AIFS ML Lecture 3: Machine Learning Basics

Suraj Narayanan Sasikumar

Hessian AI Labs

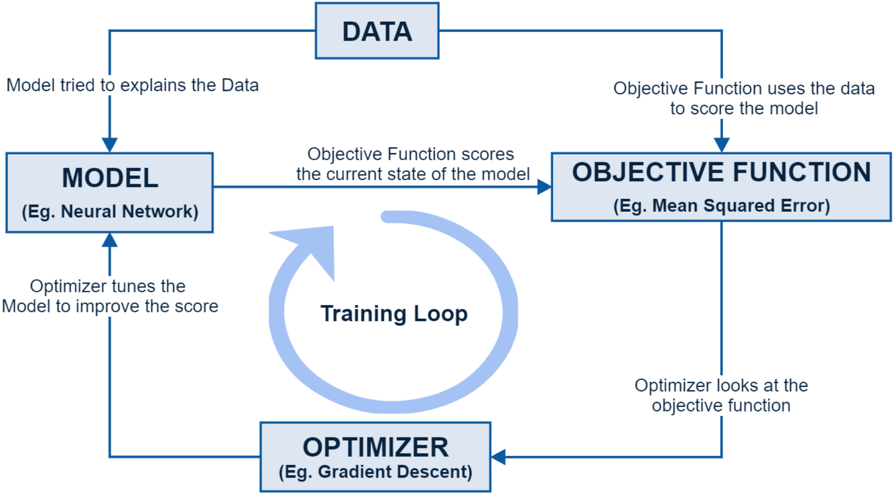
April 26, 2020

Overview

1. Recap
2. Learning Styles in Machine Learning
   - Based on Degree of Supervision (prev. lecture)
   - Based on the mode of Consuming Data
   - Based on Type of Model
3. Example: Linear Regression

Table of Contents

Recap - Training Loop



Learning Style based on Degree of Supervision

- **Supervised:** The dataset provides a supervisory signal in the form of a label (*aka* target) to inform the learning algorithm what the true prediction should have been. Example tasks: *Classification*, *Regression*
- **Unsupervised:** No explicit prediction or inference, the goal of the algorithm is to extract patterns about the underlying data generating process from the dataset. Example tasks: *Clustering*, *Dimensionality Reduction*
- **Reinforcement:** The science of sequential decision making for artificial agents, such that the decisions taken by the agent in an environment maximizes a notion of cumulative reward. Due to the presence of a reward signal that does not fully inform the agent of the absolute true response, Reinforcement Learning is neither supervised nor unsupervised.

Table of Contents

Based on mode of Consuming Data

Learning algorithms can be categorized based on their capability to process data that becomes available sequentially.

- Batch Learning - incapable of handling new data on-the-fly, needs to look at the entire dataset.
- Online Learning - can handle new data on-the-fly.

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:
  - For algorithms that leverage iterative methods, the convergence of the optimizer is more stable since the data is static.

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:
  - For algorithms that leverage iterative methods, the convergence of the optimizer is more stable since the data is static.
  - No risk from new data affecting the performance of a previously trained model.

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:
  - For algorithms that leverage iterative methods, the convergence of the optimizer is more stable since the data is static.
  - No risk from new data affecting the performance of a previously trained model.
- Disadvantages:

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:
  - For algorithms that leverage iterative methods, the convergence of the optimizer is more stable since the data is static.
  - No risk from new data affecting the performance of a previously trained model.
- Disadvantages:
  - For large datasets the optimizer has to go through all the data points before tuning the parameters of the model. This makes the process very slow and prohibitive in most real world situations.

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:
    - For algorithms that leverage iterative methods, the convergence of the optimizer is more stable since the data is static.
    - No risk from new data affecting the performance of a previously trained model.
- Disadvantages:
    - For large datasets the optimizer has to go through all the data points before tuning the parameters of the model. This makes the process very slow and prohibitive in most real world situations.
    - Having to go through the entire dataset during a single training loop means that the data has to be fully loaded onto the RAM. In most real world situations where machine learning is useful, the size of the dataset far exceeds the size of the RAM.

Batch Learning

- During each iteration of the training loop, the learning algorithm processes the entire dataset. These algorithms cannot learn incrementally from new data. In order to learn from the new data, the model has to be trained from scratch using the old and new data.
- Advantages:
  - For algorithms that leverage iterative methods, the convergence of the optimizer is more stable since the data is static.
  - No risk from new data affecting the performance of a previously trained model.
- Disadvantages:
  - For large datasets the optimizer has to go through all the data points before tuning the parameters of the model. This makes the process very slow and prohibitive in most real world situations.
  - Having to go through the entire dataset during a single training loop means that the data has to be fully loaded onto the RAM. In most real world situations where machine learning is useful, the size of the dataset far exceeds the size of the RAM.
  - Cannot be used in situation where we have streaming data like live CCTV footage, stock prices etc.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:
  - Can break down huge datasets into mini-batches to separately load them into memory for training.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:
  - Can break down huge datasets into mini-batches to separately load them into memory for training.
  - Faster learning since the model can be updated after only looking at a small portion of the dataset.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:
    - Can break down huge datasets into mini-batches to separately load them into memory for training.
    - Faster learning since the model can be updated after only looking at a small portion of the dataset.
    - Able to handle streaming data.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:
  - Can break down huge datasets into mini-batches to separately load them into memory for training.
  - Faster learning since the model can be updated after only looking at a small portion of the dataset.
  - Able to handle streaming data.
- Disadvantages:

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:
  - Can break down huge datasets into mini-batches to separately load them into memory for training.
  - Faster learning since the model can be updated after only looking at a small portion of the dataset.
  - Able to handle streaming data.
- Disadvantages:
  - When processing streaming data, there is a possibility of adversely affecting the performance of the AI system if the data is bad. Garbage in, Garbage out.

Online Learning

- During each iteration of the training loop the learning algorithm processes a small subset of the dataset called a mini-batch. This allows the algorithm to learn from new data on-the-fly.
- The mini-batch can just be a single data-point, thus enabling online learning algorithms to incrementally learn from steaming data.
- Advantages:
  - Can break down huge datasets into mini-batches to separately load them into memory for training.
  - Faster learning since the model can be updated after only looking at a small portion of the dataset.
  - Able to handle streaming data.
- Disadvantages:
  - When processing streaming data, there is a possibility of adversely affecting the performance of the AI system if the data is bad. Garbage in, Garbage out.
  - Since at any given moment the learning algorithm is seeing only partial data, the optimizer takes a longer time to converge.

Based on Type of Model

The model used by the learning algorithm can be broadly categorized as:

- **Parametric**: When the model has a finite set of adaptable parameters (knobs) that can be adjusted to adequately model the data generating process.
- **Non-parametric**: These models cannot be parameterized by a finite number of parameters, and hence their structure is determined from the data rather than specified apriori.

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours
- Advantages:

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours
- Advantages:
  - The capability of the model to generalize to the true data generating process increases as the size of the dataset increases.

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours
- Advantages:
  - The capability of the model to generalize to the true data generating process increases as the size of the dataset increases.
  - Learning new data is as simple as storing it in memory.

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours
- Advantages:
  - The capability of the model to generalize to the true data generating process increases as the size of the dataset increases.
  - Learning new data is as simple as storing it in memory.
- Disadvantages

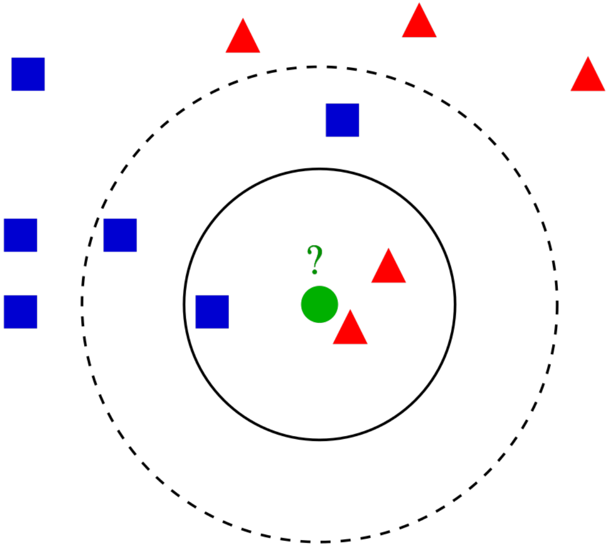
Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours
- Advantages:
  - The capability of the model to generalize to the true data generating process increases as the size of the dataset increases.
  - Learning new data is as simple as storing it in memory.
- Disadvantages
  - Computational complexity of performing prediction (eg. classification) increases linearly with the size of the dataset. Prohibitive in real-world datasets

Non-parametric Model

- Rather than learning about the the underlying data-generation process, these model uses the data directly to perform inferences and predictions.
- It's also called memory-based learning because these model memorize the data to compare new data with similar data seen during training.
- These models use a *measure of similarity* to compare new a data point to already learned data points.
- Eg. Algorithm: K-Nearest Neighbours
- Advantages:
  - The capability of the model to generalize to the true data generating process increases as the size of the dataset increases.
  - Learning new data is as simple as storing it in memory.
- Disadvantages
  - Computational complexity of performing prediction (eg. classification) increases linearly with the size of the dataset. Prohibitive in real-world datasets
  - In high dimensional datasets, the space would only be sparsely populated.

K-Nearest Neighbours



Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.
- Eg. Algorithms: Linear Regression, Logistic Regression

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.
- Eg. Algorithms: Linear Regression, Logistic Regression
- Advantages:

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.
- Eg. Algorithms: Linear Regression, Logistic Regression
- Advantages:
  - Interpretability: In most cases (except Deep learning), we can look at the parameter values and understand the results.

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.
- Eg. Algorithms: Linear Regression, Logistic Regression
- Advantages:
  - Interpretability: In most cases (except Deep learning), we can look at the parameter values and understand the results.
  - Does not need the entire dataset for doing prediction.

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.
- Eg. Algorithms: Linear Regression, Logistic Regression
- Advantages:
  - Interpretability: In most cases (except Deep learning), we can look at the parameter values and understand the results.
  - Does not need the entire dataset for doing prediction.
- Disadvantages:

Parametric Model

- An explicit mathematical model with finite adaptable parameters for modelling the data generating process.
- The goal of the learning algorithm is to find the values of the adaptable parameters so that the model best fits the data.
- Eg. Algorithms: Linear Regression, Logistic Regression
- Advantages:
  - Interpretability: In most cases (except Deep learning), we can look at the parameter values and understand the results.
  - Does not need the entire dataset for doing prediction.
- Disadvantages:
  - The initial choice of the mathematical model constraints the class of functions it can model, called the model-class. If the data generating process is outside the model-class, then it doesn't matter how much ever data is available, the learning algorithm cannot learn the true data generating process.

Table of Contents

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.
- The goal of linear regression is to find the model that best predicts output label, y, given input vector, $\mathbf{x}$, using the dataset $\boldsymbol{X}$

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.
- The goal of linear regression is to find the model that best predicts output label, y, given input vector, $\mathbf{x}$, using the dataset $\boldsymbol{X}$
- Components of the Linear Regression algorithm

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.
- The goal of linear regression is to find the model that best predicts output label, y, given input vector, $\mathbf{x}$, using the dataset $\boldsymbol{X}$
- Components of the Linear Regression algorithm
  - **Data:** $\boldsymbol{X}$

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.
- The goal of linear regression is to find the model that best predicts output label, y, given input vector, $\mathbf{x}$, using the dataset $\boldsymbol{X}$
- Components of the Linear Regression algorithm
  - **Data:** $\boldsymbol{X}$
  - **Model:** $\hat{y} = \mathbf{w}^T\mathbf{x} + w_0 x_0$, where $x_0 = 1$ and w_0 is the y-intercept.

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.
- The goal of linear regression is to find the model that best predicts output label, y, given input vector, $\mathbf{x}$, using the dataset $\boldsymbol{X}$
- Components of the Linear Regression algorithm
  - **Data:** $\boldsymbol{X}$
  - **Model:** $\hat{y} = \mathbf{w}^T\mathbf{x} + w_0x_0$, where $x_0 = 1$ and w_0 is the y-intercept.
  - **Loss Function:** Minimize Mean Squared Error

$$\frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||^2$$

Linear Regression

- Linear Regression is a machine learning algorithm that finds the best line (or the equivalent in higher dimensions) that fits the data.
- The goal of linear regression is to find the model that best predicts output label, y, given input vector, $\mathbf{x}$, using the dataset $\boldsymbol{X}$
- Components of the Linear Regression algorithm
  - **Data:** $\boldsymbol{X}$
  - **Model:** $\hat{y} = \mathbf{w}^T\mathbf{x} + w_0x_0$, where $x_0 = 1$ and w_0 is the y-intercept.
  - **Loss Function:** Minimize Mean Squared Error

$$\frac{1}{N}||\hat{\mathbf{y}} - \mathbf{y}||^2$$

  - **Optimizer:** Closed Form Solution

$$\mathbf{w}^* = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\mathbf{y}$$