Week4: Supervised and Unsupervised Learning

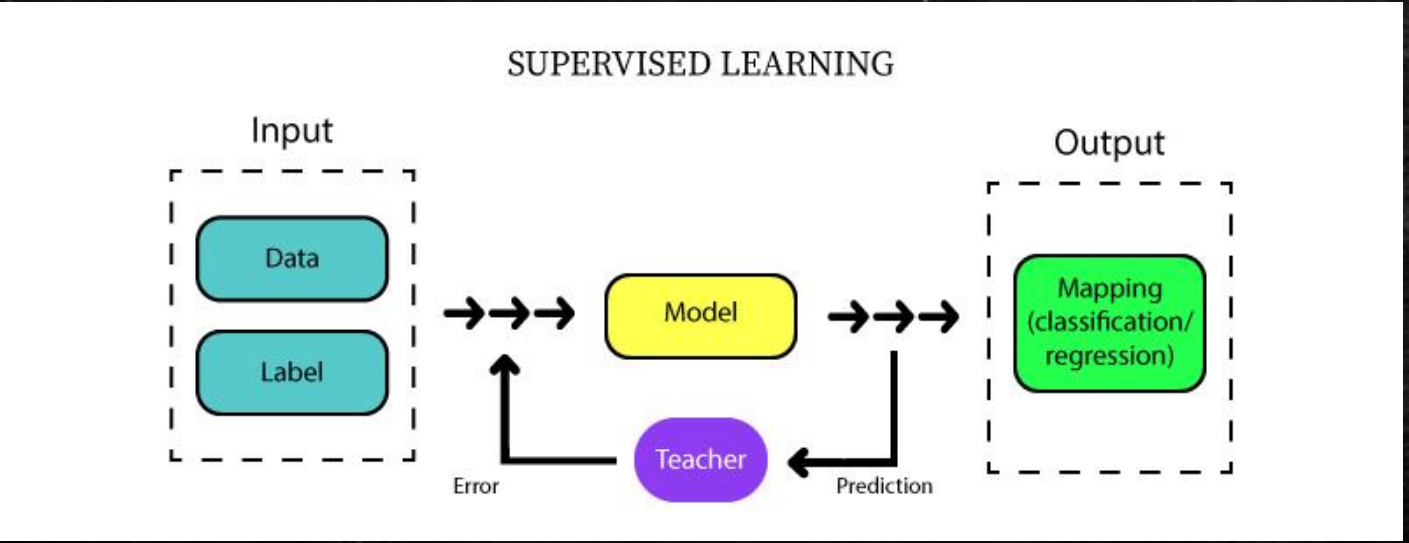
Shayan Dadman, PhD candidate
UiT, Narvik

Supervised learning

- Supervised learning is the standard approach in machine learning problems.
- It is called supervised learning, as a teacher supervises the learning process.
- In supervised learning, we provide input variables (x) and output variables (Y).
- The algorithms are used to learn the mapping function from the input to the output.

Supervised learning

- Indeed, the goal is to obtain a mapping function that can predict output variables (Y), given the input variables (x).
- During the training process, we know the correct answers. Thus, the algorithm iteratively makes predictions on the given input variables (x) and is corrected by the teacher.



Supervised learning

- Supervised learning is typically utilized in the context of classification and regression problems.
- Traditional algorithms in supervised learning are k-nearest neighbors, logistic regression, support vector machines, random forests.
- When applying supervised learning, the main concerns are the model complexity and the bias-variance tradeoff.

Supervised learning: model complexity

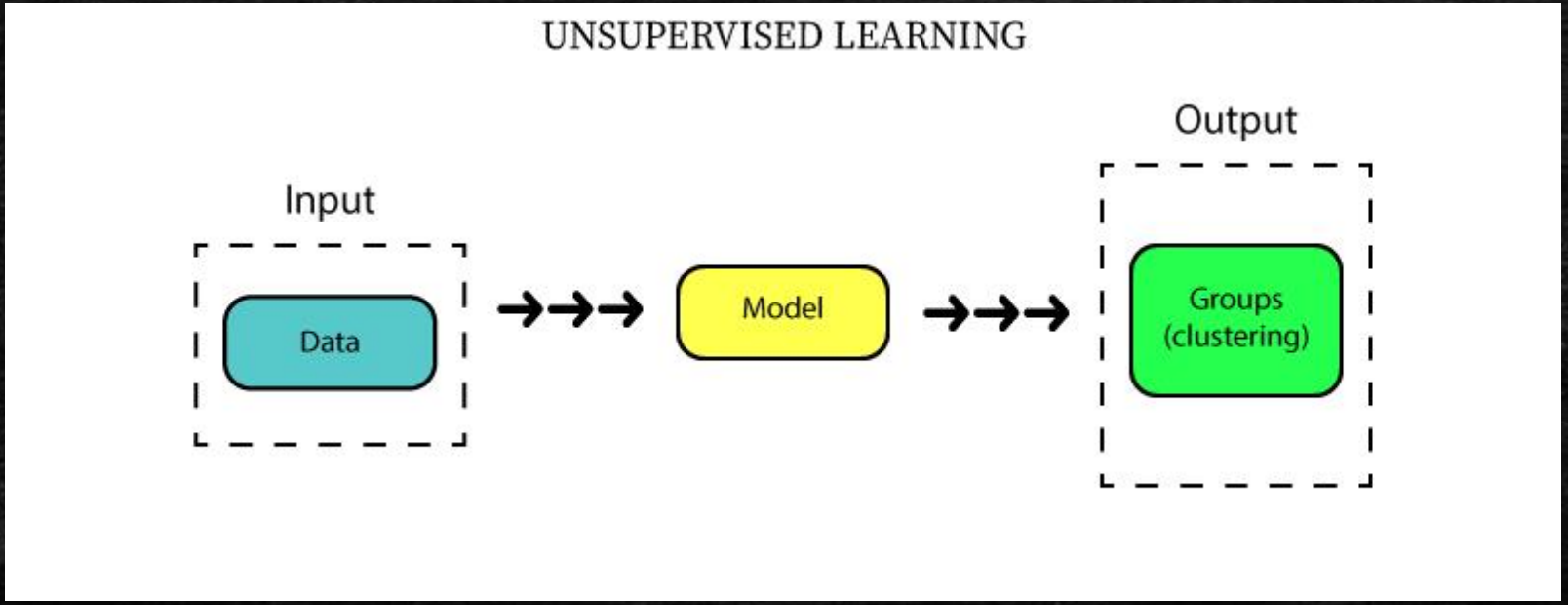
- The model's complexity refers to the complexity of the function we are trying to learn.
- If the dataset is small or it is not uniformly spread out, then the low-complexity model is the way to go.
- This is because the high-complexity model will overfit if utilized on a small amount of data.
- Indeed, the model's learning function fits the training data very well that it cannot generalize to other data points.

Supervised learning: bias-variance tradeoff

- The bias-variance tradeoff relates to model generalization.
- Bias is the constant error term. In other words, it is the difference between the average prediction of our model and the expected value.
- Variance indicates the distribution of model prediction for a given data point.
- The model with high bias pays little attention to the training data and oversimplifies the model. In fact, it results in a high error on training and test data.
- On the contrary, the model with high variance pays a lot of attention to training data. It performs well on training data but poorly on test data. Indeed, it lacks the ability to generalize on unseen data.

Unsupervised learning

- Unsupervised learning is when you only have input variable (x).
- It is called unsupervised learning, because in contrary to supervised learning, there is no correct answers and a teacher.



Unsupervised learning

- In unsupervised learning, the goal is to model the structure in the given input variables (x) by extracting the features and analyzing its structure.
- In this approach, the algorithms are left to ther own to find the existing structure in the given inputs.

Unsupervised learning

- The typical tasks within unsupervised learning are clustering, representation learning, and density estimation.
- Common algorithms utilized for such tasks include k-mean clustering, principal component analysis, and autoencoders.
- There is no explicit way to examine model performance in unsupervised learning. It is because no labels are provided.

Unsupervised learning

- Two prevalent use cases for unsupervised learning are exploratory analysis and dimensionality reduction.
- In an exploratory analysis, the unsupervised learning algorithms can be utilized to identify structure in data when it is either impossible or impractical for humans.
- In such situations, unsupervised learning can provide initial insights into the relationship between the given data points.
- In the case of dimensionality reduction, the unsupervised learning methods can be used to represent data in fewer columns or features. The principal component analysis is commonly used for this purpose.

Questions

1. Email me at Shayan.Dadman@uit.no
2. My office D3430