DEEP LEARNING

JOHN D. KELLEHER

The MIT Press Essential Knowledge Series

A complete list of the titles in this series appears at the back of this book.

The MIT Press | Cambridge, Massachusetts | London, England

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030

© 2019 The Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Chaparral Pro by Toppan Best-set Premedia Limited.
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Names: Kelleher, John D., 1974- author.
Title: Deep learning / John D. Kelleher.
Description: Cambridge, MA : The MIT Press, [2019] | Series: The MIT press essential knowledge series | Includes bibliographical references and index.
Identifiers: LCCN 2018059550 | ISBN 9780262537551 (pbk. : alk. paper)
Subjects: LCSH: Machine learning. | Artificial intelligence.
Classification: LCC Q325.5 .K454 2019 | DDC 006.3/1—dc23 LC record available at https://lccn.loc.gov/2018059550

10 9 8 7 6 5

CONTENTS

Series Foreword vii
Preface ix
Acknowledgments xi

1 Introduction to Deep Learning 1
2 Conceptual Foundations 39
3 Neural Networks: The Building Blocks of Deep Learning 65
4 A Brief History of Deep Learning 101
5 Convolutional and Recurrent Neural Networks 159
6 Learning Functions 185
7 The Future of Deep Learning 231

Glossary 251
Notes 257
References 261
Further Readings 267
Index 269

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030

1

INTRODUCTION TO DEEP LEARNING

Deep learning is the subfield of artificial intelligence that focuses on creating large neural network models that are capable of making accurate *data-driven decisions*. Deep learning is particularly suited to contexts where the data is complex and where there are large datasets available. Today most online companies and high-end consumer technologies use deep learning. Among other things, Facebook uses deep learning to analyze text in online conversations. Google, Baidu, and Microsoft all use deep learning for image search, and also for machine translation. All modern smart phones have deep learning systems running on them; for example, deep learning is now the standard technology for speech recognition, and also for face detection on digital cameras. In the healthcare sector, deep learning is used to process medical images (X-rays, CT, and MRI scans) and diagnose health conditions. Deep learning is also at the core of self-driving cars, where it is used for

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030

and each neuron implements a function: it takes a number of values as input and maps them to an output value. The arrows in the network show how the outputs of each neuron are passed as inputs to other neurons. In this network, information flows from left to right. For example, if this network were trained to predict a person's credit solvency, based on their income and debt, it would receive the income and debt as inputs on the left of the network and output the credit solvency score through the neuron on the right.

A neural network uses a divide-and-conquer strategy to learn a function: each neuron in the network learns a simple function, and the overall (more complex) function, defined by the network, is created by combining these simpler functions. Chapter 3 will describe how a neural network processes information.

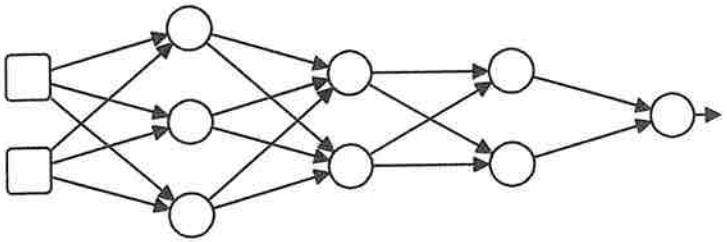

Figure 1.2 Schematic illustration of a neural network.

What Is Machine Learning?

A machine learning algorithm is a search process designed to choose the best function, from a set of possible functions, to explain the relationships between features in a dataset. To get an intuitive understanding of what is involved in extracting, or learning, a function from data, examine the following set of sample inputs to an unknown function and the outputs it returns. Given these examples, decide which arithmetic operation (addition, subtraction, multiplication, or division) is the best choice to explain the mapping the unknown function defines between its inputs and output:

function(Inputs) = Output

function(5,5) = 25

function(2,6) = 12

function(4,4) = 16

function(2,2) = 04

Most people would agree that multiplication is the best choice because it provides the best match to the observed relationship, or mapping, from the inputs to the outputs:

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030

$5 \times 5 = 25$

$2 \times 6 = 12$

$4 \times 7 = 28$

$2 \times 2 = 04$

In this particular instance, choosing the best function is relatively straightforward, and a human can do it without the aid of a computer. However, as the number of inputs to the unknown function increases (perhaps to hundreds or thousands of inputs), and the variety of potential functions to be considered gets larger, the task becomes much more difficult. It is in these contexts that harnessing the power of machine learning to search for the best function, to match the patterns in the dataset, becomes necessary.

Machine learning involves a two-step process: training and inference. During training, a machine learning algorithm processes a dataset and chooses the function that best matches the patterns in the data. The extracted function will be encoded in a computer program in a particular form (such as if-then-else rules or parameters of a specified equation). The encoded function is known as a model, and the analysis of the data in order to extract the function is often referred to as training the model.

Essentially, models are functions encoded as computer programs. However, in machine learning the concepts of function and model are so closely related that the distinction is often skipped over and the terms may even be used interchangeably.

In the context of deep learning, the relationship between functions and models is that the function extracted from a dataset during training is represented as a neural network model, and conversely a neural network model encodes a function as a computer program. The standard process used to train a neural network is to begin training with a neural network where the parameters of the network are randomly initialized (we will explain network parameters later; for now just think of them as values that control how the function the network encodes works). This randomly initialized network will be very inaccurate in terms of its ability to match the relationship between the various input values and target outputs for the examples in the dataset. The training process then proceeds by iterating through the examples in the dataset, and, for each example, presenting the input values to the network and then using the difference between the output returned by the network and the correct output for the example listed in the dataset to update the network's parameters so that it matches the data more closely. Once the machine learning algorithm has found a function that is sufficiently accurate (in terms of the outputs it generates

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030

matching the correct outputs listed in the dataset) for the problem we are trying to solve, the training process is completed, and the final model is returned by the algorithm. This is the point at which the learning in machine learning stops.

Once training has finished, the model is fixed. The second stage in machine learning is inference. This is when the model is applied to new examples—examples for which we do not know the correct output value, and therefore we want the model to generate estimates of this value for us. Most of the work in machine learning is focused on how to train accurate models (i.e., extracting an accurate function from data). This is because the skills and methods required to deploy a trained machine learning model into production, in order to do inference on new examples at scale, are different from those that a typical data scientist will possess. There is a growing recognition within the industry of the distinctive skills needed to deploy artificial intelligence systems at scale, and this is reflected in a growing interest in the field known as DevOps, a term describing the need for collaboration between development and operations teams (the operations team being the team responsible for deploying a developed system into production and ensuring that these systems are stable and scalable). The terms MLOps, for machine learning operations, and AIOps, for artificial intelligence operations, are also used to describe the challenges of deploying a trained model. The questions around model deployment are beyond the scope of this book, so we will instead focus on describing what deep learning is, what it can be used for, how it has evolved, and how we can train accurate deep learning models.

One relevant question here is: why is extracting a function from data useful? The reason is that once a function has been extracted from a dataset it can be applied to unseen data, and the values returned by the function in response to these new inputs can provide insight into the correct decisions for these new problems (i.e., it can be used for inference). Recall that a function is simply a deterministic mapping from inputs to outputs. The simplicity of this definition, however, hides the variety that exists within the set of functions. Consider the following examples:

- Spam filtering is a function that takes an email as input and returns a value that classifies the email as spam (or not).
- Face recognition is a function that takes an image as input and returns a labeling of the pixels in the image that demarcates the face in the image.
- Gene prediction is a function that takes a genomic DNA sequence as input and returns the regions of the DNA that encode a gene.

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030

- Speech recognition is a function that takes an audio speech signal as input and returns a textual transcription of the speech.
- Machine translation is a function that takes a sentence in one language as input and returns the translation of that sentence in another language.

It is because the solutions to so many problems across so many domains can be framed as functions that machine learning has become so important in recent years.

Why Is Machine Learning Difficult?

There are a number of factors that make the machine learning task difficult, even with the help of a computer. First, most datasets will include noise[3] in the data, so searching for a function that matches the data exactly is not necessarily the best strategy to follow, as it is equivalent to learning the noise. Second, it is often the case that the set of possible functions is larger than the set of examples in the dataset. This means that machine learning is an ill-posed problem: the information given in the problem is not sufficient to find a *single* best solution; instead multiple possible solutions will match the data. We can use the problem of selecting the arithmetic operation (addition, subtraction, multiplication, or division) that best matches a set of example input-output mappings for an unknown function to illustrate the concept of an ill-posed problem. Here are the example mappings for this function selection problem:

$$\text{function}(\text{Inputs}) = \text{Output}$$

$$\text{function}(1,1) = 1$$

$$\text{function}(2,1) = 2$$

$$\text{function}(3,1) = 3$$

Given these examples, multiplication and division are better matches for the unknown function than addition and subtraction. However, it is not possible to decide whether the unknown function is actually multiplication or division using this sample of data, because both operations are consistent with all the examples provided. Consequently, this is an ill-posed problem: it is not possible to select a single best answer given the information provided in the problem.

One strategy to solve an ill-posed problem is to collect more data (more examples) in the hope that the new examples will help us to discriminate between the correct underlying function and the remaining alternatives. Frequently, however, this strategy is not feasible, either

MIM Software Inc. v. Progenics Pharmaceuticals, Inc.
IPR2025-00630
Progenics EX2030