What is ML? Data Preparation & How do Machines Learn?

BST.209 Collaborative Data Science in Healthcare

Eptehal Nashnoush

Introduction – What is ML?

Objectives

- Recognize what is meant by machine learning.
- Have an appreciation of the difference between supervised and unsupervised learning.

End of Module Questions

- What is machine learning?
- What is the relationship between machine learning, AI, and statistics?
- What is meant by supervised and unsupervised learning?

Rule-Based Programming

- The idea is applying rules to data to gain insights and make decisions.
  - For example, we learn that human body temperature is ~37 °C (~98.5 °F), and that higher or lower temperatures can be cause for concern.
- In programming we translate this logic into if-else statements.
- See example.

Machine Learning

- In ML the model/framework **learns the insights from the data for itself**. As the **volume and complexity** of data **increases**, so does the value of having models that can generate new rules.
- See example.

Exercise

A. What is the most time-consuming aspect of developing a predictive model, according to the authors?

B. How have "traditional" predictive models dealt with high numbers of predictor variables, according to the authors?

Solutions

A. 80% of effort in building models is in "preprocessing, merging, customizing, and cleaning".

B. Traditional modeling approaches have dealt with complexity by choosing a very limited number of variables to consider.

Statistics, machine learning, and "AI"

Statistics, machine learning, and "AI"

- *Statistics*: A well-established field of mathematics concerned with methods for collecting, analyzing, interpreting and presenting empirical data.
- *Machine learning*: A set of computational methods that learn rules from data, often with the goal of prediction.
- *Deep learning*: A subfield of machine learning that focuses on more complex "artificial neural network" algorithms.
- *Artificial intelligence*: The goal of conferring human-like intelligence to machines. Researchers working on the goal of intelligent machines are now using "Artificial General Intelligence" (A.G.I.) instead of AI

Supervised vs Unsupervised Learning

- Supervised learning is a category of machine learning that involves the use of labelled datasets to train models for classification and prediction. **Target variable is labeled.**
- Unsupervised machine learning, on the other hand attempts to identify meaningful patterns within unlabeled datasets. **Target variable is unlabeled.**

Exercises

A. We have laboratory test data on patients admitted to a critical care unit and we are trying to identify patients with an emerging, rare disease. There are no labels to indicate which patients have the disease, but we believe that the infected patients will have very distinct characteristics. Do we look for a supervised or unsupervised machine learning approach?

B. We would like to predict whether or not patients will respond to a new drug that is under development based on several genetic markers. We have a large corpus of clinical trial data that includes both genetic markers of patients and their response to the new drug. Do we use a supervised or unsupervised approach?

Solutions

A. The prediction targets are not labeled, so an unsupervised learning approach would be appropriate. Our hope is that we will see a unique cluster in the data that pertains to the emerging disease.

B. We have both genetic markers and known outcomes, so in this case supervised learning is appropriate.

Key points

- Machine learning borrows heavily from fields such as statistics and computer science.
- In machine learning, models learn rules from data.
- In supervised learning, the target in our training data is labeled.
- AI has become a synonym for machine learning.
- AGI is the loftier goal of achieving human-like intelligence.

Data Preparation

Objectives

- Explore characteristics of our dataset.
- Partition data into training and test sets.
- Encode categorical values.
- Use scaling to pre-process features.

End of Module Questions

- Why are some common steps in data preparation?
- What is SQL and why is it often needed?
- What do we partition data at the start of a project?
- What is the purpose of setting a random state when partitioning?
- Should we impute missing values before or after partitioning?"

R Demo

		Missing	Overall	ALIVE	EXPIRED
n			235	195	40
gender, n (%)	Female	0	116 (49.4)	101 (51.8)	15 (37.5)
	Male		118 (50.2)	94 (48.2)	24 (60.0)
	Unknown		1 (0.4)		1 (2.5)
age, mean (SD)		9	61.9 (15.5)	60.5 (15.8)	69.3 (11.5)
admissionweight, mean (SD)		5	87.6 (28.0)	88.6 (28.8)	82.3 (23.3)
length of stay, mean (SD)		0	9.2 (8.6)	9.6 (7.5)	6.9 (12.5)
acutephysiologyscore, mean (SD)		0	59.9 (28.1)	54.5 (23.1)	86.7 (34.7)
apachescore, mean (SD)		0	71.2 (30.3)	64.6 (24.5)	103.5 (34.9)
heartrate, mean (SD)		0	108.7 (33.1)	107.9 (30.6)	112.9 (43.2)
mean blood pressure, mean (SD)		0	93.2 (47.0)	92.1 (45.4)	98.6 (54.5)
creatinine, mean (SD)		0	1.0 (1.7)	0.9 (1.7)	1.7 (1.6)
temperature, mean (SD)		0	35.2 (6.5)	36.1 (3.9)	31.2 (12.4)
respiratoryrate, mean (SD)		0	30.7 (15.2)	29.9 (15.1)	34.3 (15.6)
white cell count, mean (SD)		0	10.5 (8.4)	10.7 (8.2)	9.7 (9.7)
admissionheight, mean (SD)		2	168.0 (12.8)	167.7 (13.4)	169.4 (9.1)

Exercises

A. What is the approximate percent mortality in the eICU cohort?

B. Which variables appear noticeably different in the "Alive" and "Expired" groups?

C. How does the in-hospital mortality differ between the eICU cohort and the ones in Rajkomar et al?

Solutions

A. Approximately 17% (40/235)
B. Several variables differ, including age, length of stay, acute physiology score, heart rate, etc.
C. The Rajkomar et al dataset has significantly lower in-hospital mortality (~2% vs 17%).

Encoding

- What is encoding? Many ML models do not work with categorical variables; therefore, we use label encoding to convert the categorical values to numerical representations.
- e.g., In hospital mortality column entries “ALIVE” or “EXPIRED”
  - Other examples of categorical variables:
    - Blood Type: Categories could include A, B, AB, and O
    - Gender
- One hot encoding - categorical variables are represented as a binary columns for each category.
- See example.

Splitting the Data – Partitioning

- In ML data is split into a training set and "held-out" test set.
- The training set is used for building the model and the test set is used to evaluate model performance on new **unseen data**.
- A split of ~70% training, 30% test is common.
- See example.

How to Deal with Missing Data?

Missing Data

- Imputing (estimating missing values) – In our case, we will take a simple approach of replacing missing values in numerical columns with the median.
  - ****With physiological data, imputing the median typically implies that the missing observation is not a cause for concern.**
- To avoid data leaking between our training and test sets, we take the median from the training set only. **The training median is then used to impute missing values in the held-out test set.**
- See example.

Normalization (known as feature scaling)

- Scaling variables such that they span consistent ranges.
- This can be important, particularly for **models that rely on distance-based** optimization metrics such as kNN. However, it's **less crucial** for models like **Decision Trees**, which do not depend on the scale of the features.
- There are plenty of ways to scale features between zero and one. E.g., Min-Max normalization.
- See example.

Key points

- Data pre-processing is arguably the most important task in machine learning.
- SQL is the tool that we use to extract data from database systems.
- Data is typically partitioned into training and test sets.
- Setting random states helps to promote reproducibility.

End of Module Questions

- What are some common steps in data preparation?
- What is SQL and why is it often needed?
- What do we partition data at the start of a project?
- What is the purpose of setting a random state when partitioning?
- Should we impute missing values before or after partitioning?

The learning part in ML

Objectives

- Understand the importance of quantifying error.
- Code a linear regression model that takes inputs, weights, and bias.
- Code a loss function that quantifies model error.

End of Module Questions

- How do machines learn?
- How can machine learning help us to make predictions?
- Why is it important to be able to quantify the error in our models?
- What is an example of a loss function?

How do machines learn?

- Inspired by the way humans learn (learning rules through trial and error). Machines aren't that different! They go through an iterative learning process to improve their performance.
- In machine learning, we often speak about how a model 'fits' the data. This essentially means how well our model's predictions align with the actual outcomes it's trying to predict. To accomplish this, our model has numerous adjustable parameters that it can tweak and optimize.

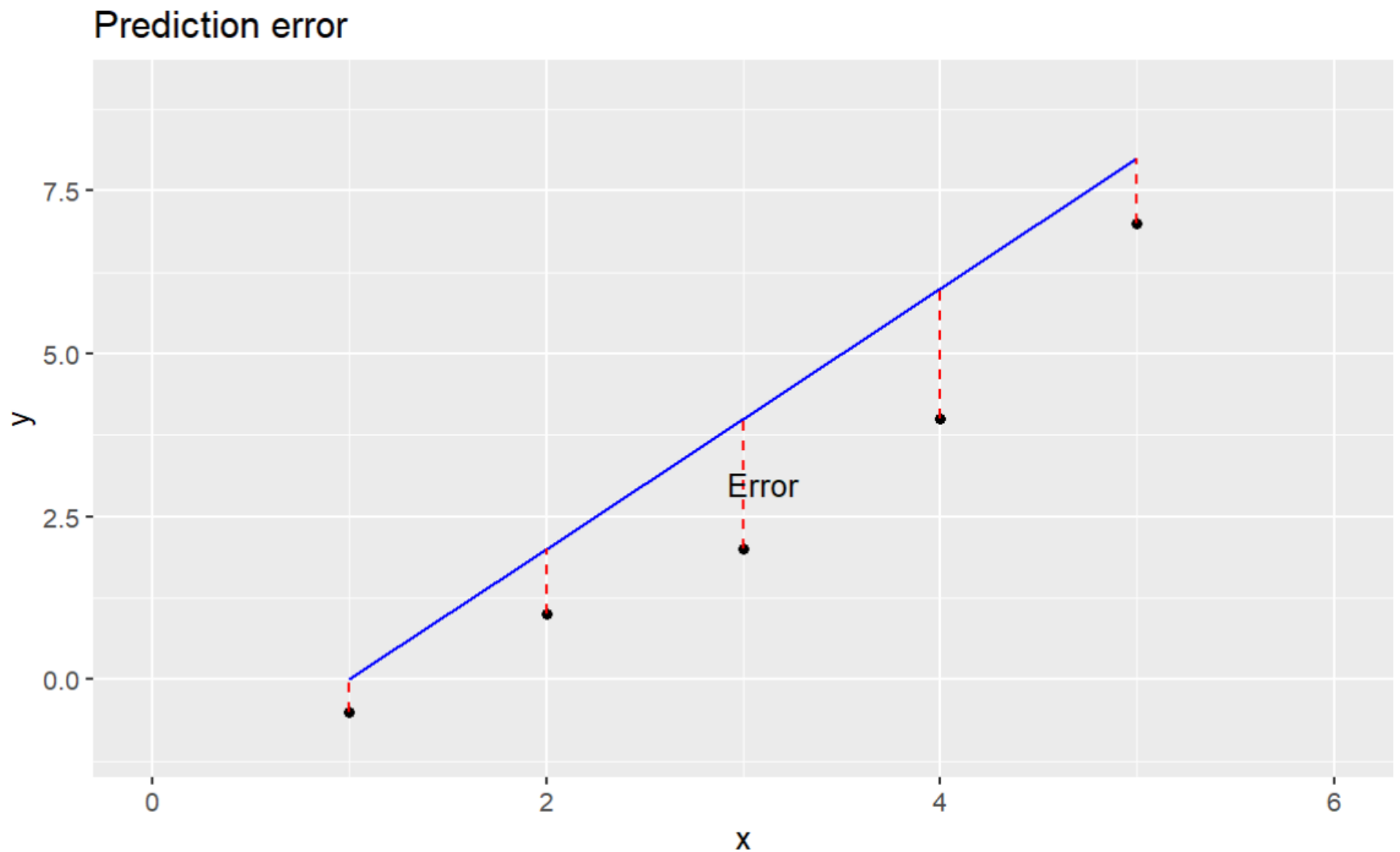
Data Generating Process

- 'data generating process' – a process that's responsible for creating the data we're analyzing. This process could be anything from patient health outcomes in a hospital to transaction patterns in an e-commerce database.
- Understanding this data generating process is crucial because it gives us insights into the underlying patterns and structures in our data. This understanding informs our choice of model, our feature engineering strategies, and even our understanding of what kind of errors our model might make.
- Our goal is to adjust these parameters such that our model best captures the underlying data generating process. This optimization process, often carried out via methods like gradient descent, is at the heart of machine learning.

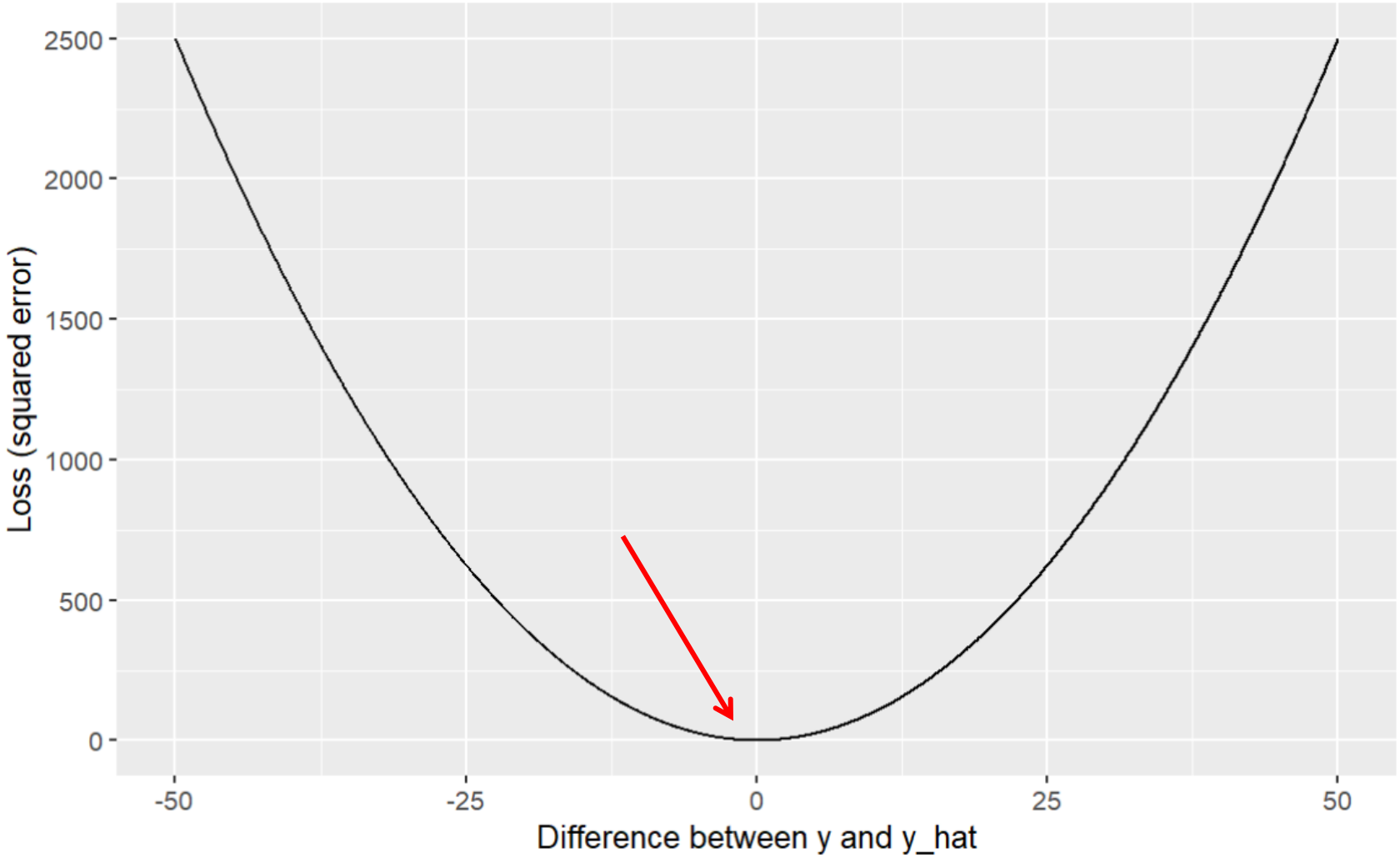
Loss Function

- We need to have some way of quantifying the difference between a “good” model vs a bad” model.
- They allow us to quantify how closely our predictions fit to the known target values.
- In regression models, Mean Squared Error (MSE) is a common example of a loss function. For each prediction, we measure the distance between the known target value Y and our prediction Ŷ, then we take the square.
- See demo.

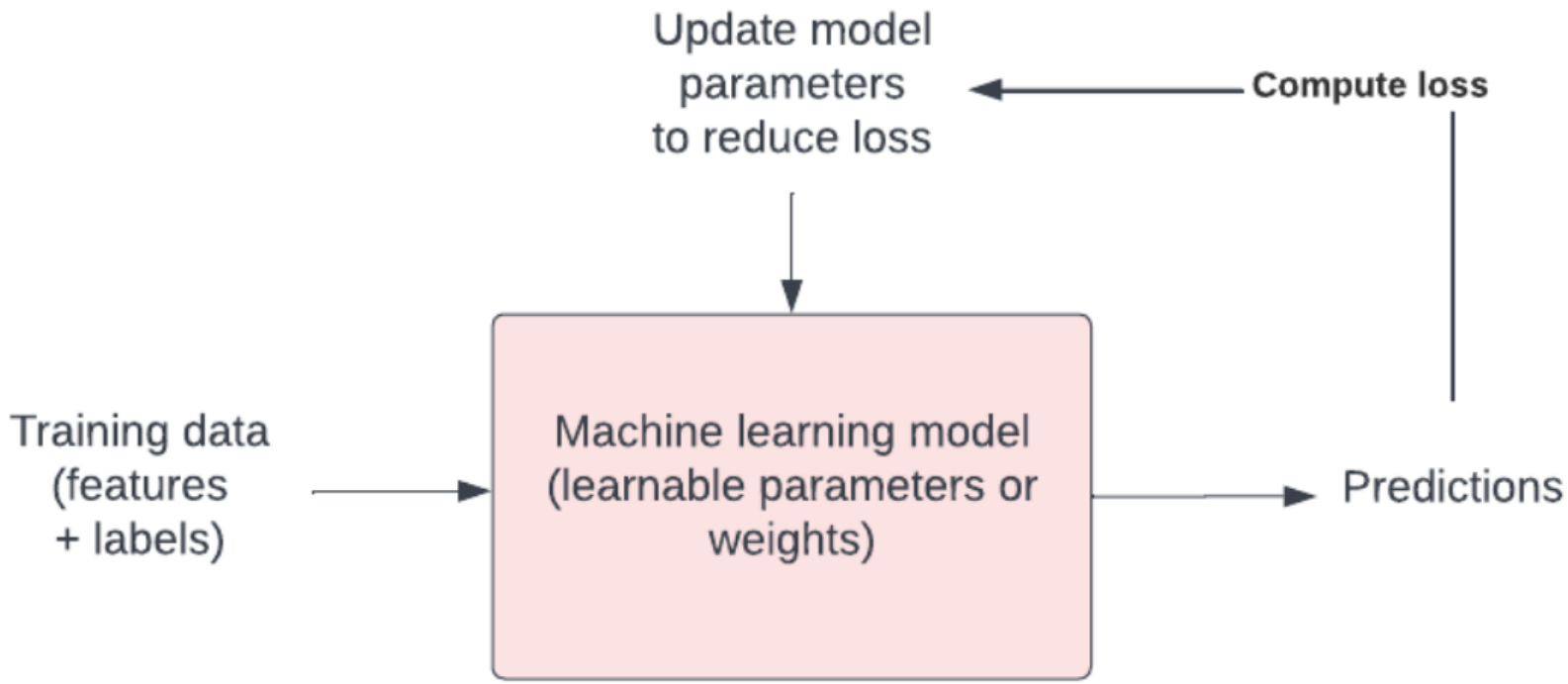
$$mse = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$


Prediction error
7.5
5.0
2.5
0.0
y
Error
0
2
4
6
x

Minimizing the Error | $argmin\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

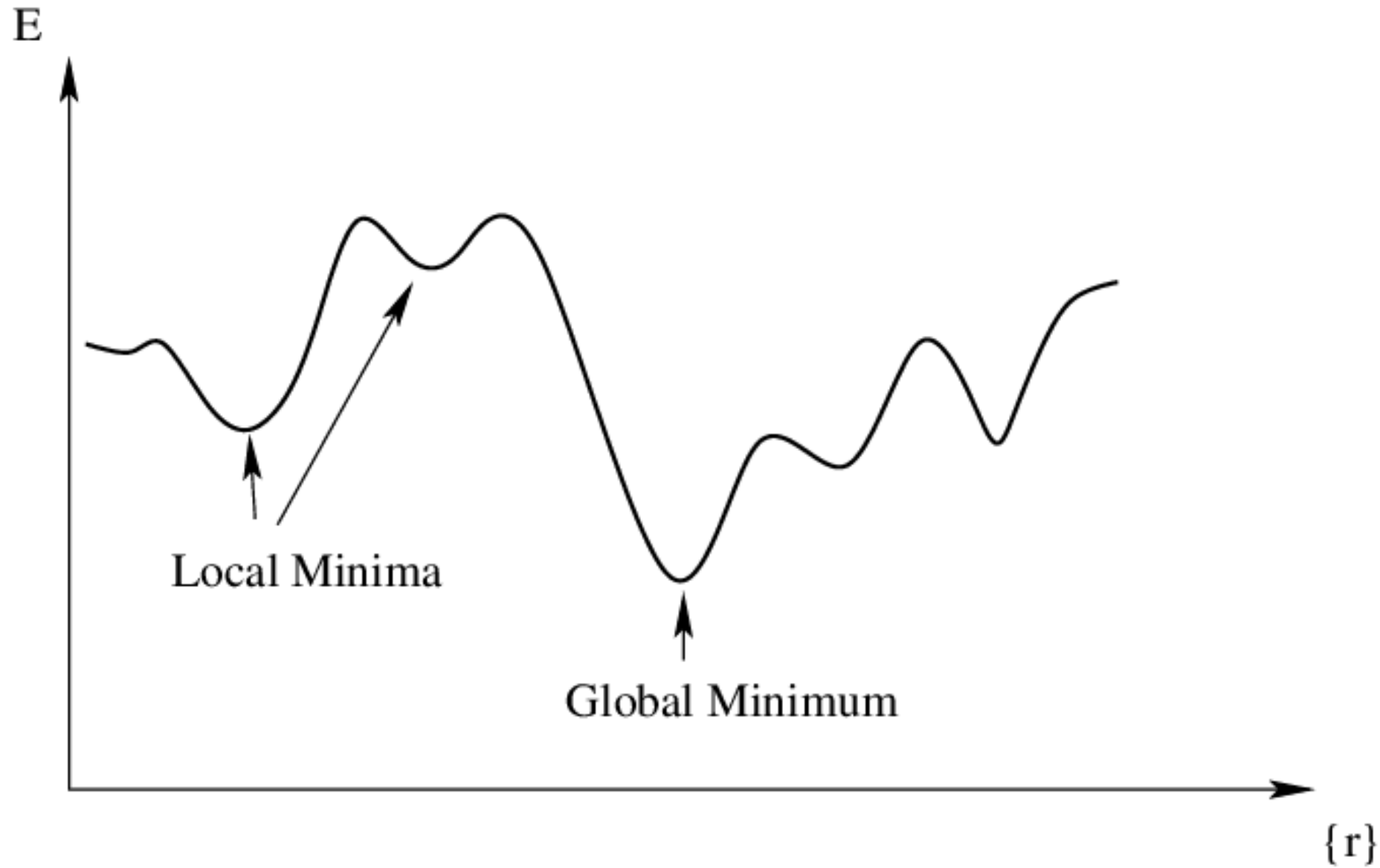


Optimization



Gradient Descent - Finding the ✨global minimum✨

- Gradient Descent is a popular optimization method used in machine learning, particularly in training deep learning models.
- It works by taking iterative steps in the opposite direction of the function's gradient, aiming to reach the point where the function attains its minimum value (i.e., the lowest error).
- In the context of machine learning, the 'function' is usually a loss function and the 'gradient' refers to the derivative of this loss function.


E
Local Minima
Global Minimum
{r}

Key points

- Loss functions allow us to define a good model.
- Y is a known target. $\hat{Y}$ is a prediction.
- Mean squared error is an example of a loss function.
- After defining a loss function, we search for the optimal solution (global minimum) in a process known as 'training’.
- Optimization is at the heart of machine learning.

Exercise

A. What does a loss function quantify?

B. What is an example of a loss function?

C. What are some other names used for loss functions?

D. What is happening when a model is trained?

Solutions

A. A loss function quantifies the goodness of fit of a model (i.e. how closely its predictions match the known targets).

B. One example of a loss function is mean squared error (MSE), RMSE, MAE, etc.

C. Objective function, error function, and cost function.

D. When a model is trained, we are attempting to find the optimal model parameters in process known as "optimization".

End of Module Questions

- How do machines learn?
- How can machine learning help us make predictions?
- Why is it important to be able to quantify the error in our models?
- What is an example of a loss function?