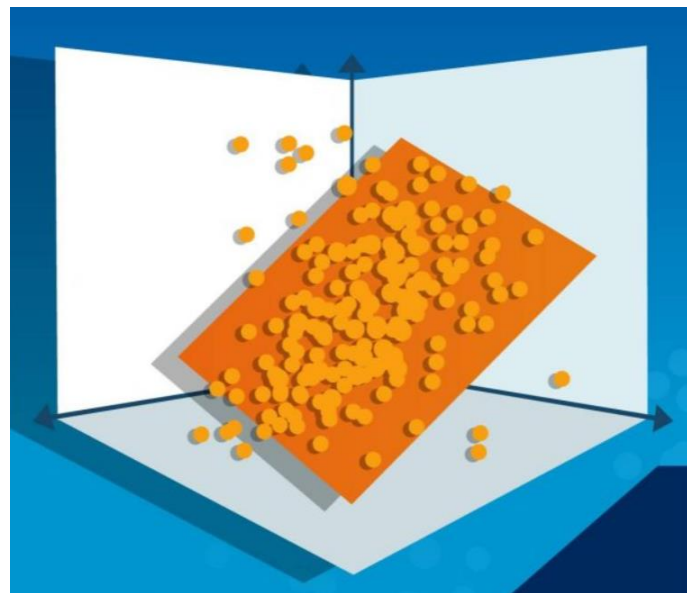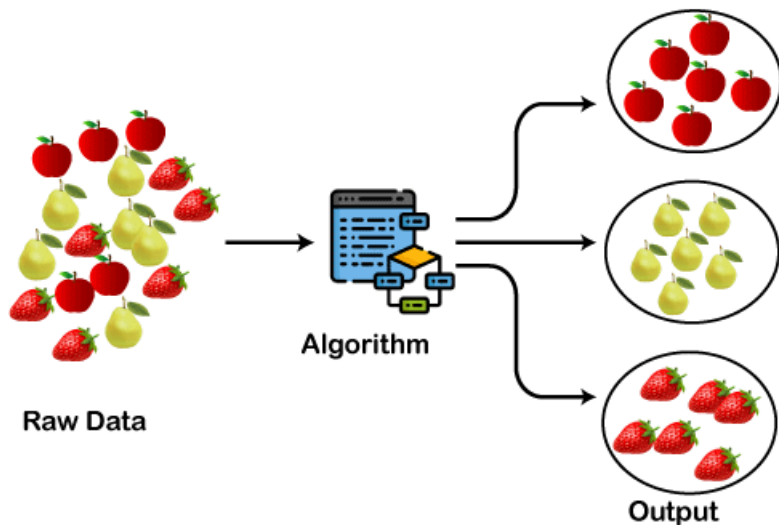

Machine Learning Phases

机器学习分类
Categories of Machine Learning (ML)

▶ 非监督学习（Unsupervised Learning）

   ▶ 无监督学习是从不包含标签的数据集中得出推论的算法，主要包括聚类分析和降维等方法。

   Unsupervised learning (UL) is a type of learning that learns patterns from untagged data by building an appropriate internal representation of the data. Methods belonging to UL include clustering and dimension-reduction.

机器学习分类
Categories of Machine Learning (ML)

▶ 非监督学习（Unsupervised Learning）

  ▶ 聚类分析主要用于根据数据的内在属性分组；降维主要用于减小特征的维度，创建更简明的机器学习模型。这两种方法均可用于如数据的可视化。注意非监督学习得到的结果并不准确或唯一，主要，因此要求待解决的问题具有一定的容错性。

  Clustering is mainly used for grouping or categorization, and dimension reduction is to merge correlated features to create more concise machine learning models. Both methods can be used for data visualization. Notably, Unsupervised Learning usually leads to results which is not unique, therefore, it tends to require that the problems to be solved can tolerant the imprecision;

  ▶ 基于监督学习的例子包括K均值聚类，主成分分析，自组织映射等

  Algorithms and techniques used for unsupervised learning include K-means clustering, principal component analysis (PCA), self-organizing maps (SOM), etc.
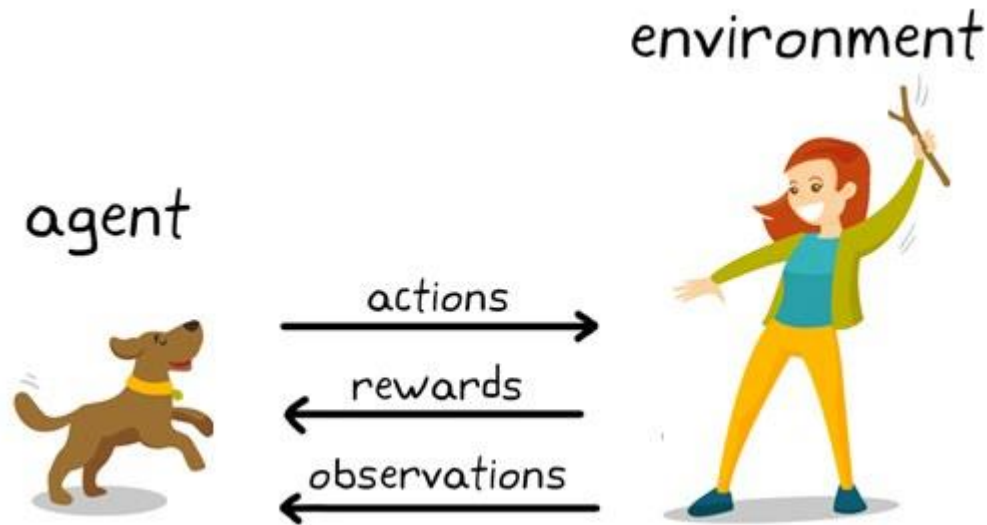
机器学习分类
Categories of Machine Learning (ML)

▶ 强化学习（Reinforcement Learning）

  ▶ 强化学习是智能体通过与环境进行交互，即执行某些操作并观察从这些操作中获得的奖励来学习如何响应环境中的特定状态。

  Reinforcement learning (RL) concerns with the approaches that how intelligent agents effectively interact with the environments by taking actions.

机器学习分类
Categories of Machine Learning (ML)

- 强化学习（Reinforcement Learning）

  - 强化学习主要通过试错方式，学习从环境状态到动作的映射，以使动作从环境中获得的累积奖励最大化。

    By a trial-and-error way, the main goal of reinforcement learning is to learn the optimal mapping between states and the corresponding actions, in order to maximize the notion of cumulative reward in the long run.

  - 强化学习主要包括算法主要包括Q学习算法，动作器-评判器算法，近端策略优化算法等等。
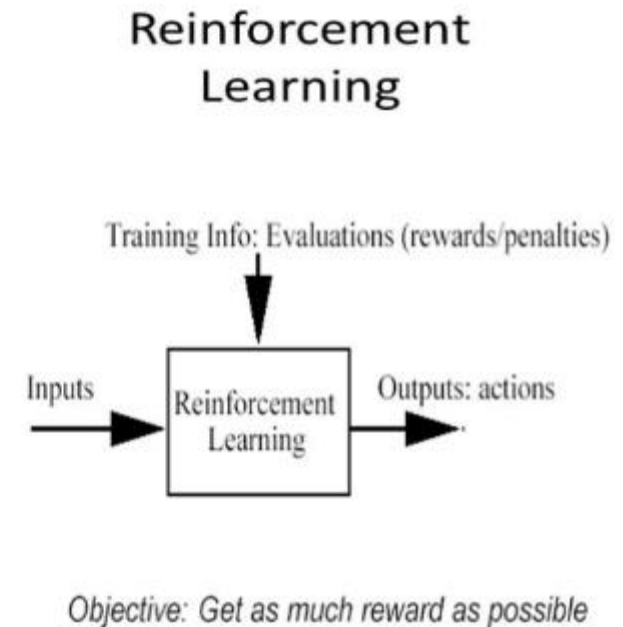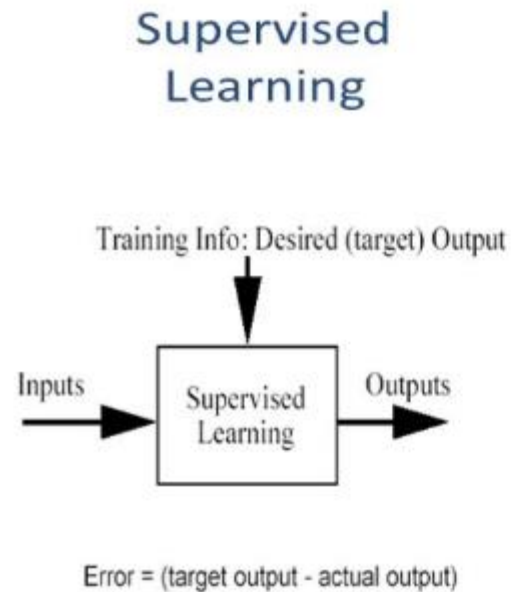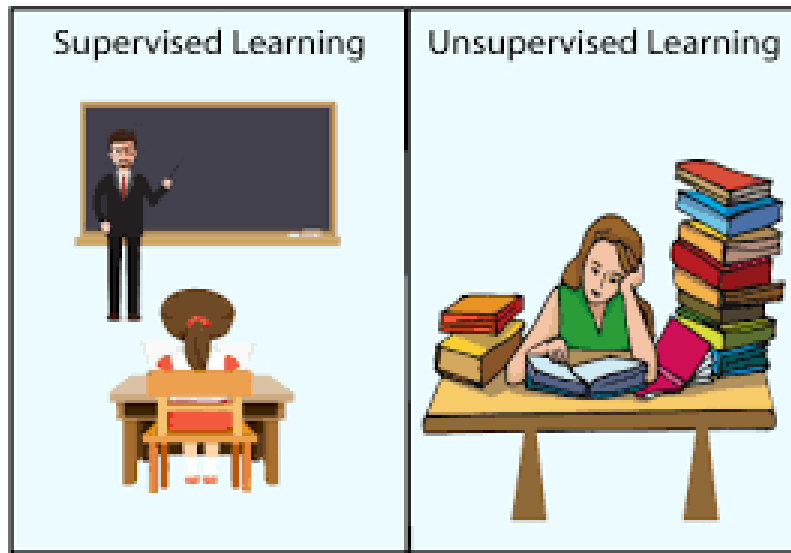
    Methods of reinforcement learning includes Q-Learning, Actor-Critic algorithm and Proximal Policy Optimization method, etc.

机器学习分类
Categories of Machine Learning (ML)

▶ 三种学习方式的比较

Comparison of these three learning paradigms

机器学习要素
Elements of Machine Learning (ML)

▶ 解决机器学习问题的组成要素：

Elements of machine learning problems:

❖ 数据（data）；

数据即模型输入，数据的特征取决于问题领域（Domain）；

Data are the inputs to the models, might contain ground-truth labels or annotations. The features of the data are domain dependent.

❖ 模型（model）

模型即解决问题所采用的算法；

Model are the adopted algorithms to solve the problems.

❖ 目标或损失函数（Loss function）；

目标或损失函数依机器学习的范式而定，如果是监督学习，则一般有显示的（Explicit）损失函数，若是非监督学习，损失内蕴于（Intrinsic）数据或样本本身

机器学习要素
Elements of Machine Learning (ML)

Loss function usually depends on the learning paradigm. For supervised learning, there exists explicit loss function by comparing the model outputs with ground-truth information. For unsupervised learning, the loss is intrinsic to data structures or representations.

❖ 学习方法（Learning algorithm）；

学习方法即为使模型达到目标采取行动时所依照的法则

Learning method is the rules to take for the model to reach optimal performance during learning.

机器学习要素
Elements of Machine Learning (ML)

▶ 示例（Example）

  ▶ 以神经网络为例，神经网络本身构成了对问题进行建模的模型；损失函数则依据是分类问题或回归问题，通常采用对数损失函数或平方损失函数；学习方法即反向传播算法。

  Exemplified by neural networks, which themselves are the models targeting the problems. And the categorization of loss functions depends on the properties of the problems, namely, they are for classification or for regression, and consequently can be minus logarithm function or mean squared error function. For learning method, it is backpropagation.
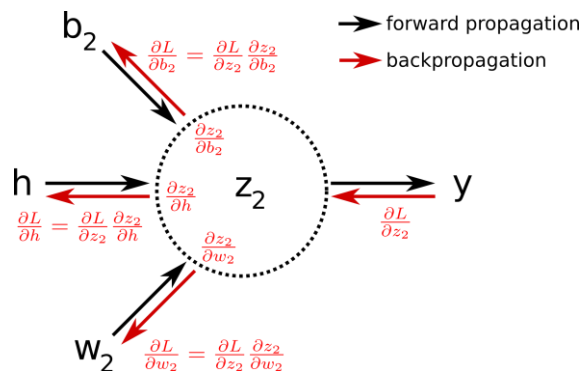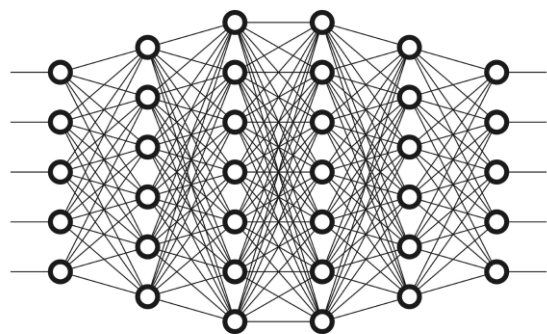
$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

习题
Problems

1. 利用神经网络的监督学习的模型可以为概率模型或者非概率模型，考虑样本 $(\vec{x}, y)$，则当模型分别表示为 $y = p(y|\vec{x})$ 与 $y = f(\vec{x})$ 时，请指出哪个是概率模型，哪个是非概率模型。

    The supervised learning by using neural networks can be interpreted as probabilistic model or non-probabilistic model. Considering the sample $(\vec{x}, y)$, when the model is represented as $y = p(y|\vec{x})$ or $y = f(\vec{x})$, then decide which is probabilistic model and which is non-probabilistic model according to your understanding.

2. 阅读相关文献，解释什么是半监督学习（Semi-supervised Learning）。

    By doing literature survey to explain the method of semi-supervised learning.

3. 监督学习又分为生成方法和判别方法，所用到的模型分别称为生成模型和判别模型。请阅读相关文献，解释什么是生成模型，什么是判别模型。

    Supervised learning can also be categorized as generative methods and discriminative methods, and the corresponding models are called generative models and discriminative models. By doing literature survey to explain what's generative model and what's discriminative model.

4. 深度学习作为机器学习的一个子域，也继承了机器学习中的一些概念，如参数和超参，请阅读相关文献解释这两个概念。

    As a subset of machine learning, deep learning also inherits some concepts of machine learning, for instance, parameters and hyper-parameters. Try to explain and distinguish these two concepts by doing literature survey.