

Machine Learning with PyTorch and Scikit-Learn

From Coursera

Cristo Daniel Alvarado

January 26, 2026

Contents

1 Module 1	6
1.1 Overview	6
1.2 Giving computers the ability to learn from data	6
1.3 Introduction	6
1.3.1 The Three Different Types of Machine Learning	7
1.3.2 Supervised Learning	7
1.3.3 Classification for Predicting Class Labels	8
1.3.4 Regression for Predicting Continuous Outcomes	9
1.4 Solving Problems with Reinforcement Learning	10
1.5 Unsupervised Learning	11
1.5.1 Finding Subgroups With Clustering	11
1.5.2 Dimensionality Reduction for Data Compression	11
1.6 Terminology and Notations	12
1.6.1 Machine Learning Terminology	14
1.7 Machine Learning Systems	14
1.7.1 Preprocessing: Getting Data into Shape	14
1.7.2 Training and Selecting a Predictive Model	16
1.7.3 Evaluating Models and Predicting Unseen Data Instances	17
2 Training Simple Machine Learning Algorithms for Classification	18
2.1 Introduction	18
2.2 Artificial Neurons: Early Stages of Machine Learning	18
2.3 Formal Definition of an Artificial Neuron	19
2.4 Perceptron Learning Rule	20
2.5 Implementing a Perceptron Learning Algorithm in Python	22
2.6 An Object-Oriented Perceptron API	22
2.7 Training of the Perceptron Algorithm on the Iris Dataset	26
2.7.1 OvA method for multi-class classification	26
2.7.2 Training the Algorithm	28
2.8 Adaptive Linear Neurons and the Convergence of Learning	32
2.8.1 Main Differences Between Perceptron and Adaline	32

2.8.2	Minimizing Loss Function with Gradient Descent	32
2.9	Implementing Adaline in Python	35
2.9.1	Hyperparameters	38
2.10	Improving Gradient Descent Through Feature Scaling	38
2.10.1	Large-Scale Machine Learning and Stochastic Gradient Descent	40
3	A Tour of Machine Learning Classifiers Using Scikit-Learn	45
3.1	Introduction	45
3.1.1	Choosing a Classification Algorithm	45
3.1.2	First Steps with scikit-learn Training a Perceptron	46
3.1.3	Training a Model Using Scikit-Learn	46
3.2	Modeling Class Probabilities Via Logistic Regression	51
3.2.1	Logistic Regression and Conditional Probabilities	51
3.3	Learning the Model Weights via the Logistic Loss Function	54
3.3.1	Deriving the likelihood function	54
3.4	Converting an Adaline Implementation Into an Algorithm for Logistic Regression	56
3.5	Gradient Descent	58
3.6	Training a Logistic Regression Model with Scikit-Learn	58
3.6.1	Algorithms for convex optimization	61
3.7	Tackling Overfitting via Regularization	62
3.7.1	The Bias-Variance Tradeoff	62
3.7.2	Implementation of Regularization	64
3.8	Maximum Margin Classification with Support Vector Machines	65
3.8.1	Dealing with a Nonlinearly Separable Case Using Slack Variables	67
3.8.2	Training a Model with a Support Vector Machine	67
3.8.3	Alternative Implementations in Scikit-learn	69
3.9	Solving Nonlinear Problems Using a Kernel SVM	70
3.9.1	Kernel Methods for Linearly Inseparable Data	70
3.9.2	Using the Kernel Trick to Find Separating Hyperplanes in a High-Dimensional Space	71
3.10	Decision Tree Learning	78
3.10.1	Maximizing IG Getting the Most Bang for Your Buck	78
3.11	Building a Decision Tree	84
3.11.1	Combining Multiple Decision Trees via Random Forests	87
3.12	<i>K</i> -Nearest Neighbours	89
3.12.1	Advantages and disadvantages of memory-based approaches	90
4	Building Good Trainign Datasets: Data Preprocessing	94
4.1	Introduction	94
4.2	Dealing With Missing Data	94

4.3 Identifying Missing Values in Tabular Data	94
4.3.1 Eliminating Training Examples or Features with Missing Values	96
A Using Python for Machine Learning	97
A.1 Basic Configuration	97
A.2 Anaconda Python Distribution and Package Manager	98
A.3 Packages for Scientific Computing, Data Science, and Machine Learning	98
B A Note on Distributions	100
B.1 Normal Distribution	100
B.2 Bernoulli's Distribution	100

Code List

2.1	Python Perceptron Rule Implementation.	23
2.2	Read Dataset from UCI.	27
2.3	Extraction and Visualization of Iris Dataset.	27
2.4	Training the Perceptron Algorithm with the Iris Dataset	28
2.5	Plotting Regions on the Iris Dataset	29
2.6	Plotting the Perceptron Info on the Iris Dataset	30
2.7	Adaline Implementation in Python.	36
2.8	Computation of Partial Derivatives	37
2.9	Weight Update Optimization.	37
2.10	Loss for Two Different Learning Rates.	38
2.11	Standarization Using Numpy	40
2.12	Adaline With Standarization Procedure	40
2.13	Adaline Implemented Using Stochastic Gradient Descent	42
3.1	Class Labels of Matrix <code>X</code>	46
3.2	Train Test Split Function	47
3.3	Use of <code>bincount</code> Method in Python.	47
3.4	<code>StandardScaler</code> Method in Python.	47
3.5	Perceptron Trained Model using Sklearn	48
3.6	Sigmoid Function Graph.	52
3.7	Implementation of Logistic Regression in Python.	56
3.8	Logistic Regression Iris Dataset Python.	59
3.9	Predicting Probabilities.	61
3.10	Logistic Regression Regularized with different values of <code>c</code> on the Iris Dataset Python.	64
3.11	Implementation of SVM on Python and Graph of Regions.	67
3.12	caption	70
3.13	<code>XOR</code> in Python.	70
3.14	Decision Boundary using RBF Kernel on the XOR Dataset.	73
3.15	Decision Boundary on the Iris Dataset using RBF.	75
3.16	Entropy Computation.	80
3.17	Entropy Computation.	82
3.18	Code for Decision Tree Implementation.	85

3.19	caption	86
3.20	Code for <i>k</i> -Nearest Neighbours.	90
4.1	CSV with Missing Data.	94
4.2	CSV with Missing Data Output	95
4.3	Check Missing Data per Row	95
4.4	Output Missing Data Per Row	95
4.5	Dataframe <code>df</code>	95
4.6	Values of Dataframe <code>df</code>	96
4.7	Dropping Rows from <code>df</code>	96
4.8	Dropping Columns from a <code>DataFrame df</code>	96
A.1	Check Python Version	97
A.2	<code>pip</code> Package Installation.	97
A.3	Conda Package Installation.	98
A.4	caption	98
A.5	Check Version of Packages	99

CHAPTER 1

MODULE 1

The goal of this chapter is to explore the foundational concepts of machine learning, focusing on how algorithms can transform data into knowledge. We delve into the practical applications of supervised and unsupervised learning, equipping you with the skills to implement these techniques using Python tools for effective data analysis and prediction.

Observation 1.0.1 (Learning Objectives)

Learning Objectives:

- Analyze data patterns to make future predictions.
- Design systems for supervised and unsupervised learning.
- Implement machine learning algorithms using Python tools.

1.1 OVERVIEW

The goal basically is to learn machine learning using Python libraries such as Scikit-Learn and PyTorch.

1.2 GIVING COMPUTERS THE ABILITY TO LEARN FROM DATA

So, basically in this scenario we will do the following:

- Implement machine learning algorithms using Python tools.
- Design systems for supervised and unsupervised learning.
- Analyze data patterns to make future predictions.

1.3 INTRODUCTION

Definition 1.3.1 (Machine Learning)

Machine learning is the *application and science of algorithms that make sense of data*.

The goal of this section is to introduce some of the main concepts and different types of machine learning, together with a basic introduction to the relevant terminology. The goal is to establish the groundwork for successfully using machine learning techniques for practical problem solving.

1.3.1 THE THREE DIFFERENT TYPES OF MACHINE LEARNING

In this age, we have a large amount of structured and non-structured data.

Definition 1.3.2 (Structured and Non-structured Data)

Structured data refers to information that is organized in a predefined format, typically within a fixed schema, such as rows and columns in a database or spreadsheet

In contrast, **non-structured data** lacks a predefined format or structure and exists in its native, raw state. It is typically qualitative and encompasses a wide variety of formats such as text documents, emails, social media posts, images, videos, and audio files.

One of the main goals of machine learning is to extract knowledge from data in order to make predictions.

Observation 1.3.1 (Use of Machine Learning)

Machine Learning offers a more efficient alternative for capturing the knowledge in data to improve the performance of predictive models and make data-driven decisions.

Idea 1.3.1 (Note on the use of Machine Learning and some of its Applications)

Also, notable progress has been made in medical applications; for example, researchers demonstrated that deep learning models can detect skin cancer with near-human accuracy [link to article](#).

Another milestone was recently achieved by researchers at DeepMind, who used deep learning to predict 3D protein structures, outperforming physics-based approaches by a substantial margin [link to article](#).

There are three types of machine learning: **supervised learning**, **unsupervised learning**, and **reinforcement learning**. There are fundamental differences between these three types of machine learning and we will look at each of them in detail with some notes on its possible applications.

Three Types of Machine Learning	
Supervised learning	<ul style="list-style-type: none">• Labeled data• Direct feedback• Predict outcome/future
Unsupervised learning	<ul style="list-style-type: none">• No labels/targets• No feedback• Find hidden structure in data
Reinforcement learning	<ul style="list-style-type: none">• Decision process• Reward system• Learn series of actions

Table 1.1: Three Types of Machine Learning.

1.3.2 SUPERVISED LEARNING

The main goal of **supervised learning** is to learn a model from labeled training data that allows us to make predictions about unseen or future data.

The term *supervised* refers to a set of training examples (data inputs), where the desired output signals (labels) are already known.

Definition 1.3.3 (Supervised Learning)

Supervised learning is the *process of modeling the relationship between data inputs and the labels*.

We can think of supervised learning as *label learning*.

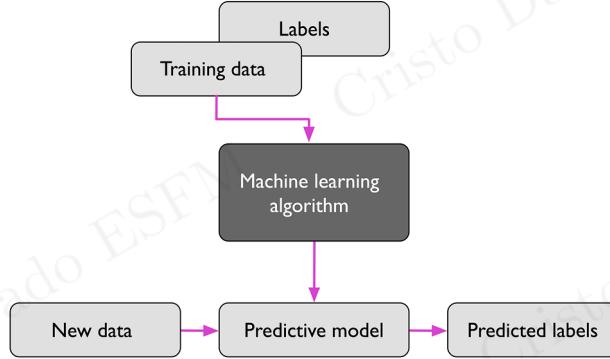


Figure 1.1: General Process of Supervised Learning

Example 1.3.1 (Example of Supervised Learning)

Lets consider the example of email spam filtering. We can train a model using a supervised machine learning algorithm on a corpus of labeled emails, which are correctly marked as spam or non-spam.

To predict whether a new email belongs to either of the two categories, a supervised learning task with discrete class labels, such as in the previous email spam filtering example.

This is called a **classification task**. Another subcategory of supervised learning is **regression**, where the outcome signal is a continuous value.

1.3.3 CLASSIFICATION FOR PREDICTING CLASS LABELS

Definition 1.3.4 (Classification)

Classification is a subcategory of supervised learning, where the goal is to predict the categorical class labels of new instances or data points based on past observations.

Those class labels are discrete, unordered values that can be understood as the group memberships of the data points. The previously mentioned example of email spam detection represents a typical example of a binary classification task, where the machine learning algorithm learns a set of rules to distinguish between two possible classes: spam and no-spam emails.

The next illustration shows the concept of binary classification task given 30 training examples; 15 training are labeled as **class A** and the other 15 labeled as **class B**.

Our dataset is two-dimensional, meaning that each example has two values associated with it: x_1 and x_2 . We can use a supervised machine learning algorithm to learn a rule (decision boundary represented as the dashed line) that can separate two classes and classify new data into each of those two categories given its x_1 and x_2 values.

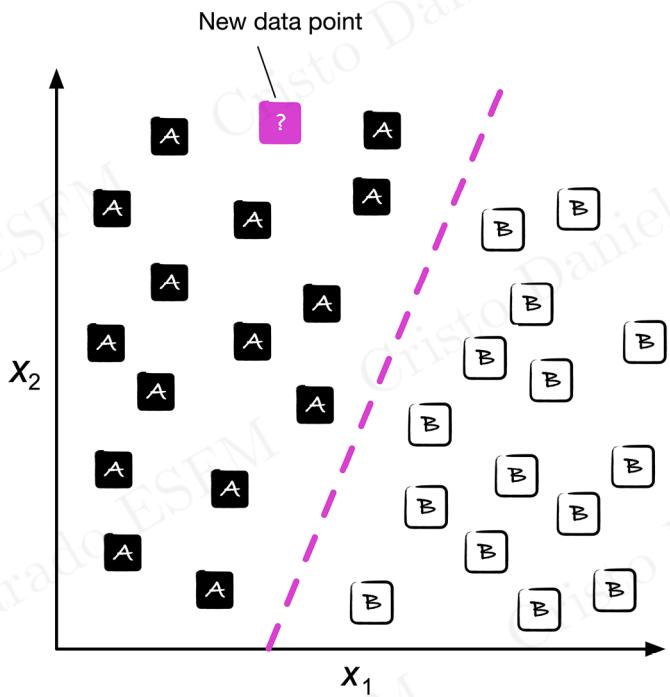


Figure 1.2: Classification in Machine Learning

Observation 1.3.2 (Nature of classification of class labels)

When a set of class labels does not have to be of a binary nature. The predictive model learned by a supervised learning algorithm can assign any class label that was presented in the training dataset to a new, unlabeled data point or instance.

Example 1.3.2

An example of a multiclass classification task is handwritten character recognition.

1.3.4 REGRESSION FOR PREDICTING CONTINUOUS OUTCOMES

Definition 1.3.5 (Regression Analysis)

In a **regression analysis**, we are given a number of predictor (**explanatory**) variables and a continuous response variable (**outcome**) and we try to find a relationship between those variables that allows us to predict an outcome.

Observation 1.3.3 (Feature and Target Variables)

In the field of machine learning, the predictor variables are commonly called *features*, and the response variables are referred to as *target variables*.

Example 1.3.3

Let's assume we want to predict the SAT scores of students. If there is a relationship between time spent studying for the test and the final scores, we could use it as training data to learn a model that uses the study time to predict the test scores of future students.

Given a feature variable x and a target variable y , we fit a straight line to the data points that minimizes the error in predicting y from x .

We can now use the intercept and slope learned from this data to predict the target variable of new data.

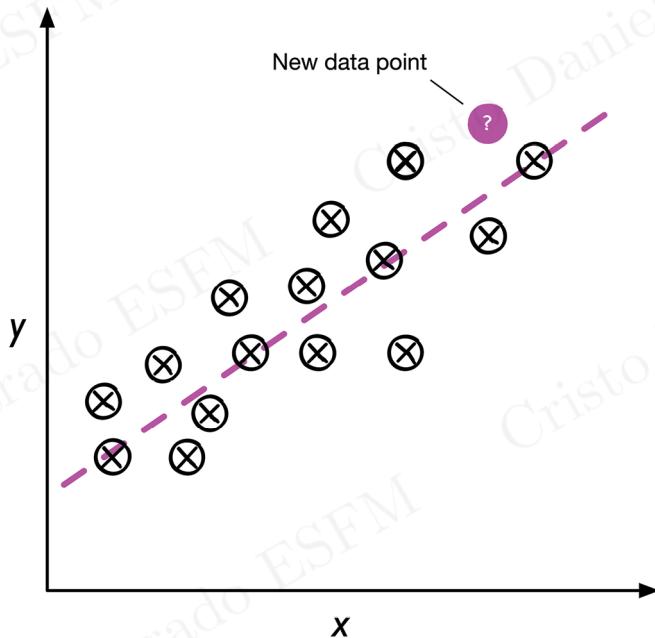


Figure 1.3: Linear Approximation

1.4 SOLVING PROBLEMS WITH REINFORCEMENT LEARNING

Definition 1.4.1 (Reinforcement Learning)

Reinforcement Learning has the *goal to develop a system* (called **agent**) *that improves its performance through interactions with the environment*.

Observation 1.4.1

Typically, the information coming from a current state of the environment also includes a **reward signal**. With this in mind, reinforcement learning can be viewed as being related to supervised learning.

This feedback is not the correct ground-truth label or value, but a measure of how well the action is evaluated by a reward function.

Using a reward signal, a system can measure this reward via an exploratory trial-and-error approach or deliberative planning.

Example 1.4.1

One of the most popular examples of reinforcement learning is **chess program**. The agent decides upon a series of moves and the reward can be defined as win or lose at the end of the game.

In synthesis, the reinforcement learning can be described with the following diagram.

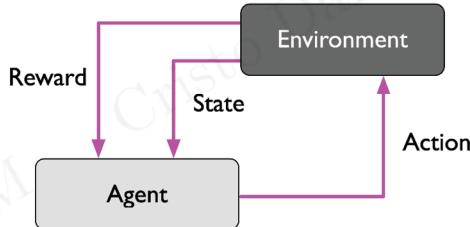


Figure 1.4: Reinforcement Learning Process

There are many different subtypes of reinforcement learning, but in a general scheme the agent tries to maximize the reward through a series of interactions with the environment.

Idea 1.4.1

In summary, reinforcement learning is concerned with learning to choose a series of actions that maximizes the total reward which could be earned immediately after taking an action or via *delayed* feedback.

1.5 UNSUPERVISED LEARNING

Unsupervised learning tries to find the hidden structures of our data. Using supervised learning or reinforcement learning, we have the right answer or a measure of a reward for particular actions, respectively. Unsupervised learning explores the structure of our data to extract meaningful information without the guidance of a known outcome variable.

1.5.1 FINDING SUBGROUPS WITH CLUSTERING

Definition 1.5.1 (Clustering)

Clustering is an exploratory data analysis or pattern discovery technique that organizes information into meaningful subgroups (called **clusters**) without any prior knowledge of group memberships.

Each cluster defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters which is why clustering is called (sometimes) **unsupervised classification**.

The following scatter plot shows how clustering can organize unlabeled data into three distinct groups or clusters.

1.5.2 DIMENSIONALITY REDUCTION FOR DATA COMPRESSION

Another subfield of unsupervised learning is dimensionality reduction.

Definition 1.5.2 (Dimensionality Reduction)

Dimensionality reduction is the process to reduce the dimension of the data we are working with, preserving the original and meaningful data needed to interpret it.

Often, we work with data of high dimensionality that can present challenges for limited storage space and the computational performance of machine learning algorithms. The unsupervised dimensionality reduction is commonly used in feature preprocessing to remove noise from data, which can degrade the predictive performance of certain algorithms.

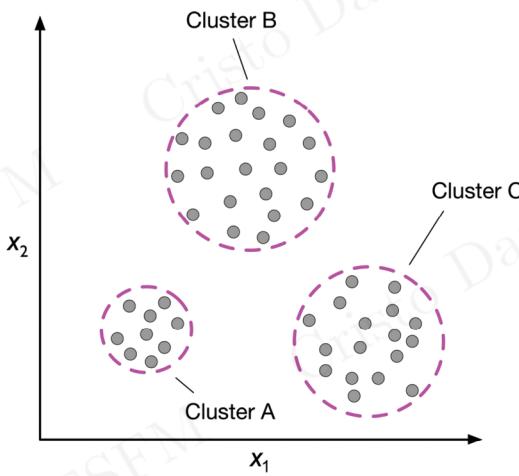


Figure 1.5: Caption

Dimensionality reduction compresses the data into a smaller dimensional subspace while retaining most of the relevant information.

Observation 1.5.1

Sometimes, dimensionality reduction is useful for visualizing data. The following figure is a good and useful example of it.

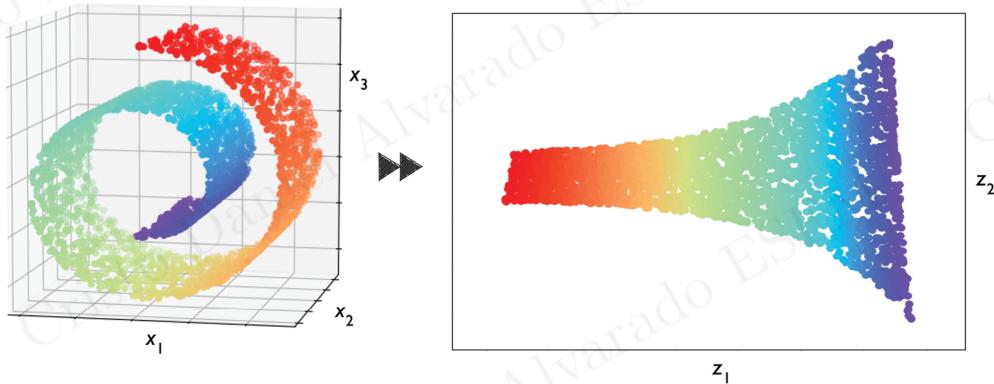


Figure 1.6: Example of Dimensionality Reduction in a Dataset.

1.6 TERMINOLOGY AND NOTATIONS

Definition 1.6.1 (Dataset)

A **dataset** is a *collection of related sets of information that is composed of separate elements but can be manipulated as a unit by a computer*.

One image that contains most of the information that we'll be using throughout this course is the following:

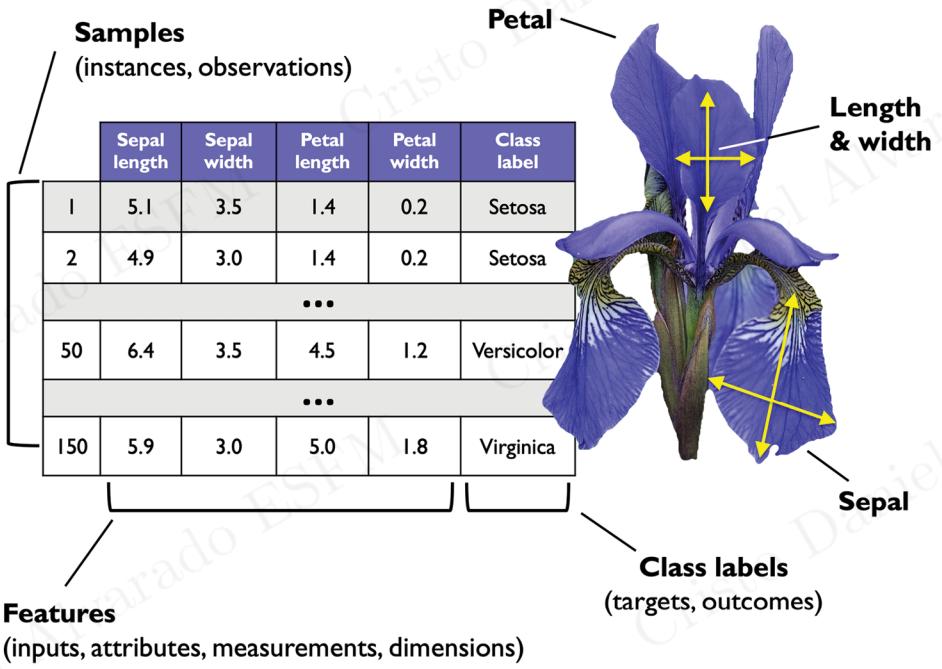


Figure 1.7: Samples, Features, and Class labels.

Example 1.6.1

The Iris dataset contains the measurements of 150 iris flowers from three different species (setosa, versicolor and virginica).

Each flower example represents one row in our dataset and the flower measurements in centimeters are stored as columns, called **features of the dataset**.

Definition 1.6.2 (Feature)

In a dataset, a **feature** is an input, attribute, measurement or dimensions of something.

Idea 1.6.1 (Notation)

The iris dataset consists of 150 examples and four features, we will write them as a 150×4 matrix:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & x_3^{(1)} & x_4^{(1)} \\ x_1^{(2)} & x_2^{(2)} & x_3^{(2)} & x_4^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{(150)} & x_2^{(150)} & x_3^{(150)} & x_4^{(150)} \end{bmatrix}$$

Vectors will be identified as the column matrix $\vec{x} \in \mathbb{R}^{n \times 1}$. To refer to the elements of the matrix we will use the notation $x_i^{(j)}$.

Observation 1.6.1 (Use of Notation)

In the latter idea, $x_1^{(150)}$ refers to the first dimension flower example 150, corresponding to the sepal length.

Each row in the matrix X represents one flower instance and can be written as a four-

dimensional row vector:

$$x^{(i)} = \begin{bmatrix} x_1^{(i)} & x_2^{(i)} & x_3^{(i)} & x_4^{(i)} \end{bmatrix}$$

Each feature is a 150-dimensional column vector, denoted by x_j :

$$x_j = \begin{bmatrix} x_j^{(1)} \\ x_j^{(2)} \\ \vdots \\ x_j^{(150)} \end{bmatrix}$$

We represent the target variables (here, class labels) as 150-dimensional column vector:

$$y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(150)} \end{bmatrix}$$

where $y^{(i)} \in \{\text{Setosa, Versicolor, Virginica}\}$.

1.6.1 MACHINE LEARNING TERMINOLOGY

Definition 1.6.3 (Terminology in Machine Learning)

Some of the following terms are largely used in machine learning:

- **Training example:** A row in a table representing the dataset and synonymous with an observation, record, instance, or sample (in most contexts, sample refers to a collection of training examples).
- **Training:** Model fitting, for parametric models similar to parameter estimation.
- **Feature (x):** A column in a data table or data (design) matrix. Synonymous with predictor, variable, input, attribute, or covariate.
- **Target (y):** Synonymous with outcome, output, response variable, dependent variable, (class) label, and ground truth.
- **Loss function:** Often used synonymously with a cost function. Sometimes the loss function is also called an error function. In some literature, the term 'loss' refers to the loss measured for a single data point, and the cost is a measurement that computes the loss (average or summed) over the entire dataset.

1.7 MACHINE LEARNING SYSTEMS

In this section we'll discuss the other important parts of a machine learning system accompanying the learning algorithm.

1.7.1 PREPROCESSING: GETTING DATA INTO SHAPE

Raw data rarely comes in the form and shape that is necessary for the optimal performance of a learning algorithm. Thus, the preprocessing of the data is one of the most crucial steps in any machine learning application.

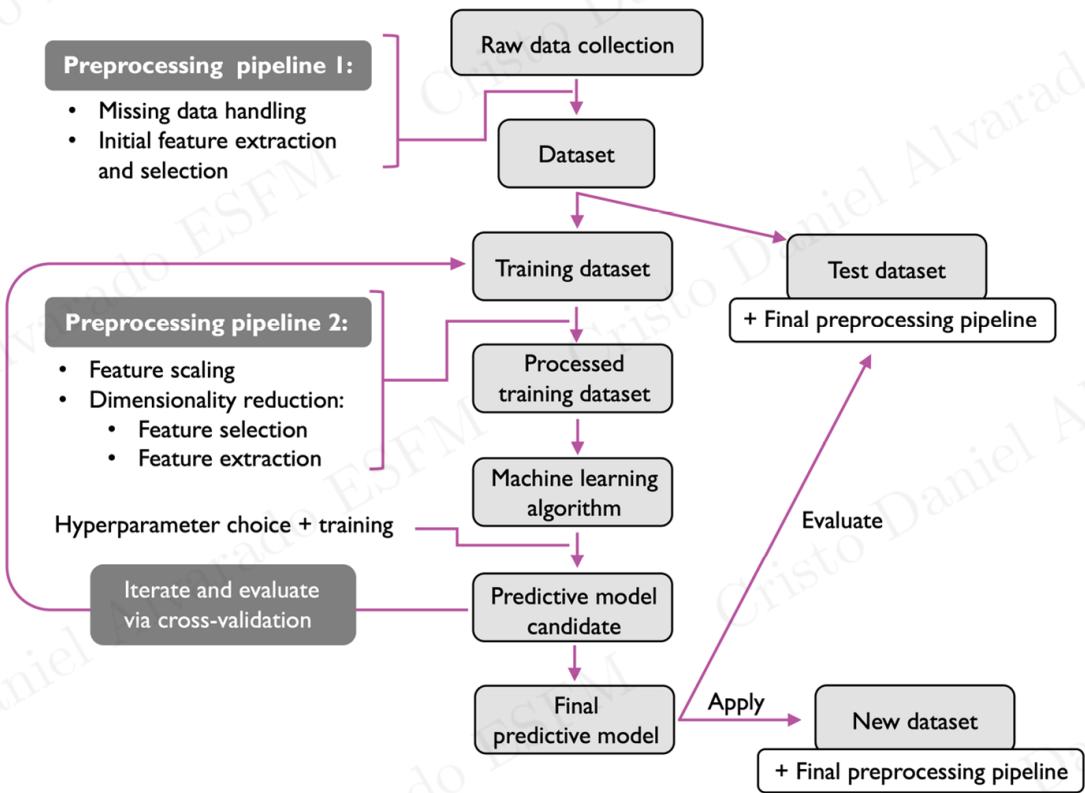


Figure 1.8: Machine Learning System.

Definition 1.7.1 (Preprocessing)

Preprocessing is the process of converting raw data into data suitable for a machine learning algorithm.

Example 1.7.1 (Optimal Features in the Iris Dataset)

Let's consider the iris dataset from the previous example. We can think of raw data as a series of flower images from which we want to extract **meaningful features**. Useful features could be:

- Centered around the color of the flowers.
- Height.
- Length.
- Width.

Observation 1.7.1 (Data Convesion)

Many machine learning models require that the selected features are on the same scale for optimal performance, which is achieved by transforming the features to the range $[0, 1]$.

Other way to achieve this is use standard normal distribution with zero mean and unit variance.

Observation 1.7.2 (Use of Dimensionaly Reduction)

When certain features are highly correlated and therefore, redundant, dimensionaly techniques may be useful for compressing the features onto a lower dimensional subspace.

The main advantage of dimensionality reduction is that less storage space is required, and therefore the learning algorithm can run much faster. This can also improve the performance of a model if the dataset contains a large number of irrelevant features (or noise).

Definition 1.7.2 (Noise)

In data science and statistics, **noise** refers to the *random, uncontrollable, and unexplained variations in the data. It is the part of the data that does not represent the underlying phenomenon you are trying to study or model.*

Observation 1.7.3 (How to Deal with Noise?)

You can rarely eliminate noise completely, but you can manage it.

(1) **Data Cleaning & Preprocessing:** This is the first line of defense.

- **Smoothing:** Applying algorithms to average out random fluctuations in time-series or signal data.
- **Binning:** Grouping numerical values into bins to smooth out small irregularities.
- **Correcting Typos:** Standardizing text entries.

(2) **Collecting More Data:** A larger dataset can sometimes "drown out" the noise, as the true signal becomes stronger with more examples.

(3) **Using Robust Algorithms:** Some machine learning models (like Random Forests) are inherently more resistant to noise than others (like Decision Trees).

(4) **Feature Selection:** Removing irrelevant or redundant features that mostly contain noise.

Idea 1.7.1

To determine whether our machine learning algorithm performs well not only on the training dataset but also generalizes well to new data, *we also want to randomly divide the dataset into separate training and test datasets.* In this way, we can measure whether our machine learning model is performing well.

1.7.2 TRAINING AND SELECTING A PREDICTIVE MODEL

We cannot get learning for free.

Observation 1.7.4

Meany different machine learning models have been developed to solve different problem tasks.

The goal is that sometimes is easier to use a certain machine learning model for some tasks than use it for other ones.

Example 1.7.2

Each classification algorithm has its inherent biases, and no single classification model enjoys superiority if we do not make any assumptions about the task.

In summary, it's essential to compare at least a handful of different learning algorithms in order to train and select the best performing model.

Observation 1.7.5

Before deciding which models we shall use, we have to decide upon a metric to measure performance. One commonly used metric is classification accuracy, which is defined as the proportion of correctly classified instances.

Sometimes we may have a machine learning model which was trained using a dataset for model selection, but one question arises: what if we do not use this training dataset? To address this issue, we can use different techniques such as cross-validation.

Definition 1.7.3 (Cross-Validation)

Cross-validation *we divide further a dataset into training and validation subsets in order to estimate the generalization performance of the model.*

Observation 1.7.6 (Hyperparameters)

We also cannot expect that the default parameters of the different learning algorithms provided by software libraries are optimal for our specific problem task. Therefore, we will make frequent use of hyperparameter optimization techniques that help us fine-tune the performance of our model in later sections.

We can think of those hyperparameters as parameters that are not learned from the data but represent the knobs of a model that we can turn to improve its performance.

1.7.3 EVALUATING MODELS AND PREDICTING UNSEEN DATA INSTANCES

After selecting a model that has been fitted on the training dataset, we can use the test dataset to estimate how well it performs on unseen data to determine the **generalization error**.

Definition 1.7.4 (Generalization Error)

Generalization Error *is formally the expected error of the model on a new, random instance drawn from the same underlying data distribution.*

Observation 1.7.7 (Dataset Dependancy)

It is important to note that the parameters for previously mentioned procedures, such as feature scaling and dimensionality reduction, **are solely obtained from the training dataset**, and the **same parameters are later reapplied to transform the test dataset as well as any new data instances—the performance measured on the test data may be overly optimistic otherwise**.

CHAPTER 2

TRAINING SIMPLE MACHINE LEARNING ALGORITHMS FOR CLASSIFICATION

So, we will see basic algorithms for classification and some other useful things.

2.1 INTRODUCTION

In this section we will make use of the first two algorithmically described machine learning algorithms for classification: the **perceptron** and **adaptive linear neurons**.

First, we'll start by implementing a perceptron step by step in Python and training it to classify different flower species in the iris dataset.

Observation 2.1.1

Using this algorithm and discussing the basics of optimization using adaptive linear neurons layse the groundwork for using more sophisticated classifiers.

The goals of this chapter are the following:

- Building an understanding of machine learning algorithms
- Using pandas, NumPy, and Matplotlib to read in, process, and visualize data
- Implementing linear classifiers for 2-class problems in Python

2.2 ARTIFICIAL NEURONS: EARLY STAGES OF MACHINE LEARNING

Warren McCulloch and Walter Pitts published the first concept of a simplified brain cell, the so-called McCulloch-Pitts (MCP) neuron, in 1943 (*A Logical Calculus of the Ideas Immanent in Nervous Activity* by W. S. McCulloch and W. Pitts, *Bulletin of Mathematical Biophysics*, 5(4): 115-133, 1943).

Observation 2.2.1 (Biological Neurons)

Biological Neurons are interconnected nerve cells in the brain that are involved in the processing and transmitting of chemical and electrical signals.

Idea 2.2.1

They described a nerve cell as a **simple logic gate with binary outputs**.

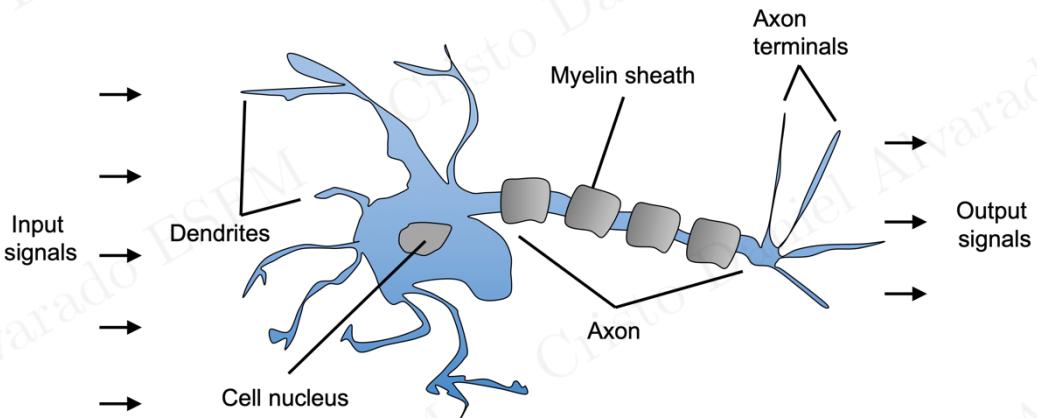


Figure 2.1: Biological Neuron.

Multiple signals arrive at the dendrites, they are then integrated into the cell body and, if the accumulated signals exceeds a certain threshold, an output signal is generated that will be passed on by the axon.

Observation 2.2.2 (Perceptron Learning Rule)

Only a few years later, Frank Rosenblatt published the first concept of the **perceptron learning rule based** on the MCP neuron model (*The Perceptron: A Perceiving and Recognizing Automaton* by F. Rosenblatt, Cornell Aeronautical Laboratory, 1957). With his perceptron rule, Rosenblatt proposed an algorithm that would automatically learn the optimal weight coefficients that would then be multiplied with the input features in order to make the decision of whether a neuron fires (transmits a signal) or not.

In the context of supervised learning and classification, such an algorithm could then be used to predict whether a new data point belongs to one class or the other.

2.3 FORMAL DEFINITION OF AN ARTIFICIAL NEURON

We can put the idea of **artificial neurons** into the context of a binary classification task with two classes: 0 and 1.

Definition 2.3.1 (Decision Function)

A **decision function** is a function $\sigma : X \rightarrow \{0, 1\}$ that takes a linear combination of a certain input values (called x) and a corresponding weight vector (called w), where z is the net input:

$$z = w_1x_1 + w_2x_2 + \cdots + w_mx_m = \sum_{i=1}^m w_i x_i$$

so, $\sigma(z) \in \{0, 1\}$.

If the net input of a particular example is greater than a defined threshold (lets say $\vartheta \in \mathbb{R}$, so $z \geq \vartheta$), we predict class 1, and class 0 otherwise. In the perceptron algorithm, the decision function, $\sigma(z)$ is a variant of a **unit step function**:

$$\sigma(z) = \begin{cases} 1 & \text{if } z \geq \vartheta \\ 0 & \text{otherwise} \end{cases}$$

Observation 2.3.1

We can modify the setup with a couple of steps. So, we make the **bias unit** $b = -\vartheta$, and we make:

$$z' = w_1x_1 + w_2x_2 + \cdots + w_mx_m + b = \vec{w}^T \vec{x} + b$$

And now replacing every input z by z' we obtain that:

$$\sigma(z') = \begin{cases} 1 & \text{if } z' \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Observation 2.3.2

The transpose A^T of matrix $A \in \mathcal{M}_{m \times n}$ is defined as:

$$A^T = \begin{bmatrix} a_1^{(1)} & a_2^{(1)} & \cdots & a_n^{(1)} \\ a_1^{(2)} & a_2^{(2)} & \cdots & a_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{(m)} & a_2^{(m)} & \cdots & a_n^{(m)} \end{bmatrix}$$

where:

$$A = \begin{bmatrix} a_1^{(1)} & a_2^{(1)} & \cdots & a_m^{(1)} \\ a_1^{(2)} & a_2^{(2)} & \cdots & a_m^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ a_1^{(n)} & a_2^{(n)} & \cdots & a_m^{(n)} \end{bmatrix}$$

Observation 2.3.3 (Notation Problem)

This notation is horrible and differs from the one Mathematicians use in all his textbooks, please, don't use this notation.

2.4 PERCEPTRON LEARNING RULE

The whole idea behind the MCP neuron and Rosenblatt's thresholded perceptron model is to *use a reductionist approach to mimic how a single neuron in the brain works*: **it either fires or it doesn't**. Thus, Rosenblatt's classic perceptron rule is fairly simple, and the perceptron algorithm can be summarized by the following steps:

1. Initialize the weights and bias unit to 0 or small random numbers
2. For each training example, $x^{(i)}$.
3. Compute the output value, $\hat{y}^{(i)}$.
4. Update the weight and bias unit.

Observation 2.4.1 (Output Value)

Here, the **output value** is the **class label** predicted by the unit step function defined earlier.

The simultaneous update of each weight w_j in the weight vector \vec{w} can be formally written as:

$$w_j \equiv w_j + \Delta w_j$$

and,

$$b \equiv b + \Delta b$$

Observation 2.4.2 (Use of the Symbol \equiv)

Here, the symbol \equiv states that the value of the variable to the left is updated to that of the variable on the right.

Where, the updated values (or **deltas**) are computed as follows:

$$\Delta w_j = \eta(y^{(i)} - \hat{y}^{(i)})x_j^{(i)}$$

and,

$$\Delta b = \eta(y^{(i)} - \hat{y}^{(i)})$$

Observation 2.4.3 (Notation Abuse)

This thing is absolutely ridiculous in terms of notation, the correct notation should be something like $\Delta w_j^{(i)}$, but since this variable is not going to be used anymore, we simply forgot about it, this is because in the process this variable is going to be updated every single time and we don't care about the different values depending on the i .

Here, η is a learning rate (usually between 0 and 1). $y^{(i)}$ is the **true class label** and $\hat{y}^{(i)}$ is the **predicted class label**.

Observation 2.4.4

The bias unit and all weights in the weight vector are updated simultaneously, which means that the predicted label $\hat{y}^{(i)}$ is not recomputed before the bias unit and all the weights are updated via the respective update values, Δw_j and Δb .

Example 2.4.1

Concretely, for a two-dimensional dataset, we would write the update as:

$$\begin{aligned}\Delta w_1 &= \eta(y^{(i)} - \text{output}^{(i)})x_1^{(i)} \\ \Delta w_2 &= \eta(y^{(i)} - \text{output}^{(i)})x_2^{(i)} \\ \Delta b &= \eta(y^{(i)} - \text{output}^{(i)})\end{aligned}$$

1. Before we implement the perceptron rule in Python, let's go through a simple thought experiment to illustrate how beautifully simple this learning rule really is. In the two scenarios where the perceptron predicts the class label correctly, the bias unit and weights remain unchanged, since the update values are 0:

2. When prediction is correct:

- If $y^{(i)} = 0$ and $\hat{y}^{(i)} = 0$:

$$\begin{aligned}\Delta w_j &= \eta(0 - 0)x_j^{(i)} = 0 \\ \Delta b &= \eta(0 - 0) = 0\end{aligned}$$

- If $y^{(i)} = 1$ and $\hat{y}^{(i)} = 1$:

$$\begin{aligned}\Delta w_j &= \eta(1 - 1)x_j^{(i)} = 0 \\ \Delta b &= \eta(1 - 1) = 0\end{aligned}$$

3. When prediction is wrong:

- If $y^{(i)} = 0$ and $\hat{y}^{(i)} = 1$:

$$\begin{aligned}\Delta w_j &= \eta(0 - 1)x_j^{(i)} = -\eta x_j^{(i)} \\ \Delta b &= \eta(0 - 1) = -\eta\end{aligned}$$

- If $y^{(i)} = 1$ and $\hat{y}^{(i)} = 0$:

$$\begin{aligned}\Delta w_j &= \eta(1 - 0)x_j^{(i)} = \eta x_j^{(i)} \\ \Delta b &= \eta(1 - 0) = \eta\end{aligned}$$

4. To get a better understanding of the feature value as a multiplicative factor $x_j^{(i)}$, consider another example where $y^{(i)} = 1$, $\hat{y}^{(i)} = 0$, $\eta = 1$. Assume that $x_j^{(i)} = 1.5$ and this example is misclassified as class 0. In this case, we would increase the corresponding weight so that the net input $z = x_j^{(i)}w_j + b$ would be more positive the next time we encounter this example and thus more likely to be above the threshold of the unit step function to classify the example as class 1:

$$\begin{aligned}\Delta w_j &= (1 - 0) \times 1.5 = 1.5 \\ \Delta b &= (1 - 0) = 1\end{aligned}$$

5. The weight update Δw_j is proportional to the value of $x_j^{(i)}$. For instance, if we have another example $x_j^{(i)} = 2$ that is incorrectly classified as class 0, we will push the decision boundary by an even larger extent to classify this example correctly the next time:

$$\begin{aligned}\Delta w_j &= (1 - 0) \times 2 = 2 \\ \Delta b &= (1 - 0) = 1\end{aligned}$$

6. The convergence of the perceptron is only guaranteed if the two classes are linearly separable, which means that the two classes can be perfectly separated by a linear decision boundary. If the two classes cannot be separated by a linear decision boundary, we can set a maximum number of passes over the training dataset (epochs) and/or a threshold for the number of tolerated misclassifications—the perceptron would never stop updating the weights otherwise.
7. The perceptron receives the inputs of an example (x) and combines them with the bias unit (b) and weights (w) to compute the net input. The net input is then passed on to the threshold function, which generates a binary output of 0 or 1—the predicted class label of the example. During the learning phase, this output is used to calculate the error of the prediction and update the weights and bias unit.

2.5 IMPLEMENTING A PERCEPTRON LEARNING ALGORITHM IN PYTHON

Now, we will do an implementation of the Perceptron Rule in Python.

2.6 AND OBJECT-ORIENTED PERCEPTRON API

We will take an object-oriented approach to defining the perceptron interface as a Python class, which will allow us to initialize new Perceptron objects that can learn from data via a `fit` method and make predictions via a separate `predict` method.

Observation 2.6.1

We append an underscore `_` to attributes that are not created upon the initialization of the object, but we do this by calling the object's other methods, for example, `self._w_`.

The code is the following:

```
1 import numpy as np
2
3 class Perceptron:
4     """Perceptron classifier.
5
6     Parameters
7     -----
8     eta : float
9         Learning rate (between 0.0 and 1.0)
10    n_iter : int
11        Passes over the training dataset.
12    random_state : int
13        Random number generator seed for random weight
14        initialization.
15
16     Attributes
17     -----
18     w_ : 1d-array
19         Weights after fitting.
20     b_ : Scalar
21         Bias unit after fitting.
22     errors_ : list
23         Number of misclassifications (updates) in each epoch.
24
25 """
26 def __init__(self, eta=0.01, n_iter=50, random_state=1):
27     self.eta = eta
28     self.n_iter = n_iter
29     self.random_state = random_state
30
31 def fit(self, X, y):
32     """Fit training data.
33
34     Parameters
35     -----
36     X : {array-like}, shape = [n_examples, n_features]
37         Training vectors, where n_examples is the number of
38         examples and n_features is the number of features.
39     y : array-like, shape = [n_examples]
40         Target values.
41
42     Returns
43     -----
44     self : object
45
46     """
```

```

47     rgen = np.random.RandomState(self.random_state)
48     self.w_ = rgen.normal(loc=0.0, scale=0.01, size=X.shape
49                           [1])
50     self.b_ = np.float_(0.)
51     self.errors_ = []
52
53     for _ in range(self.n_iter):
54         errors = 0
55         for xi, target in zip(X, y):
56             update = self.eta * (target - self.predict(xi))
57             self.w_ += update * xi
58             self.b_ += update
59             errors += int(update != 0.0)
60         self.errors_.append(errors)
61     return self
62
63     def net_input(self, X):
64         """Calculate net input"""
65         return np.dot(X, self.w_) + self.b_
66
67     def predict(self, X):
68         """Return class label after unit step"""
69         return np.where(self.net_input(X) >= 0.0, 1, 0)

```

Code 2.1: Python Perceptron Rule Implementation.

First, we will list the parameters and attributes of the code, in order to understand it better:

- **eta**: Is a float, which represents the learning rate (a number between 0.0 and 1.0).
- **n_iter**: Is an integer, which is the number of passes over the training dataset.
- **random_state**: It is an integer, which is a **random number generator seed** for random weight initialization.

Definition 2.6.1 (Random Number Generator (RNG))

A **random number generator (RNG) seed** is *an initial value used by a computer algorithm to generate a sequence of pseudo-random numbers*. For random weight *initialization in a neural network, this seed sets the starting point for the sequence of numbers used to initialize the model's weights and biases*.

The attributes are the following:

- **w_**: It is a 1-dimensional array, which represents the weights after fitting.
- **b_**: It is a float, which represents the bias unit after fitting.
- **errors_**: It is a list, which is the number of misclassifications (or updates) in each **epoch**.

Definition 2.6.2 (Epoch)

In machine learning, an **epoch** is *one complete pass through the entire training dataset*. During a single epoch, *the learning algorithm processes every training example once to update the model's internal parameters*. Training a model over multiple epochs *allows it to progressively learn and*

refine its understanding of the data, with a well-tuned number of epochs balancing performance between underfitting and overfitting.

In this code, we have three methods in the class Perceptron:

- `__init__`: It initializes the parameters of the class, in this case, the `eta` is setted to 0.01, `n_iter`=50 and `random_state`=1.
- `fit`: It is the code that fits the data in a dataset, given an array `X` and another array `y` of target values.
- `net_input`: It calculates the net input of a dataset `X`.
- `predict`: It calculates the class label after a unit step of a dataset `X`.

Observation 2.6.2 (Use of the Perceptron Implementation)

Using this perceptron implementation, we can initialize new Perceptron objects with a given learning rate, `eta`, and the number of epochs, `n_iter` (or iterations over the training dataset).

Via the `fit` method, we initialize the bias `self.b_` to an initial value 0 and the weights in `self.w_` to a vector in \mathbb{R}^m . Here, m stands for the number of dimensions (features) in the dataset.

Observation 2.6.3 (Decision of Initialization of Weight Vectors)

The **initial weight** vector contains small random numbers drawn from a normal distribution with a standard deviation of 0.01 via `rgen.normal(loc=0.0, scale=0.01, size=1 + X.shape[1])`, where `rgen` is a NumPy random number generator seeded with a user-specified random seed so that results can be reproduced when desired.

Technically, we could initialize the weights to zero (this is done in the original perceptron algorithm). However, if we did that, the **learning rate** η (`eta`) would have no effect on the decision boundary. If all the weights are initialized to zero, the learning rate parameter affects only the scale of the weight vector, not the direction.

Observation 2.6.4 (Use of Normal Distribution)

Our decision to draw the random numbers from a normal distribution—for example, instead of from a uniform distribution—and to use a standard deviation of 0.01 was arbitrary; remember, we are just interested in small random values to avoid the properties of all-zero vectors, as discussed earlier.

Excercise 2.6.1

As an optional exercise, you can change `self.w_ = rgen.normal(loc=0.0, scale=0.01, size=X.shape[1])` to `self.w_ = np.zeros(X.shape[1])` and run the perceptron training code presented in the next section with different values for `eta`. You will observe that the decision boundary does not change.

After the weights have been initialized, the fit method loops over all individual examples in the training dataset and updates the weights according to the perceptron learning rule discussed in the previous section.

Observation 2.6.5 (Prediction)

The class labels are predicted by the `predict` method, which is called in the `fit` method during training to get the class label for the weight update; but `predict` can also be used to predict the class labels of new data after the model has been fitted. Furthermore, we collect the number of misclassifications during each epoch in the `self.errors_` list so that we can later analyze how well the perceptron performed during training. The `np.dot` function used in the `net_input` method calculates the vector dot product, $wTx + bwTx + b$.

Idea 2.6.1

Instead of using NumPy to calculate the vector dot product between two arrays, `a` and `b`, via `a.dot(b)` or `np.dot(a, b)`, we could also perform the calculation in pure Python via `sum([i * j for i, j in zip(a, b)])`. However, the advantage of using NumPy over classic Python for loop structures is that its arithmetic operations are vectorized. Vectorization means that an elemental arithmetic operation is automatically applied to all elements in an array. By formulating our arithmetic operations as a sequence of instructions on an array, rather than performing a set of operations for each element at a time, we can make better use of our modern central processing unit (CPU) architectures with single instruction, multiple data (SIMD) support. Furthermore, NumPy uses highly optimized linear algebra libraries, such as **Basic Linear Algebra Subprograms (BLAS)** and **Linear Algebra Package (LAPACK)**, that have been written in C or Fortran. Lastly, NumPy also allows us to write our code in a more compact and intuitive way using the basics of linear algebra, such as vector and matrix dot products.

2.7 TRAINING OF THE PERCEPTRON ALGORITHM ON THE IRIS DATASET

Observation 2.7.1 (Restriction of Dimensions)

To test our perceptron implementation, we will restrict the following analyses and examples in the remainder of this section to two feature variables (dimensions). Although the perceptron rule is not restricted to two dimensions, considering only two features—sepal length and petal length—allows us to visualize the decision regions of the trained model in a scatterplot.

Idea 2.7.1 (Restriction to Only Two Classes)

We will also only consider two flower classes, *setosa* and *versicolor*, from the Iris dataset for practical reasons (remember, the **perceptron is a binary classifier**). However, the perceptron algorithm can be extended to multi-class classification using the one-versus-all (OvA) technique.

2.7.1 OvA METHOD FOR MULTI-CLASS CLASSIFICATION

We can extend any binary classifier to a multi-class problems using the method called **OvA** or **one-versus-rest (OvR)**.

Definition 2.7.1 (One-versus-rest (OvR))

The method **OvR** allows to extend any binary classifier to a multiple classifier. To classify a new, unlabeled instance, we use our n classifiers (where n is the number of class labels) and assign the class label with the highest confidence to the instance. For the perceptron, we choose the class label associated with the largest absolute net input value.

First, we use the `pandas` library to load the Iris dataset directly from the **UCI Machine Learning Repository** into a `DataFrame` object and *print the last five lines via the `tail` method to confirm that the data loaded correctly*:

```

1 import os
2 import pandas as pd
3 s = 'https://archive.ics.uci.edu/ml/' \
4     'machine-learning-databases/iris/iris.data'
5 print('From URL:', s)
6 #From URL: https://archive.ics.uci.edu/ml/machine-learning-
    databases/iris/iris.data
7 df = pd.read_csv(s,
8                  header=None,
9                  encoding='utf-8')
10 print(df.tail())

```

Code 2.2: Read Dataset from UCI.

```

Drive/Conocimiento/Conocimiento/Knowledge/Programacion/Courses/Machine Learning with
From URL: https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
      0   1   2   3        4
145  6.7  3.0  5.2  2.3  Iris-virginica
146  6.3  2.5  5.0  1.9  Iris-virginica
147  6.5  3.0  5.2  2.0  Iris-virginica
148  6.2  3.4  5.4  2.3  Iris-virginica
149  5.9  3.0  5.1  1.8  Iris-virginica

```

Figure 2.2: Tail of the Iris Dataset.

Next, we extract the first 100 class labels that correspond to the 50 `Iris-setosa` and 50 `Iris-versicolor` flowers and convert the class labels into the two integer class labels 1 (`versicolor`) and 0 (`setosa`), assigning them to a vector `y`. Similarly, we extract the first feature column (`sepal length`) and the third feature column (`petal length`) of those 100 training examples and assign them to a feature matrix `X`, which we then visualize via a two-dimensional scatterplot:

```

1 import matplotlib.pyplot as plt
2 import numpy as np
3 # select setosa and versicolor
4 y = df.iloc[0:100, 4].values
5 y = np.where(y == 'Iris-setosa', 0, 1)
6 # extract sepal length and petal length
7 X = df.iloc[0:100, [0, 2]].values
8 # plot data
9 plt.scatter(X[:50, 0], X[:50, 1],
10             color='red', marker='o', label='Setosa')
11 plt.scatter(X[50:100, 0], X[50:100, 1],
12             color='blue', marker='s', label='Versicolor')
13 plt.xlabel('Sepal length [cm]')
14 plt.ylabel('Petal length [cm]')
15 plt.legend(loc='upper left')
16 plt.show()

```

Code 2.3: Extraction and Visualization of Iris Dataset.

Observation 2.7.2 (Use of the Code)

Add this code to the remaining code.

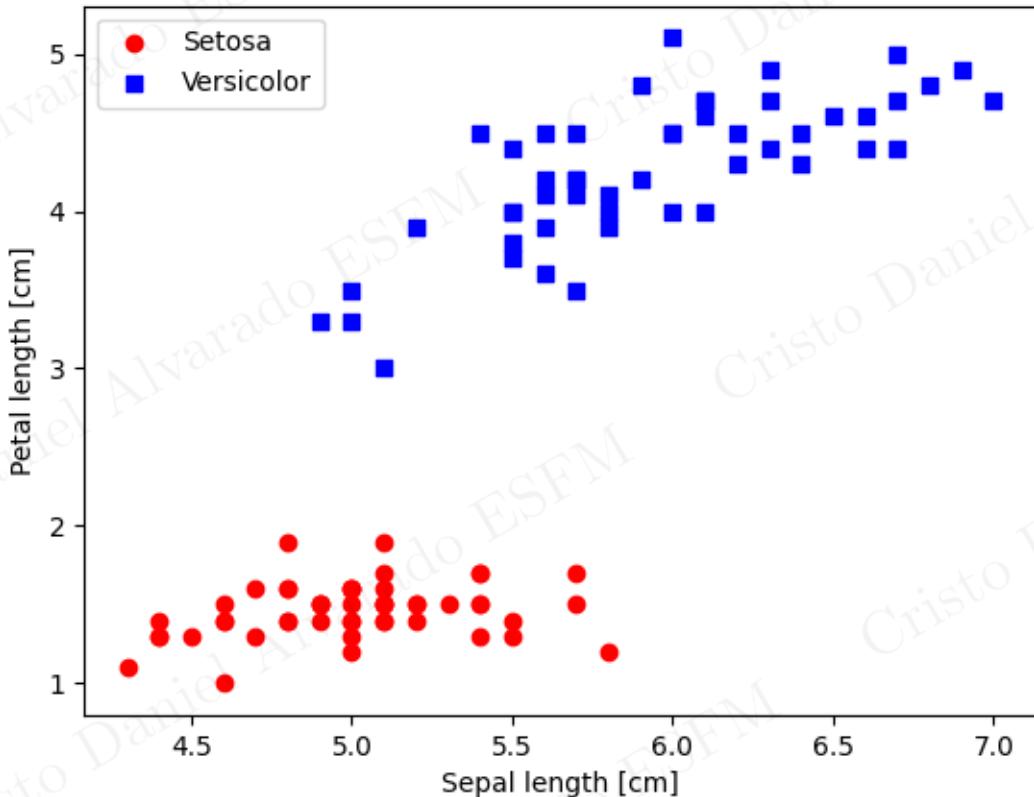


Figure 2.3: Plot of the Seritosa and Versicolor on the Iris Dataset.

This plot shows the distribution of flower examples in the Iris dataset, along with the two feature axes: petal and sepal length.

Observation 2.7.3

In this subspace, a linear decision boundary should be sufficient to separate setosa from versicolor flowers.

Thus, a linear classifier such as the perceptron should be able to classify the flowers in the Iris dataset perfectly.

2.7.2 TRAINING THE ALGORITHM

We will also plot the misclassification error for each epoch to check whether the algorithm converged and found a decision boundary that separates the two Iris flower classes:

```
1 import Perceptron
2 ppn = Perceptron(eta=0.1, n_iter=10)
3 ppn.fit(X, y)
4 plt.plot(range(1, len(ppn.errors_) + 1),
5           ppn.errors_, marker='o')
6 plt.xlabel('Epochs')
```

```
7 plt.ylabel('Number of updates')
8 plt.show()
```

Code 2.4: Training the Perceptron Algorithm with the Iris Dataset

Observation 2.7.4

In this scenario, since I put Perceptron in a different Python file, I have to import it in order to be able to use it.

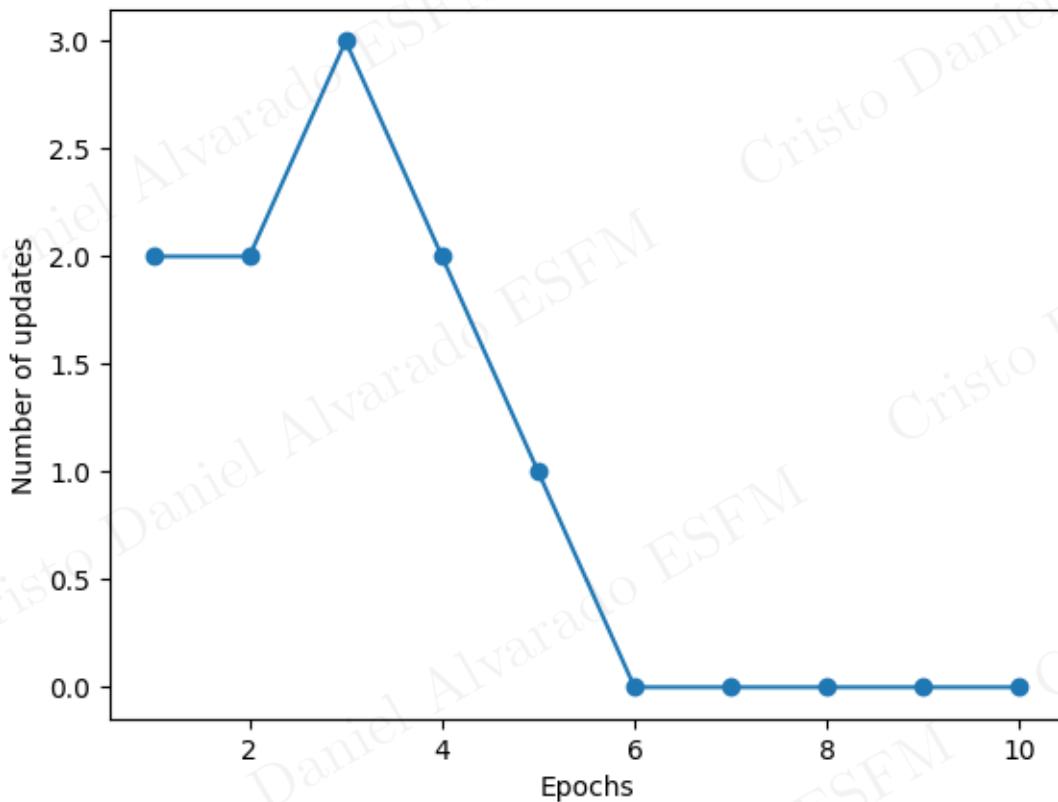


Figure 2.4: Epoch Plot and Number of Updates on the Iris Dataset.

Observation 2.7.5 (Convergence of the Perceptron)

As we can see, our perceptron converged after the sixth epoch and should now be able to classify the training examples perfectly.

Let's implement a small convenience function to visualize the decision boundaries for two-dimensional datasets:

```
1 from matplotlib.colors import ListedColormap
2
3 def plot_decision_regions(X, y, classifier, resolution=0.02):
4     # setup marker generator and color map
5     markers = ('o', 's', '^', 'v', '<')
6     colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
7     cmap = ListedColormap(colors[:len(np.unique(y))])
```

```

8
9     # plot the decision surface
10    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
11    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
12    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution)
13                           ,
14                           np.arange(x2_min, x2_max, resolution)
15                           )
16
17    lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
18                             ).T
19
20    lab = lab.reshape(xx1.shape)
21    plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
22    plt.xlim(xx1.min(), xx1.max())
23    plt.ylim(xx2.min(), xx2.max())
24
25    # plot class examples
26    for idx, cl in enumerate(np.unique(y)):
27        plt.scatter(x=X[y == cl, 0],
28                    y=X[y == cl, 1],
29                    alpha=0.8,
30                    c=colors[idx],
31                    marker=markers[idx],
32                    label=f'Class {cl}',
33                    edgecolor='black')

```

Code 2.5: Plotting Regions on the Iris Dataset

First, we define a number of *colors* and *markers* and create a colormap from the list of colors via `ListedColormap`. Then, we determine the minimum and maximum values for the two features and use those feature vectors to create a pair of grid arrays, `xx1` and `xx2`, via NumPy's `meshgrid` function. Since we trained our perceptron classifier on two feature dimensions, we flatten the grid arrays and create a matrix with the same number of columns as the Iris training subset so that we can use the `predict` method to predict the class labels `lab` of the corresponding grid points.

After reshaping the predicted class labels into a grid with the same dimensions as `xx1` and `xx2`, we draw a contour plot via Matplotlib's `contourf` function, which maps the different decision regions to different colors for each predicted class in the grid array:

```

1 plot_decision_regions(X, y, classifier=ppn)
2 plt.xlabel('Sepal length [cm]')
3 plt.ylabel('Petal length [cm]')
4 plt.legend(loc='upper left')
5 plt.show()

```

Code 2.6: Plotting the Perceptron Info on the Iris Dataset

As we can see in the plot, the perceptron learned a decision boundary that can classify all flower examples in the Iris training subset perfectly.

Observation 2.7.6 (Convergence of the Perceptron)

Although the perceptron classified the two Iris flower classes perfectly, convergence is one of its biggest challenges. Rosenblatt proved mathematically that the **perceptron learning rule converges if the two classes can be separated by a linear hyperplane**. However, if the

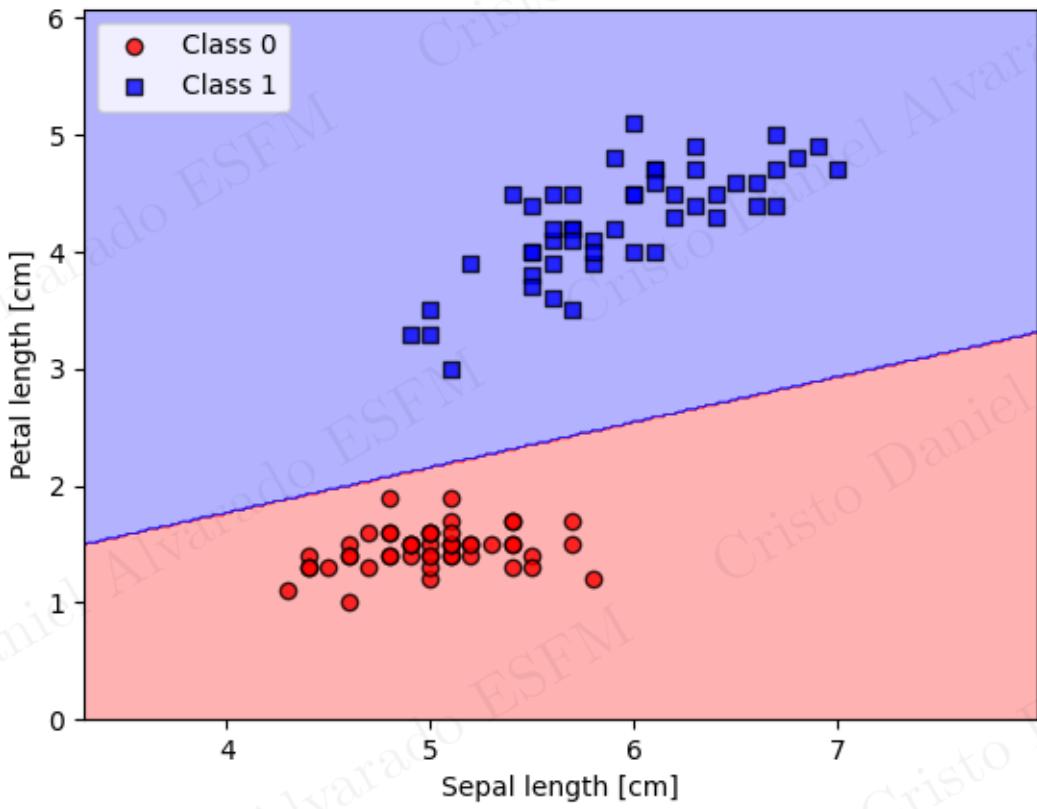


Figure 2.5: Perceptron Visual Classification on the Iris Dataset.

classes cannot be separated perfectly by such a linear decision boundary, the weights will never stop updating unless we set a maximum number of epochs. Interested readers can find a summary of the proof in the lecture notes at [Perceptron Slides](#).

2.8 ADAPTIVE LINEAR NEURONS AND THE CONVERGENCE OF LEARNING

ADaptive LInear NEuron (Adaline). Adaline was published by Bernard Widrow and his doctoral student Tedd Hoff only a few years after Rosenblatt's perceptron algorithm, and it can be considered an improvement on the latter (An Adaptive "Adaline" Neuron Using Chemical "Memistors", Technical Report Number 1553-2 by B. Widrow and colleagues, Stanford Electron Labs, Stanford, CA, October 1960).

This algorithm is interesting because it illustrates the key concepts of defining and minimizing a continuous loss function.

Observation 2.8.1 (Use of Adaline)

This lays the groundwork for understanding other machine learning algorithms for classification, such as logistic regression, support vector machines, and multilayer neural networks, as well as linear regression models.

Idea 2.8.1 (Widrow-Hoff Rule)

The key difference between Adaline and the Perceptron is that weights are updated based on a linear activation function rather than a unit step function.

This is called **Widrow-Hoff rule**.

In Adaline, the linear activation function is the identity function of the net input, that is: $\sigma(z) = z$.

Observation 2.8.2

While the linear activation function is used for learning the weights, a threshold function is still applied to make the final prediction, which is similar to the unit step function covered earlier.

2.8.1 MAIN DIFFERENCES BETWEEN PERCEPTRON AND ADALINE

Observation 2.8.3 (Perceptron vs Adeline)

The diagram shows that the Adaline algorithm compares the true class labels with the linear activation function's continuous-valued output to compute the model error and update the weights, whereas the perceptron compares the true class labels to the predicted class labels.

2.8.2 MINIMIZING LOSS FUNCTION WITH GRADIENT DESCENT

One of the key ingredients of supervised machine learning algorithms is a *defined objective function that is optimized during learning*. This *objective function is often a loss or cost function that we want to minimize*.

Observation 2.8.4 (Loss Function in Adeline)

In the case of Adaline, the **loss function**, L , can be defined as the mean squared error (MSE) between the calculated outcome and the true class label:

$$L(w, b) = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \sigma(z^{(i)}))^2$$

It turns out that the continuous linear activation function makes the loss function differentiable.

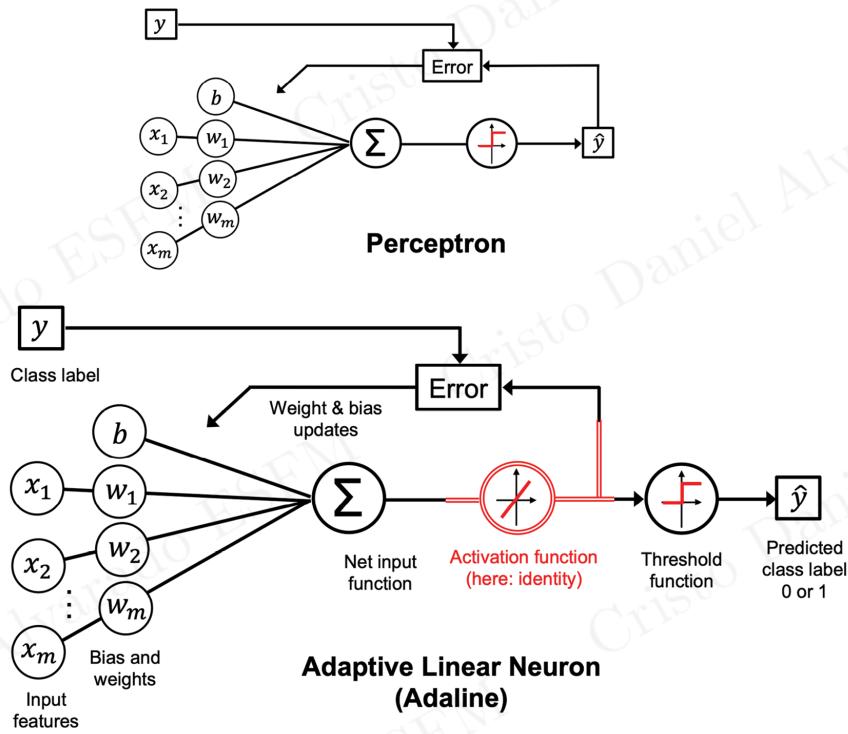


Figure 2.6: Perceptron vs Adeline Work Flux.

Also, this function is **convex**.

Definition 2.8.1 (Convex Function)

Let X be a convex subset of a real vector space, and $f : X \rightarrow \mathbb{R}$ a function. f is called **convex** if for all $x_1, x_2 \in X$ we have that:

$$f(tx_1 + (1 - t)x_2) \leq tf(x_1) + (1 - t)f(x_2), \quad \forall t \in [0, 1]$$

The following image represents what a convex continuous function should look like:

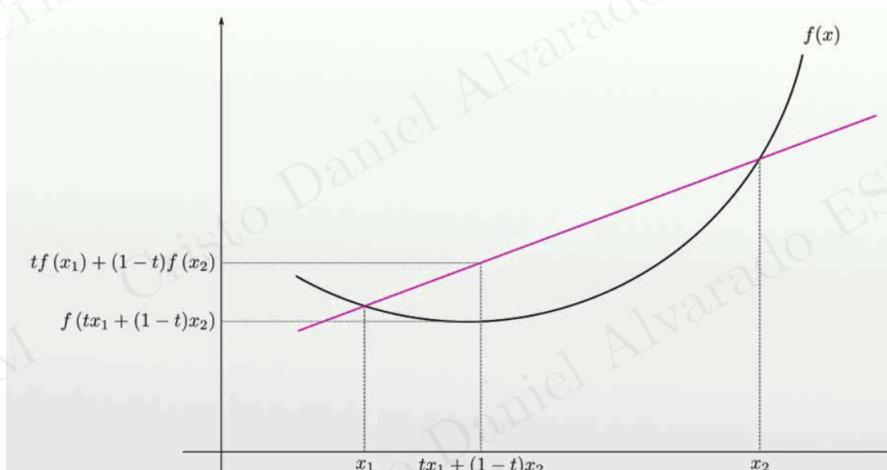


Figure 2.7: Convex Function.

Observation 2.8.5 (About Convex Functions)

When having a convex function and the inequality of the latter definition, the right side represents a straight line between $f(x_1)$ and $f(x_2)$. The argument of the function represents the line joining x_1 and x_2 , so we compute its image under f .

Due to the fact that the loss function is convex, we have a simple yet powerful optimization algorithm called **gradient descent** can be used to find the weights that minimize the loss function when classifying examples in the Iris dataset.

Idea 2.8.2 (Gradient Descent)

The idea behind **gradient descent** is to climb down a hill until a local or global loss minimum is reached.

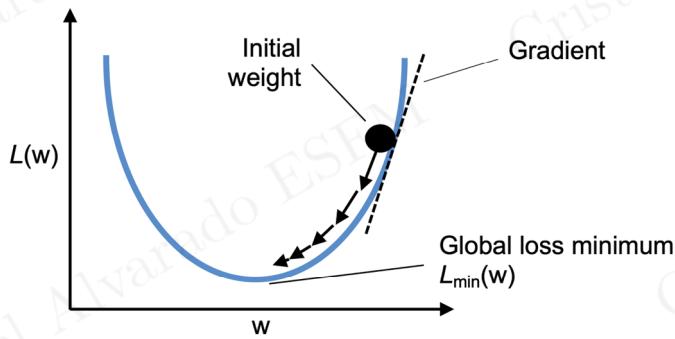


Figure 2.8: Gradient Descent.

The idea behind the algorithm is that on each iteration, a step is taken in the opposite direction of the gradient, where the step size is determined by the learning rate value and the slope of the gradient.

Observation 2.8.6 (How Does Gradient Descent Works?)

Gradient descent updates each parameter by moving it in the direction that reduces the loss:

$$\vartheta - \nabla_{\vartheta} L(\vartheta) \rightarrow \vartheta$$

Here, ϑ represents any parameter of the model. But since we have two different parameters w and b , we need to compute their individual gradients.

Using gradient descent, the model parameters are updated by taking a step in the opposite direction of the gradient, $\nabla L(w, b)$ of the loss function $L(w, b)$:

$$w = w + \Delta w$$
$$b = b + \Delta b$$

The parameter changes, Δw and Δb are defined as the negative gradient multiplied by the learning rate η :

$$\Delta w = -\eta \nabla_w L(w, b)$$
$$\Delta b = -\eta \nabla_b L(w, b)$$

To compute the gradient of the loss function, compute the partial derivative of the loss function with respect to each weight, w_j :

$$\frac{\partial L}{\partial w_j} = -\frac{2}{n} \sum_{i=1}^n (y^{(i)} - \sigma(z^{(i)})) x_j^{(i)}$$

Similarly, compute the partial derivative of the loss with respect to the bias:

$$\frac{\partial L}{\partial b} = -\frac{2}{n} \sum_{i=1}^n (y^{(i)} - \sigma(z^{(i)}))$$

So, the weight updates becomes:

$$\begin{aligned}\Delta w_j &= -\eta \frac{\partial L}{\partial w_j} \\ \Delta b &= -\eta \frac{\partial L}{\partial b}\end{aligned}$$

Observation 2.8.7

Let's recall that $L : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}$, so its gradient with respect to w and b will be:

$$\begin{aligned}\nabla_w L(w, b) &= \left[\begin{array}{c} \frac{\partial L}{\partial w_1}(w, b) \\ \frac{\partial L}{\partial w_2}(w, b) \\ \vdots \\ \frac{\partial L}{\partial w_j}(w, b) \\ \vdots \\ \frac{\partial L}{\partial w_n}(w, b) \end{array} \right] \\ \nabla_b L(w, b) &= \left[\frac{\partial L}{\partial b}(w, b) \right] = \frac{\partial L}{\partial b}(w, b)\end{aligned}$$

respectively. This makes perfect sense with the latter observation.

Note: The MSE derivative involves calculus for weight adjustments.

Observation 2.8.8 (Batch Gradient Descent)

Although the Adaline learning rule looks identical to the perceptron rule, $\sigma(z^{(i)})$ where $z^{(i)} = w^T x^{(i)} + b$ is a real number, not an integer class label.

Furthermore, the *weight update is calculated based on all examples in the training dataset (instead of updating the parameters incrementally after each training example)*, which is why this approach is often called **batch gradient descent**. To be explicit and avoid confusion when discussing related concepts later in this section and throughout this course, this process is referred to as full **batch gradient descent**.

2.9 IMPLEMENTING ADALINE IN PYTHON

Since the perceptron rule and Adaline are very similar, we will take the perceptron implementation defined earlier and change the `fit` method so that the *weight and bias parameters are now updated by minimizing the loss function via gradient descent*:

```

 1 class AdalineGD:
 2     """ADAptive LInear NEuron classifier.
 3
 4     Parameters
 5     -----
 6     eta : float
 7         Learning rate (between 0.0 and 1.0)
 8     n_iter : int
 9         Passes over the training dataset.
10     random_state : int
11         Random number generator seed for random weight
12             initialization.
13
14     Attributes
15     -----
16     w_ : 1d-array
17         Weights after fitting.
18     b_ : Scalar
19         Bias unit after fitting.
20     losses_ : list
21         Mean squared error loss function values in each epoch.
22     """
23     def __init__(self, eta=0.01, n_iter=50, random_state=1):
24         self.eta = eta
25         self.n_iter = n_iter
26         self.random_state = random_state
27
28     def fit(self, X, y):
29         """ Fit training data.
30
31         Parameters
32         -----
33         X : {array-like}, shape = [n_examples, n_features]
34             Training vectors, where n_examples
35             is the number of examples and
36             n_features is the number of features.
37         y : array-like, shape = [n_examples]
38             Target values.
39
40         Returns
41         -----
42         self : object
43     """
44         rgen = np.random.RandomState(self.random_state)
45         self.w_ = rgen.normal(loc=0.0, scale=0.01,
46                               size=X.shape[1])
47         self.b_ = np.float_(0.)
48         self.losses_ = []
49
50         for i in range(self.n_iter):

```

```

51             net_input = self.net_input(X)
52             output = self.activation(net_input)
53             errors = (y - output)
54             self.w_ += self.eta * 2.0 * X.T.dot(errors) / X.
55                 shape[0]
56             self.b_ += self.eta * 2.0 * errors.mean()
57             loss = (errors**2).mean()
58             self.losses_.append(loss)
59             return self
60
61     def net_input(self, X):
62         """Calculate net input"""
63         return np.dot(X, self.w_) + self.b_
64
65     def activation(self, X):
66         """Compute linear activation"""
67         return X
68
69     def predict(self, X):
70         """Return class label after unit step"""
71         return np.where(self.activation(self.net_input(X))
72                         >= 0.5, 1, 0)

```

Code 2.7: Adaline Implementation in Python.

Observation 2.9.1 (Key Difference Between Adaline and Perceptron)

Instead of updating the weights after evaluating each individual training example, as in the perceptron, we calculate the gradient based on the whole training dataset.

For the bias unit, *this is done via `self.eta * 2.0 * errors.mean()`*, where `errors` is an array containing the partial derivative values $\frac{\partial L}{\partial b}$. Similarly, we update the weights. However, note that the weight updates via the partial derivatives $\frac{\partial L}{\partial w_j}$ involve the feature values x_j , which we can compute by multiplying `errors` with each feature value for each weight:

```

1 for w_j in range(self.w_.shape[0]):
2     self.w_[w_j] += self.eta * \
3         (2.0 * (X[:, w_j]*errors)).mean()

```

Code 2.8: Computation of Partial Derivatives

To implement the weight update more efficiently without using a for loop, we can use a matrix-vector multiplication between the feature matrix and the error vector instead:

```

1 self.w_ += self.eta * 2.0 * X.T.dot(errors) / X.shape[0]

```

Code 2.9: Weight Update Optimization.

Observation 2.9.2 (Activation Method)

The `activation` method has no effect on the code since it is simply an identity function. It is included to illustrate how information flows through a single-layer neural network: features from the input data, net input, activation, and output.

Later, we will learn about a logistic regression classifier that uses a *non-identity, nonlinear activation function*.

Observation 2.9.3 (Logistic Regression Model)

A logistic regression model is closely related to Adaline, with the only difference being its activation and loss function.

Similar to the previous perceptron implementation, we collect the loss values in a `self.losses_list` to check whether the algorithm converges after training.

Observation 2.9.4 (Note)

Matrix multiplication uses a vectorized approach for efficient computation.

In practice, it often requires some experimentation to find a good learning rate, η , for optimal convergence. Let's choose two different learning rates, $\eta = 0.1$ and $\eta = 0.0001$, and plot the loss values versus the number of epochs to see how well the Adaline implementation learns from the training data.

2.9.1 HYPERPARAMETERS

The learning rate, η (`eta`), and the number of epochs (`n_iter`) are hyperparameters of the perceptron and Adaline learning algorithms. Various techniques can help automatically find hyperparameter values that yield optimal classifier performance.

Let's now plot the loss against the number of epochs for the two different learning rates:

```
1 > fig, ax = plt.subplots(nrows=1, ncols=2, figsize=(10, 4))
2 ...           np.log10(ada1.losses_), marker='o')
3 > ax[0].set_xlabel('Epochs').set_ylabel('log(Mean squared
   error)').set_title('Adaline Loss vs Epochs')
4 ...           ada2.losses_, marker='o')
5 Set labels and title for the Adaline plot, then display it.
```

Code 2.10: Loss for Two Different Learning Rates.

Using a learning rate that is too large can cause the mean squared error (MSE) to increase in every epoch because the updates overshoot the global minimum. Conversely, a learning rate that is too small can lead to very slow convergence, requiring a large number of epochs to reach the minimum loss.

2.10 IMPROVING GRADIENT DESCENT THROUGH FEATURE SCALING

Many machine learning algorithms require some sort of feature scaling for optimal performance.

Definition 2.10.1 (Feature Scaling)

Feature Scaling is a crucial data preprocessing step in machine learning. Its primary goal is to normalize the range of independent variables or features of your data.

Example 2.10.1 (Feature Scaling)

When your dataset contains features that are on vastly different scales (e.g., age (0-100) and annual salary (\$30,000 - \$200,000)), many machine learning algorithms can behave poorly. Feature

scaling transforms these features onto a similar, comparable scale

Observation 2.10.1 (Feature Scaling)

Gradient descent is one of the many algorithms that benefit from **feature scaling**.

Definition 2.10.2 (Standarization)

Standarization is a normalization procedure that helps gradient descent learning converge more quickly; however it does not make the original dataset normally distributed.

Standarization shifts the mean of each feature so that it is centered at zero, and each feature has a standar deviation of 1 (unit variance).

Example 2.10.2 (Standarization Procedure)

For instance, to standarize the j -th feature, subtract the sample mean, μ_j from every training example and divide it by its standar deviation, σ_j :

$$x'_j = \frac{x_j - \mu_j}{\sigma_j}$$

Here, x_j is a vector consistent of the j -th feature values of all training examples, n , and this standarization techinque applied to each feature, j in the dataset.

Observation 2.10.2 (Use of Standarization)

One reason why standardization helps with gradient descent learning is that it **becomes easier to find a learning rate that works well for all weights** (and the bias). *If the features are on vastly different scales, a learning rate that works well for updating one weight might be too large or too small to update another weight equally well. Using standarized features stabilizes the training so that the optimizer requires fewer steps to find a good or optimal solution (the global loss minimum).*

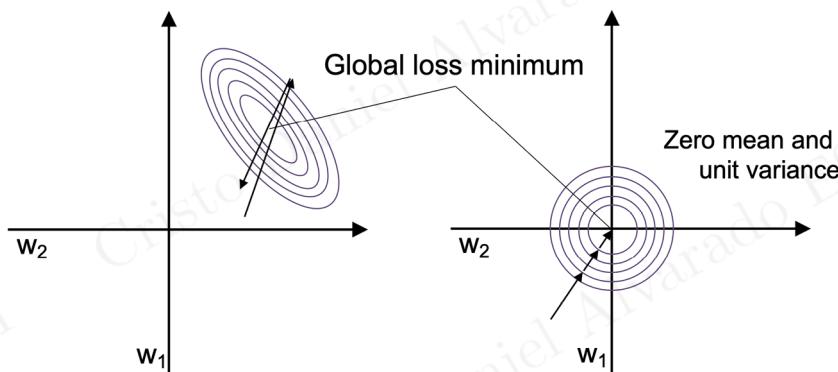


Figure 2.9: Global Loss Minimum with and without Zero Variance.

Standardization can easily be achieved by using the built-in NumPy methods `mean` and `std`:

```

1 X_std = np.copy(X)
2 X_std[:,0] = (X[:,0] - X[:,0].mean()) / X[:,0].std()
3 X_std[:,1] = (X[:,1] - X[:,1].mean()) / X[:,1].std() code

```

Code 2.11: Standardization Using Numpy

After standardization, train Adaline again; it now converges after a small number of epochs using a learning rate of $\eta = 0.5$:

```

1 ada_gd = AdalineGD(n_iter=20, eta=0.5)
2 ada_gd.fit(X_std, y)
3 plot_decision_regions(X_std, y, classifier=ada_gd)
4 plt.title('Adaline - Gradient descent')
5 plt.xlabel('Sepal length [standardized]')
6 plt.ylabel('Petal length [standardized]')
7 plt.legend(loc='upper left')
8 plt.tight_layout()
9 plt.show()
10 plt.plot(range(1, len(ada_gd.losses_) + 1),
11           ada_gd.losses_, marker='o')
12 plt.xlabel('Epochs')
13 plt.ylabel('Mean squared error')
14 plt.tight_layout()
15 plt.show()

```

Code 2.12: Adaline With Standardization Procedure

Executing this code displays the decision regions and the declining loss.

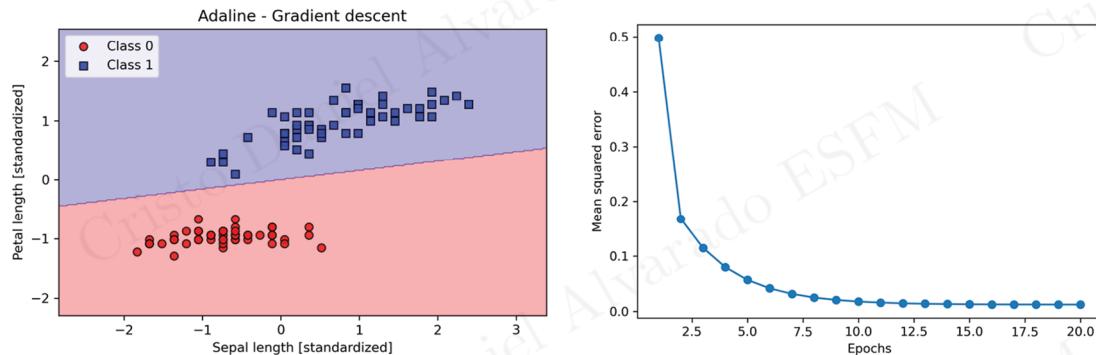


Figure 2.10: Adaline after Standardization Procedure

After training on the standardized features, Adaline has converged. However, the MSE remains non-zero even though all flower examples were classified correctly.

2.10.1 LARGE-SCALE MACHINE LEARNING AND STOCHASTIC GRADIENT DESCENT

Previously, we minimized a loss function by taking a step in the opposite direction of the loss gradient that is calculated from the whole training dataset; this approach is sometimes referred to as **full batch gradient descent**.

Observation 2.10.3 (Problem with Full Batch Gradient Descent)

For datasets with millions of data points, *running full batch gradient descent can be computationally costly because the whole training dataset must be re-evaluated each time a step toward the global minimum is taken.*

A popular alternative to batch gradient descent is **stochastic gradient descent (SGD)**, sometimes *called iterative or online gradient descent*. *Instead of updating the weights based on the sum of the accumulated errors over all training examples, $x^{(i)}$:*

$$\Delta w_j = \frac{2\eta}{n} \sum_{i=1}^n (y^{(i)} - \sigma(z^{(i)})) x_j^{(i)}$$

We update the parameters incrementally for each training example:

$$\Delta w_j = \eta(y^{(i)} - \sigma(z^{(i)})) x_j^{(i)}, \quad \text{and} \quad \Delta b = \eta(y^{(i)} - \sigma(z^{(i)}))$$

Observation 2.10.4

Although SGD can be considered an approximation of gradient descent, it typically reaches convergence much faster because of the more frequent weight updates.

Since each gradient is calculated from a single training example, the error surface is noisier than in gradient descent, which can help SGD escape shallow local minima when working with nonlinear loss functions. To obtain good results with SGD, present training data in a random order and shuffle the training dataset for every epoch to prevent cycles.

Observation 2.10.5 (Note)

Use an adaptive learning rate for better training results

Idea 2.10.1 (Advantages of SGD)

Another advantage of SGD is that *it supports online learning*. In online learning, *the model is trained on the fly as new training data arrives*. This is **especially useful when accumulating large amounts of data**, for example, customer data in web applications. *The system can immediately adapt to changes, and the training data can be discarded after updating the model if storage space is an issue.*

Observation 2.10.6 (Note)

Mini-batch gradient descent speeds up convergence with frequent updates.

Since the Adaline learning rule using gradient descent is already implemented, only minor adjustments are needed to modify the learning algorithm for weight updates via SGD.

Inside the `fit` method, weights are updated after each training example. An additional `partial_fit` method, which does not reinitialize the weights, supports online learning. To check whether the algorithm converged after training, the average loss of the training examples is calculated for each epoch. An option to shuffle the training data before each epoch avoids repetitive cycles when optimizing the loss function, and the `random_state` parameter allows specification of a random seed for reproducibility:

```

1  """ADAptive LInear NEuron classifier.
2
3 Parameters
4 -----
5 eta : float
6     Learning rate (between 0.0 and 1.0)
7 n_iter : int
8     Passes over the training dataset.
9 shuffle : bool (default: True)
10    Shuffles training data every epoch if True to prevent
11    cycles.
12 random_state : int
13    Random number generator seed for random weight
14    initialization.
15
16 Attributes
17 -----
18 w_ : 1d-array
19    Weights after fitting.
20 b_ : Scalar
21    Bias unit after fitting.
22 losses_ : list
23    Mean squared error loss function value averaged over all
24    training examples in each epoch.
25
26 """
27 def __init__(self, eta=0.01, n_iter=10,
28              shuffle=True, random_state=None):
29     self.eta = eta
30     self.n_iter = n_iter
31     self.w_initialized = False
32     self.shuffle = shuffle
33     self.random_state = random_state
34
35 def fit(self, X, y):
36     """ Fit training data.
37
38     Parameters
39     -----
40     X : {array-like}, shape = [n_examples, n_features]
41         Training vectors, where n_examples is the number of
42         examples and n_features is the number of features.
43     y : array-like, shape = [n_examples]
44         Target values.
45
46     Returns
47     -----
48     self : object
49
50     """
51     self._initialize_weights(X.shape[1])

```

```

52     self.losses_ = []
53     for i in range(self.n_iter):
54         if self.shuffle:
55             X, y = self._shuffle(X, y)
56             losses = []
57             for xi, target in zip(X, y):
58                 losses.append(self._update_weights(xi, target))
59             avg_loss = np.mean(losses)
60             self.losses_.append(avg_loss)
61     return self
62
63 def partial_fit(self, X, y):
64     """Fit training data without reinitializing the weights"""
65     if not self.w_initialized:
66         self._initialize_weights(X.shape[1])
67     if y.ravel().shape[0] > 1:
68         for xi, target in zip(X, y):
69             self._update_weights(xi, target)
70     else:
71         self._update_weights(X, y)
72     return self
73
74 def _shuffle(self, X, y):
75     """Shuffle training data"""
76     r = self.rgen.permutation(len(y))
77     return X[r], y[r]
78
79 def _initialize_weights(self, m):
80     """Initialize weights to small random numbers"""
81     self.rgen = np.random.RandomState(self.random_state)
82     self.w_ = self.rgen.normal(loc=0.0, scale=0.01,
83                               size=m)
84     self.b_ = np.float_(0.)
85     self.w_initialized = True
86
87 def _update_weights(self, xi, target):
88     """Apply Adaline learning rule to update the weights"""
89     output = self.activation(self.net_input(xi))
90     error = (target - output)
91     self.w_ += self.eta * 2.0 * xi * (error)
92     self.b_ += self.eta * 2.0 * error
93     loss = error**2
94     return loss
95
96 def net_input(self, X):
97     """Calculate net input"""
98     return np.dot(X, self.w_) + self.b_
99
100 def activation(self, X):
101     """Compute linear activation"""
102     return X
103

```

```

104 def predict(self, X):
105     """Return class label after unit step"""
106     return np.where(self.activation(self.net_input(X))
107                     >= 0.5, 1, 0)

```

Code 2.13: Adaline Implemented Using Stochastic Gradient Descent

The `_shuffle` method uses `np.random.permutation` to generate a random sequence of unique indices, which are then applied to shuffle the feature matrix and class-label vector.

Observation 2.10.7

Display the plot with labeled axes using `plt.show()`.

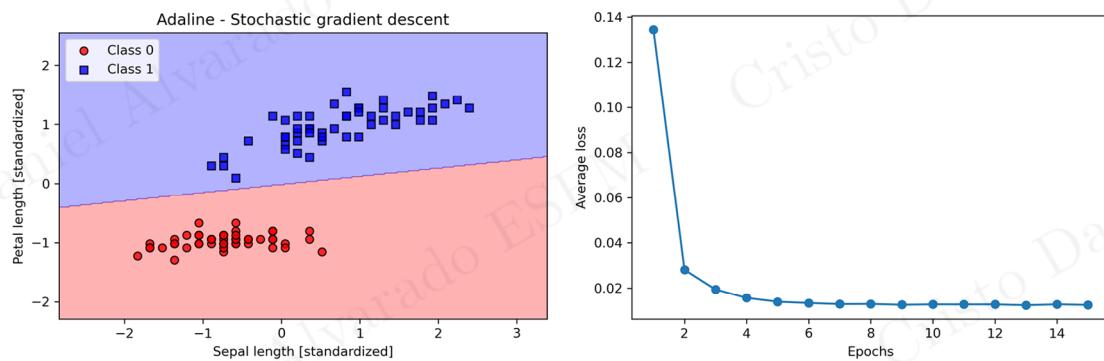


Figure 2.11: Adaline Implementation using Stochastic Gradient Descent

The *average loss decreases quickly, and the final decision boundary after 15 epochs is similar to the batch gradient descent Adaline.*

Idea 2.10.2 (Updating the Model in an Online Scenario)

To update the model in an online learning scenario with streaming data, call `ada_sgd.partial_fit(X_std[0, :], y[0])` on individual training examples.

CHAPTER 3

A TOUR OF MACHINE LEARNING CLASSIFIERS USING SCIKIT-LEARN

3.1 INTRODUCTION

The goal of this chapter is to use Scikit-Learn to implement some classifiers in order to learn how to use them, his advantages, disadvantages and different use scenarios.

Specifically, we will take a look of 4 popular machine learning models commonly used in academia and in industry. In addition, we will take a look at the Scikit-Learn library, which offers a user-friendly and consistent interface for using those algorithms efficiently and productively.

3.1.1 CHOOSING A CLASSIFICATION ALGORITHM

Each algorithm has its own quirks and relies on certain assumptions. no single classifier works best across all possible scenarios (The Lack of A Priori Distinctions Between Learning Algorithms, Wolpert, David H, Neural Computation 8.7 (1996): 1341-1390).

Observation 3.1.1 (Comparasion)

In practice, it is always recommended to compare the behaviour of different algorithms in order to find the best model suitable for a particular problem; these may differ in the number of features or examples, the amount of noise in a dataset, and whether the classes are linearly separable.

Idea 3.1.1

The performance of a classifier relies upon the data that is available for learning. The five main steps that are involved in training a supervised machine learning algorithm can be summarized as follows:

1. Selecting features and collecting labeled training examples
2. Choosing a performance metric
3. Choosing a learning algorithm and training a model
4. Evaluating the performance of the model
5. Changing the settings of the algorithm and tuning the model.

We will mainly focus on the main concepts of the different algorithms in this chapter and revisit topics such as feature selection and preprocessing, performance metrics, and hyperparameter tuning for more detailed discussions later in the book.

3.1.2 FIRST STEPS WITH SCIKIT-LEARN TRAINING A PERCEPTRON

Before we learn about two related learning algorithms: the perceptron and adaline, both implemented in Python using NumPy and other libraries by ourselves.

Now we will take a look at the **scikit-learn API**.

Observation 3.1.2

One of the advantages of using scikit-learn is that combines a user-friendly and consistent interface with a highly optimized implementation of several classification algorithms. Also, this library offers not only a large variety of learning algorithms, but also many convenient functions to preprocess data and to fine-tune and evaluate our models.

3.1.3 TRAINING A MODEL USING SCIKIT-LEARN

To get started with the scikit-learn library, we will train a perceptron model similar to the one that we implemented in Chapter 2. For simplicity, we will use the already familiar Iris dataset throughout the following sections.

Observation 3.1.3 (Iris Dataset and Its Uses)

Conveniently, the Iris dataset is already available via scikit-learn, since it is a simple yet popular dataset that is frequently used for testing and experimenting with algorithms. Similar to the previous chapter, we will only use two features from the Iris dataset for visualization purposes.

We will assign the petal length and petal width of the 150 flower examples to the feature matrix, **X**, and the corresponding class labels of the flower species to the vector array, **y**:

```
1 Class labels: [0 1 2]
2 Class labels: [0 1 2]
```

Code 3.1: Class Labels of Matrix X

The `np.unique(y)` function returned the three unique class labels stored in `iris.target`, and as we can see, the Iris flower class names, `Iris-setosa`, `Iris-versicolor`, and `Iris-virginica`, are already stored as integers (here: 0, 1, 2).

Observation 3.1.4 (Scikit-Learn with Class Labels and String Formats)

Although many scikit-learn functions and class methods also work with *class labels in string format*, *using integer labels is a recommended approach to avoid technical glitches and improve computational performance due to a smaller memory footprint*; furthermore, *encoding class labels as integers is a common convention among most machine learning libraries*.

To evaluate how well a trained model performs on unseen data, we will *further split the dataset into separate training and test datasets*.

Idea 3.1.2 (More Info About Best Practices around Model Evaluation)

In Chapter 6, *Learning Best Practices for Model Evaluation and Hyperparameter Tuning*, we will discuss the best practices around model evaluation in more detail.

Using the `train_test_split` function from scikit-learn's `model_selection` module, we randomly split the `X` and `y` arrays into 30 percent test data (45 examples) and 70 percent training data (105 examples):

```
1 train_test_split(X, y, test_size=0.3, random_state=1, stratify=y)
```

Code 3.2: Train Test Split Function

Observation 3.1.5

Note that the `train_test_split` function already shuffles the training datasets internally before splitting; otherwise, all examples from class 0 and class 1 would have ended up in the training datasets, and the test dataset would consist of 45 examples from class 2. Via the `random_state` parameter, we provided a fixed random seed (`random_state=1`) for the internal pseudo-random number generator that is used for shuffling the datasets prior to splitting. Using such a fixed `random_state` ensures that our results are reproducible.

Lastly, we took advantage of the built-in support for stratification via `stratify=y`. In this context, *stratification means that the `train_test_split` method returns training and test subsets that have the same proportions of class labels as the input dataset*. We can use NumPy's `bincount` function, which counts the number of occurrences of each value in an array, to verify that this is indeed the case:

```
1 >>> print('Labels counts in y:', np.bincount(y))
2 Labels counts in y: [50 50 50]
3 >>> print('Labels counts in y_train:', np.bincount(y_train))
4 Labels counts in y_train: [35 35 35]
5 >>> print('Labels counts in y_test:', np.bincount(y_test))
6 Labels counts in y_test: [15 15 15]
```

Code 3.3: Use of `bincount` Method in Python.

Idea 3.1.3 (Feature Scaling)

Many machine learning and optimization algorithms also require **feature scaling for optimal performance**, as we saw in the gradient descent example in Chapter 2. Here, we will standardize the features using the `StandardScaler` class from scikit-learn's `preprocessing` module:

```
1 sc = StandardScaler()
```

Code 3.4: `StandardScaler` Method in Python.

Using the preceding code, we loaded the `StandardScaler` class from the `preprocessing` module and initialized a new `StandardScaler` object that we assigned to the `sc` variable.

Using the `fit` method, `StandardScaler` estimated the parameters, μ (sample mean) and σ (standard deviation), for each feature dimension from the training data. By calling the `transform` method, we then standardized the training data using those estimated parameters, μ and σ . Note that we used the same scaling parameters to standardize the test dataset so that both the values in the training and test dataset are comparable with one another.

Example 3.1.1

Having standardized the training data, we can now train a perceptron model. Most algorithms in scikit-learn already support multiclass classification by default via the one-versus-rest (OvR) method, which allows us to feed the three flower classes to the perceptron all at once.

Finally, we have the code for the model:

```
1 from sklearn import datasets #To import Iris Dataset
2 import matplotlib.pyplot as plt
3 from matplotlib.colors import ListedColormap
4 import numpy as np
5 import sklearn.linear_model as linear_models
6 from sklearn.preprocessing import StandardScaler
7 from sklearn.model_selection import train_test_split
8
9 #Plot Regions
10
11 def plot_decision_regions(X, y, classifier, test_idx=None,
12     resolution=0.2):
13     markers = ('o', 's', '^', 'v', '<')
14     colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
15     cmap = ListedColormap(colors[:len(np.unique(y))])
16
17     x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
18     x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
19     xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),
20                           np.arange(x2_min, x2_max, resolution))
21
22     lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
23                             .T)
23     lab = lab.reshape(xx1.shape)
24     plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
25     plt.xlim(xx1.min(), xx1.max())
26     plt.ylim(xx2.min(), xx2.max())
27
28     for idx, cl in enumerate(np.unique(y)):
29         plt.scatter(x=X[y == cl, 0],
30                     y=X[y == cl, 1],
31                     alpha=0.8,
32                     c=colors[idx],
33                     marker=markers[idx],
34                     label=f'Class {cl}',
35                     edgecolor='black')
36
37     if test_idx is not None:
38         X_test, y_test = X[test_idx, :], y[test_idx]
39         plt.scatter(X_test[:, 0], X_test[:, 1],
40                     c='none', edgecolor='black', alpha=1.0,
41                     linewidth=1, marker='o',
42                     s=100, label='Test set')
```

```

42 #Import Data
43
44 iris = datasets.load_iris()
45 X = iris.data[:,2:4]
46 y = iris.target
47
48 #Scale Standard Data
49
50 sc = StandardScaler()
51 X = sc.fit_transform(X)
52
53 #Import Perceptron
54
55 clf = linear_models.Perceptron(eta0=0.1, random_state=1)
56
57 # Split training data and test data
58 indices = np.arange(len(X))
59 X_train, X_test, idx_train, idx_test = train_test_split(
60     X, indices, test_size=0.3, random_state=1
61 )
62 y_train = y[idx_train]
63 y_test = y[idx_test]
64
65 #Fit model with training data
66
67 clf.fit(X_train, y_train)
68
69 #Plot regions
70
71 plot_decision_regions(X, y, clf, idx_test, 0.2)
72 plt.xlabel(iris.feature_names[2] + " (standarized)")
73 plt.ylabel(iris.feature_names[3] + " (standarized)")
74 plt.legend()
75 plt.show()

```

Code 3.5: Perceptron Trained Model using Sklearn

With the slight modification that we made to the `plot_decision_regions` function, we can now specify the indices of the examples that we want to mark on the resulting plots.

As we can see in the resulting plot, the three flower classes can't be perfectly separated by a linear decision boundary:

However, remember from our discussion in Chapter 2 that **the perceptron algorithm never converges on datasets that aren't perfectly linearly separable**, which is *why the use of the perceptron algorithm is typically not recommended in practice*. In the following sections, we will look at more powerful linear classifiers that converge to a loss minimum even if the classes are not perfectly linearly separable.

Observation 3.1.6 (Note)

The Perceptron, as well as other scikit-learn functions and classes, *often has additional parameters that we omit for clarity. You can read more about those parameters using the help function in Python (for instance, `help(Perceptron)`) or by going through the excellent scikit-learn online*

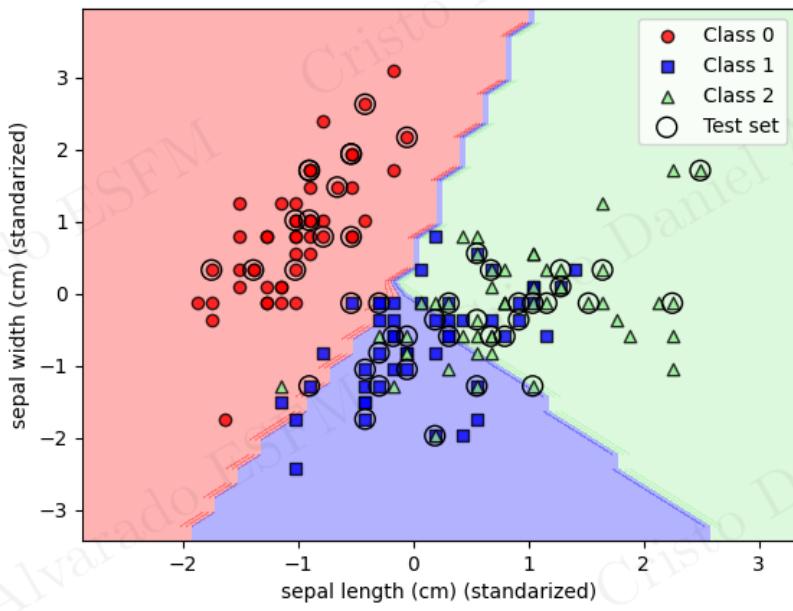


Figure 3.1: Perceptron Trained with Iris Dataset using Scikit-Learn.

documentation [here](#).

3.2 MODELING CLASS PROBABILITIES VIA LOGISTIC REGRESSION

The biggest disadvantage of the Perceptron is that it never converges if the classes aren't linearly separable. The reason for this is that the weights are continuously updated because there is always at least one misclassified training example present in each epoch.

To make better use of our time, we will now look at another simple, yet more powerful, algorithm for linear and binary classification problems: **logistic regression**. Note that, despite its name, *logistic regression is a model for classification, not regression*.

3.2.1 LOGISTIC REGRESSION AND CONDITIONAL PROBABILITIES

Definition 3.2.1 (Logistic Regression)

Logistic regression is a *classification model that is easy to implement and performs very well on linearly separable classes*. It is *one of the most widely used algorithms for classification in industry*.

Similar to the perceptron and Adaline, *the logistic regression model in this section is also a linear model for binary classification*.

Idea 3.2.1 (Note)

Logistic regression can be generalized to multinomial logistic regression.

To explain the main mechanics behind logistic regression as a probabilistic model for binary classification, let's first introduce the odds: the odds in favor of a particular event. The odds can be written as $\frac{p}{1-p}$, where p stands for the probability of the positive event. The term "positive event" does not necessarily mean "good," but refers to the event that we want to predict.

Example 3.2.1 (Conditional Probability)

Consider the **probability that a patient has a certain disease given certain symptoms**; we can think of the positive event as class label $y = 1$ and the symptoms as features \mathbf{x} . Hence, for brevity, we can define the probability p as $p := p(y = 1|\mathbf{x})$, the conditional probability that a particular example belongs to class 1 given its features, \mathbf{x} .

We can then further define the logit function, which is simply the logarithm of the odds (log-odds):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) \quad (3.1)$$

Note that *log refers to the natural logarithm, as it is the common convention in computer science*. The logit function *takes input values in the range 0 to 1 and transforms them into values over the entire real-number range*.

Under the logistic model, we **assume that there is a linear relationship between the weighted inputs and the log-odds**:

$$\text{logit}(p) = w_1x_1 + \cdots + w_nx_n + b = \sum_{i=1}^n w_i x_i + b = \mathbf{w}^T \mathbf{x} + b \quad (3.2)$$

Observation 3.2.1

While the preceding describes an assumption we make about the linear relationship between the log-odds and the net inputs, what we are actually interested in is the probability p , the *class-membership probability of an example given its features*. While the logit function maps the probability to a real-number range, we can consider the inverse of this function to map the real-number range back to a $[0, 1]$ range for the probability p .

This inverse of the logit function is typically called the **logistic sigmoid function**, which is sometimes simply abbreviated to *sigmoid function* due to its characteristic *S-shape*:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \quad \forall z \in \mathbb{R} \quad (3.3)$$

Here, z is the *net input*, the linear combination of weights and the inputs (that is, the features associated with the training examples):

$$z = \mathbf{w}^T \mathbf{x} + b \quad (3.4)$$

Example 3.2.2 (Graph of Sigmoid Function)

Graph of Sigmoid Function:

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 def sigmoid(z):
5     return 1.0 / (1.0 + np.exp(-z))
6
7 z = np.arange(-7, 7, 0.1)
8 sigma_z = sigmoid(z)
9 plt.plot(z, sigma_z)
10 plt.axvline(0.0, color='k')
11 plt.ylim(-0.1, 1.1)
12 plt.xlabel('z')
13 plt.ylabel('$\sigma(z)$')
14 # y-axis ticks and gridline
15 plt.yticks([0.0, 0.5, 1.0])
16 ax = plt.gca()
17 ax.yaxis.grid(True)
18 plt.tight_layout()
19 plt.show()
```

Code 3.6: Sigmoid Function Graph.

As a result of executing the previous code, you should now see the *S-shaped* (*sigmoidal*) curve.

Observation 3.2.2 (Comparasion with Adaline)

To build some understanding of the logistic regression model, we can relate it to Adaline. In Adaline, we used the identity function, $\sigma(z) = z$ as the activation function. In logistic regression, this activation function simply becomes the sigmoid function defined earlier.

The only difference between Adaline and logistic regression is the activation function.

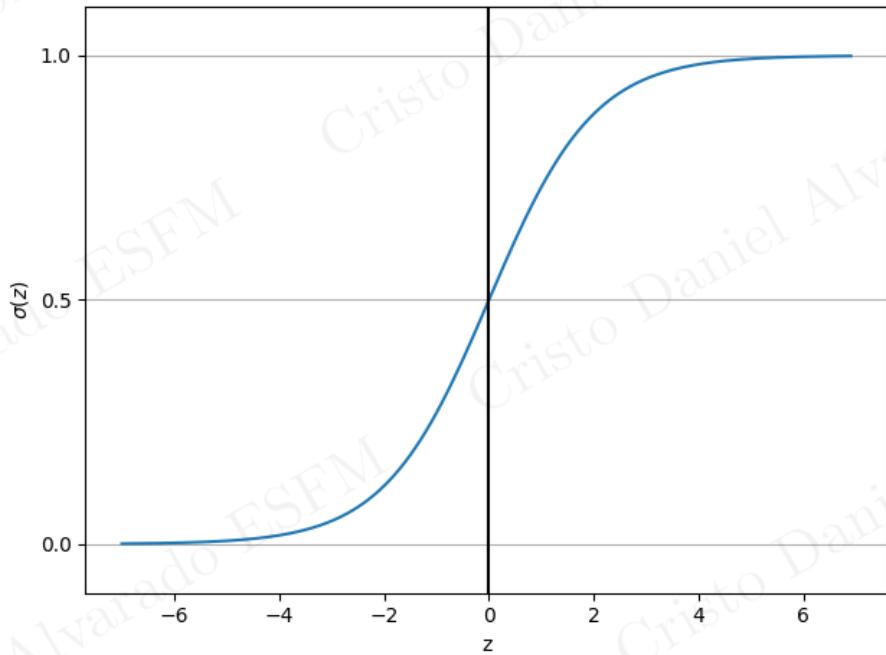


Figure 3.2: Sigmoid Function Graph.

Definition 3.2.2 (Purpose of an Activation Function)

In simple terms, the purpose of an **activation function** is to decide whether a neuron in a neural network should be "activated" or not, and to what degree. It's what allows neural networks to learn and model complex, non-linear relationships.

The output of the sigmoid function is then interpreted as the probability of a particular example belonging to class 1, $\sigma(z) = p(y = 1|\mathbf{x}; \mathbf{w}, b)$, given its features \mathbf{x} , and parameterized by the weights and bias, \mathbf{w} and b .

For example, if we compute $\sigma(z) = 0.8$ for a particular flower, it means that the chance that this example is an **Iris-versicolor** flower is 80 percent. Therefore, the probability that this flower is an **Iris-setosa** can be calculated as $p(y = 0|\mathbf{x}; \mathbf{w}, b) = 1 - p(y = 1|\mathbf{x}; \mathbf{w}, b) = 0.2$, or 20 percent.

The predicted probability can then be converted into a binary outcome via a threshold function:

$$\hat{y} = \begin{cases} 1 & \text{if } \sigma(z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

If we look at the preceding plot of the sigmoid function, this is equivalent to the following:

$$\hat{y} = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Observation 3.2.3 (Interest in Predicted Class Labels)

In many applications, we are not only interested in the predicted class labels but also in the class-membership probabilities produced by the sigmoid function before applying the threshold.

Logistic regression is used in weather forecasting, for example, not only to predict whether it will rain on a particular day but also to report the chance of rain. Similarly, logistic regression can be used to predict the chance that a patient has a particular disease given certain symptoms,

which is why logistic regression enjoys great popularity in the field of medicine.

3.3 LEARNING THE MODEL WEIGHTS VIA THE LOGISTIC LOSS FUNCTION

You have learned how we can use the logistic regression model to predict probabilities and class labels; now, let's briefly talk about how we fit the parameters of the model, for instance, the weights and bias unit, \mathbf{w} and b . Previously, we defined the mean squared error loss function as follows:

$$L(\mathbf{w}, b | \mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (y^{(i)} - \sigma(z^{(i)}))^2$$

We minimized this function in order to learn the parameters for our Adaline classification model. To explain how we can derive the loss function for logistic regression, let's first define the likelihood, \mathcal{L} , that we **want to maximize** when we build a **logistic regression model**, assuming that the individual examples in our dataset are independent of one another. The formula is as follows:

$$L(w, b | x) = p(y|x; w, b) = \prod_{i=1}^n p(y^{(i)} | x^{(i)}; w, b) = \prod_{i=1}^n (\sigma(z^{(i)}))^{y^{(i)}} (1 - \sigma(z^{(i)}))^{(1-y^{(i)})} \quad (3.5)$$

This is the **likelihood function** $L(w, b | x)$ in terms of a product over n terms. It is expressed as the product of probabilities $p(y^{(i)} | x^{(i)}; w, b)$ for i from 1 to n , which is further expanded to the product of $(\sigma(z^{(i)}))^{y^{(i)}}$ times $(1 - \sigma(z^{(i)}))^{(1-y^{(i)})}$, where σ is the sigmoid function.

Observation 3.3.1

In practice, it is easier to maximize the (natural) log of this equation, which is called the log-likelihood function:

$$l(w, b | x) = \log(\mathcal{L}(w, b | x)) = \sum_{i=1}^n [y^{(i)} \log(\sigma(z^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(z^{(i)}))] \quad (3.6)$$

Applying the log function reduces the potential for numerical underflow if the likelihoods are very small. In addition, the product of factors becomes a summation of factors, which makes it easier to obtain the derivative of this function via the addition trick.

3.3.1 DERIVING THE LIKELIHOOD FUNCTION

We can obtain the expression for the likelihood of the model given the data, $\mathcal{L}(w, b | x)$, as follows. Given that we have a binary classification problem with class labels 0 and 1, we can think of the label 1 as a Bernoulli variable—it can take on two values, 0 and 1, with the probability p of being 1: $Y \sim \text{Bernoulli}(p)$.

For a single data point, we can write this probability as:

$$P(Y = 1 | X = x^{(i)}) = \sigma(z^{(i)}) \quad (3.7)$$

$$P(Y = 0 | X = x^{(i)}) = 1 - \sigma(z^{(i)}) \quad (3.8)$$

Putting these two expressions together, and using the shorthand $P(Y = y^{(i)} | X = x^{(i)}) = p(y^{(i)} | x^{(i)})$, we get the probability mass function of the Bernoulli variable:

$$p(y^{(i)} | x^{(i)}) = (\sigma(z^{(i)}))^{y^{(i)}} (1 - \sigma(z^{(i)}))^{1-y^{(i)}}$$

Substituting the probability mass function of the Bernoulli variable, we arrive at the expression for the likelihood, which we attempt to maximize by changing the model parameters:

$$\mathcal{L}(w, b | x) = \prod_{i=1}^n (\sigma(z^{(i)}))^{y^{(i)}} (1 - \sigma(z^{(i)}))^{1-y^{(i)}}$$

Observation 3.3.2

We could use an optimization algorithm such as **gradient ascent** to maximize this log-likelihood function. (*Gradient ascent works the same way as gradient descent explained earlier, except that gradient ascent maximizes a function instead of minimizing it.*).

Alternatively, let's rewrite the log-likelihood as a loss function, L , that can be minimized using gradient descent:

$$L(w, b) = \sum_{i=1}^n [-y^{(i)} \log(\sigma(z^{(i)})) - (1 - y^{(i)}) \log(1 - \sigma(z^{(i)}))]$$

Example 3.3.1

For one single training example, looking at the equation, we can see that the first term becomes zero if $y = 0$, and the second term becomes zero if $y = 1$:

$$L(\sigma(z), y, w, b) = \begin{cases} -\log(\sigma(z)) & \text{if } y = 0 \\ -\log(1 - \sigma(z)) & \text{if } y = 1 \end{cases}$$

Let's write a short code snippet to create a plot that illustrates the loss of classifying a single training example for different values of $\sigma(z)$. The resulting plot shows the sigmoid activation on the x axis in the range 0 to 1 (the inputs to the sigmoid function were z values in the range -10 to 10) and the associated logistic loss on the y axis:

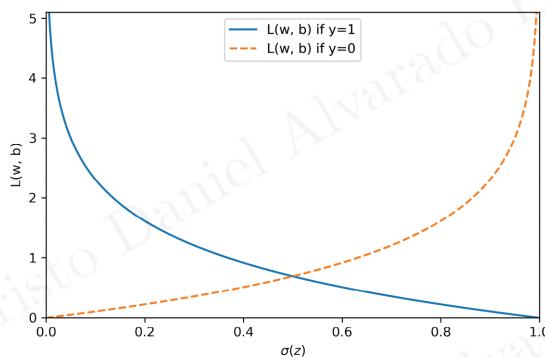


Figure 3.3: Loss Function.

We can see that the loss approaches 0 (continuous line) if we correctly predict that an example belongs to class 1. Similarly, the loss also approaches 0 if we correctly predict $y = 0$ (dashed line). However, if the prediction is wrong, the loss goes toward infinity. The main point is that we penalize incorrect predictions with an increasingly larger loss.

3.4 CONVERTING AN ADALINE IMPLEMENTATION INTO AN ALGORITHM FOR LOGISTIC REGRESSION

If we implement logistic regression from scratch, we can substitute the loss function, L , in the earlier Adaline implementation with the new loss function:

$$L(w, b) = \frac{1}{n} \sum_{i=1}^n [-y^{(i)} \log(\sigma(z^{(i)})) - (1 - y^{(i)}) \log(1 - \sigma(z^{(i)}))]$$

Observation 3.4.1

We use this to compute the loss of classifying all training examples per epoch. We also swap the linear activation function for the sigmoid. With those changes, the Adaline code becomes a working logistic regression implementation.

The following example uses full-batch gradient descent (the same changes can be applied to the stochastic variant):

```
1 import numpy as np
2
3 class LogisticRegressionGD:
4     """Gradient descent-based logistic regression classifier.
5
6     Parameters
7     -----
8     eta : float
9         Learning rate (between 0.0 and 1.0)
10    n_iter : int
11        Passes over the training dataset.
12    random_state : int
13        Random number generator seed for random weight
14        initialization.
15
16    Attributes
17    -----
18    w_ : 1d-array
19        Weights after training.
20    b_ : Scalar
21        Bias unit after fitting.
22    losses_ : list
23        Mean squared error loss function values in each epoch.
24    """
25    def __init__(self, eta=0.01, n_iter=50, random_state=1):
26        self.eta = eta
27        self.n_iter = n_iter
28        self.random_state = random_state
29
30    def fit(self, X, y):
31        """ Fit training data.
32
33        Parameters
```

```

34 -----
35     X : {array-like}, shape = [n_examples, n_features]
36         Training vectors, where n_examples is the
37         number of examples and n_features is the
38         number of features.
39     y : array-like, shape = [n_examples]
40         Target values.
41
42     Returns
43 -----
44     self : Instance of LogisticRegressionGD
45     """
46     rgen = np.random.RandomState(self.random_state)
47     self.w_ = rgen.normal(loc=0.0, scale=0.01, size=X.shape
48     [1])
49     self.b_ = np.float_(0.)
50     self.losses_ = []
51     for i in range(self.n_iter):
52         net_input = self.net_input(X)
53         output = self.activation(net_input)
54         errors = (y - output)
55         self.w_ += self.eta * 2.0 * X.T.dot(errors) / X.
56             shape[0]
57         self.b_ += self.eta * 2.0 * errors.mean()
58         loss = (-y.dot(np.log(output))
59             - ((1 - y).dot(np.log(1 - output)))
60                 / X.shape[0])
61         self.losses_.append(loss)
62     return self
63
64     def net_input(self, X):
65         """Calculate net input"""
66         return np.dot(X, self.w_) + self.b_
67
68     def activation(self, z):
69         """Compute logistic sigmoid activation"""
70         return 1. / (1. + np.exp(-np.clip(z, -250, 250)))
71
72     def predict(self, X):
73         """Return class label after unit step"""
74         return np.where(self.activation(self.net_input(X)) >=
75             0.5, 1, 0)

```

Code 3.7: Implementation of Logistic Regression in Python.

When fitting a logistic regression model, remember that it only works for binary classification tasks.

The resulting decision region plot looks as follows:

3.5 GRADIENT DESCENT

One observation we can make on the latter code is that the partial derivative of the loss function L is really simple to compute, this is due to the fact:

$$\frac{\partial L}{\partial w_j} = \frac{\partial L}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial z} \cdot \frac{\partial z}{\partial w_j}$$

where:

$$\begin{aligned}\frac{\partial L}{\partial \sigma} &= \frac{\sigma - y}{\sigma \cdot (1 - \sigma)} \\ \frac{\partial \sigma}{\partial z} &= \sigma(1 - \sigma) \\ \frac{\partial z}{\partial w_j} &= x_j\end{aligned}$$

So,

$$\frac{\partial L}{\partial w_j} = -(y - \sigma)x_j$$

Recall that gradient descent takes steps in the opposite direction of the gradient. Hence, we flip $\nabla L(w)$ and update the j-th weight as follows, including the learning rate η :

$$w_j = w_j - \eta \cdot \frac{\partial L}{\partial w_j}(w)$$

While the partial derivative of the loss function with respect to the bias unit is not shown, bias derivation follows the same concept.

Observation 3.5.1

We should compute the mean in all this process, but due we are working in this example with a single record for simplicity.

Idea 3.5.1 (Relation between Adaline and Logistical Regression)

Both the weight and bias updates are the same as for Adaline, but the change is in the activation function.

3.6 TRAINING A LOGISTIC REGRESSION MODEL WITH SCIKIT-LEARN

We just went through useful coding and math exercises in the previous subsection, which helped to illustrate the conceptual differences between Adaline and logistic regression. Now, let's learn how to use scikit-learn's more optimized implementation of logistic regression, which also supports multiclass settings off the shelf.

Observation 3.6.1

Note that in recent versions of scikit-learn, the technique used for multiclass classification—multinomial or OvR—is chosen automatically.

In the following code example, we will use the `sklearn.linear_model.LogisticRegression` class as well as the familiar `fit` method to train the model on all three classes in the standardized flower training dataset. Also, we set `multi_class='ovr'` for illustration purposes. As an exercise, you may want to compare the results with `multi_class='multinomial'`. Note that the `multinomial` setting is now the default choice in scikit-learn's `LogisticRegression` class and recommended in practice for mutually exclusive classes, such as those found in the Iris dataset.

Observation 3.6.2

Here, "mutually exclusive" means that each training example can belong to only a single class (in contrast to multilabel classification, where a training example can be a member of multiple classes).

Now, let's look at the code example:

```
1 from sklearn.linear_model import LogisticRegression
2 import numpy as np
3 import matplotlib.pyplot as plt
4 from matplotlib.colors import ListedColormap
5 from sklearn import datasets #To import Iris Dataset
6 from sklearn.preprocessing import StandardScaler
7 from sklearn.model_selection import train_test_split
8
9 def plot_decision_regions(X, y, classifier, test_idx=None,
10    resolution=0.2):
11    markers = ('o', 's', '^', 'v', '<')
12    colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
13    cmap = ListedColormap(colors[:len(np.unique(y))])
14
15    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
16    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
17    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),
18                           np.arange(x2_min, x2_max, resolution))
19
20    lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
21                            .T)
22    lab = lab.reshape(xx1.shape)
23    plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
24    plt.xlim(xx1.min(), xx1.max())
25    plt.ylim(xx2.min(), xx2.max())
26
27    for idx, cl in enumerate(np.unique(y)):
28        plt.scatter(x=X[y == cl, 0],
29                    y=X[y == cl, 1],
30                    alpha=0.8,
31                    c=colors[idx],
32                    marker=markers[idx],
33                    label=f'Class {cl}',
34                    edgecolor='black')
35
36    if test_idx is not None:
37        X_test, y_test = X[test_idx, :], y[test_idx]
38        plt.scatter(X_test[:, 0], X_test[:, 1],
39                    c='none', edgecolor='black', alpha=1.0,
40                    linewidth=1, marker='o',
41                    s=100, label='Test set')
42
43    #Import Data
```

```

41
42 iris = datasets.load_iris()
43 X = iris.data[:, 2:4]
44 y = iris.target
45
46 #Scale Standard Data
47
48 sc = StandardScaler()
49 X = sc.fit_transform(X)
50
51 # Split training data and test data
52 indices = np.arange(len(X))
53 X_train, X_test, idx_train, idx_test = train_test_split(
54     X, indices, test_size=0.3, random_state=1
55 )
56 y_train = y[idx_train]
57 y_test = y[idx_test]
58
59 lr = LogisticRegression(C=100.0, solver='lbfgs',
60                         multi_class='ovr')
61 lr.fit(X_train, y_train)
62 plot_decision_regions(X, y, lr, idx_test, 0.01)
63
64 plt.xlabel(iris.feature_names[2] + " (standarized)")
65 plt.ylabel(iris.feature_names[3] + " (standarized)")
66 plt.legend(loc='upper left')
67 plt.tight_layout()
68 plt.show()

```

Code 3.8: Logistic Regression Iris Dataset Python.

Which produces an output like this:

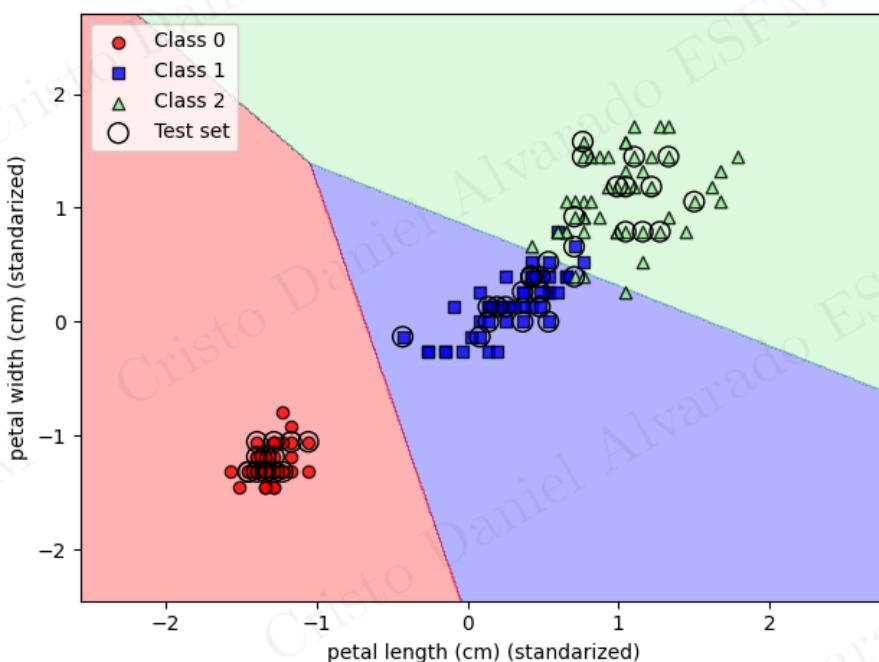


Figure 3.4: Logistic Regression on the Iris Dataset.

3.6.1 ALGORITHMS FOR CONVEX OPTIMIZATION

Many different algorithms exist for solving optimization problems. For minimizing convex loss functions, such as the logistic regression loss, it is recommended to use more advanced approaches than regular **stochastic gradient descent (SGD)**. Scikit-learn implements a range of such optimization algorithms, which can be specified via the solver parameter: '`newton-cg`', '`lbfgs`', '`liblinear`', '`sag`', and '`saga`'.

While the logistic regression loss is convex, most optimization algorithms should converge to the global loss minimum with ease. However, there are certain advantages to using one algorithm over another. For example, in earlier versions (for instance, v0.21), scikit-learn used '`liblinear`' as the default, which cannot handle the multinomial loss and is limited to the OvR scheme for multiclass classification. In scikit-learn v0.22, the default solver was changed to '`lbfgs`', which stands for the **limited-memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm** and is more flexible in this regard.

Observation 3.6.3

Looking at the preceding code that we used to train the `LogisticRegression` model, you might now be wondering, "What is this mysterious parameter `C`?" We will discuss this parameter in the next subsection, where we will introduce the concepts of overfitting and regularization. However, before we move on to those topics, let's finish our discussion of class membership probabilities.

The probability that training examples belong to a certain class can be computed using the `predict_proba` method.

Example 3.6.1

For example, we can predict the probabilities of the first three examples in the test dataset as follows:

```
1 lr.predict_proba(X_test[:3, :])
```

Code 3.9: Predicting Probabilities.

This code snippet returns the following array:

```
1 array([[3.81527885e-09, 1.44792866e-01, 8.55207131e-01],
2         [8.34020679e-01, 1.65979321e-01, 3.25737138e-13],
3         [8.48831425e-01, 1.51168575e-01, 2.62277619e-14]])
```

The first row corresponds to the class membership probabilities of the first flower, the second row corresponds to the second flower, and so forth. Notice that the column-wise sum in each row is 1, as expected (you can confirm this by executing `lr.predict_proba(X_test_std[:3, :]).sum(axis=1)`).

The highest value in the first row is approximately 0.85, which means that the first example belongs to class 3 (`Iris-virginica`) with a predicted probability of 85 percent.

Observation 3.6.4

As you may have noticed, we can obtain the predicted class labels by identifying the largest column in each row, for example, by using NumPy's `argmax` function:

```
1 lr.predict_proba(X_test_std[:3, :]).argmax(axis=1)
```

The returned class indices (corresponding to `Iris-virginica`, `Iris-setosa`, and `Iris-setosa`) are:

```
1 array([2, 0, 0])
```

In the preceding code example, we converted conditional probabilities into class labels manually by using NumPy's `argmax` function. In practice, the more convenient way of obtaining class labels when using scikit-learn is to call the `predict` method directly:

```
1 lr.predict(X_test_std[:3, :])
2 array([2, 0, 0])
```

Lastly, if you want to predict the class label of a single flower example, scikit-learn expects a two-dimensional array as input. One way to convert a single row entry into a two-dimensional data array is to use NumPy's `reshape` method to add a new dimension, as shown here:

```
1 lr.predict(X_test[0, :].reshape(1, -1))
2 array([2])
```

3.7 TACKLING OVERFITTING VIA REGULARIZATION

Definition 3.7.1 (Overfitting)

Overfitting is a *common problem in machine learning, where a model performs well on training data but does not generalize well to unseen data (test data).*

If a model suffers from overfitting, we also say that the model has **high variance**, which can be caused by having too many parameters, leading to a model that is too complex given the underlying data.

Observation 3.7.1 (Underfitting)

Similarly, a model can suffer from **underfitting (high bias)**, which means *it is not complex enough to capture the pattern in the training data well and therefore also performs poorly on unseen data.*

Although we have only encountered linear models for classification so far, the *problems of overfitting and underfitting can be illustrated by comparing a linear decision boundary to more complex, nonlinear decision boundaries.*

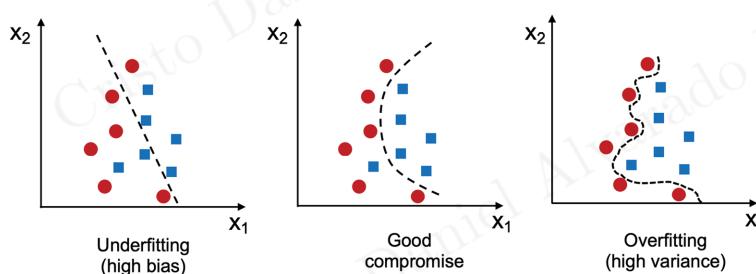


Figure 3.5: Examples of Underfitting, Good Compromise and Overfitting.

3.7.1 THE BIAS-VARIANCE TRADEOFF

Definition 3.7.2 (Tradeoff)

Tradeoff is a *balance achieved between two desirable but incompatible features; a compromise.*

Researchers often use the terms **bias** and **variance** or **bias-variance tradeoff** to describe model performance—that is, you may encounter talks, courses, or articles where people say that a model has high variance or high bias.

In general, *high variance is proportional to overfitting and high bias is proportional to underfitting.*

Observation 3.7.2 (Variance in Machine Learning)

In the context of machine learning models, *variance measures the consistency (or variability) of the model prediction for classifying a particular example if we retrain the model multiple times on different subsets of the training dataset.*

In contrast, *bias measures how far off the predictions are from the correct values in general if we rebuild the model multiple times on different training datasets; bias is the measure of the systematic error that is not due to randomness.*

One way of finding a good **bias-variance tradeoff** is to tune the complexity of the model via regularization.

Definition 3.7.3 (Regularization)

Regularization is a useful method for handling collinearity (high correlation among features), filtering out noise from data, and eventually preventing overfitting.

The concept behind regularization is to *introduce additional information to penalize extreme parameter (weight) values.*

Observation 3.7.3 (L2 Regularization)

The most common form of regularization is **L2 regularization** (sometimes called L2 shrinkage or weight decay), which can be written as follows:

$$\frac{\lambda}{2n} \|w\|^2 = \frac{\lambda}{2n} \sum_{j=1}^m w_j^2$$

Here, λ is the **regularization parameter**. The 2 in the denominator is merely a scaling factor, such that cancels when computing the loss gradient. The sample size n added to scale the regularization term similarly to the loss.

Observation 3.7.4 (Importance of Feature Scaling)

Regularization is another reason why feature scaling such as standaraization is important. For regularization to work properly, all features must be on comparable scales.

The loss function for logistic regression can be regularized by adding a regularization term, which shrinks the weights during model training:

$$\frac{\partial L}{\partial w_j} = \frac{1}{n} \sum_{i=1}^n \left(\sigma(w^T x^{(i)} - y^{(i)}) x_j^{(i)} \right)$$

by adding the regularization term to the loss changes the partial derivative to the following form:

$$\frac{\partial L}{\partial w_j} = \frac{1}{n} \sum_{i=1}^n \left(\sigma(w^T x^{(i)} - y^{(i)}) x_j^{(i)} \right) + \frac{\lambda}{n} w_j$$

Observation 3.7.5 (Information About the Regularization Parameter λ)

Vía the regularization parameter λ we can control how closely we fit the training data while keeping the weights small. By increasing λ , we increase the regularization strength. Note that the bias unit, essentially an intercept term or negative threshold, is usually not regularized.

3.7.2 IMPLEMENTATION OF REGULARIZATION

In the following example, the parameter `c` implemented for the `LogisticRegression` class in scikit-learn comes from a convention in support vector machines. The term `c` is inversely proportional to the regularization parameter λ . Consequently, decreasing `c` increases the value of λ and therefore, increases the regularization strength.

Observation 3.7.6

We can visualize this behaviour by plotting the L_2 regularization path for the two weight coefficients.

```

1  from sklearn.linear_model import LogisticRegression
2  import numpy as np
3  import matplotlib.pyplot as plt
4  from matplotlib.colors import ListedColormap
5  from sklearn import datasets #To import Iris Dataset
6  from sklearn.preprocessing import StandardScaler
7  from sklearn.model_selection import train_test_split
8
9  #Import Data
10
11 iris = datasets.load_iris()
12 X = iris.data[:,2:4]
13 y = iris.target
14
15 #Scale Standard Data
16
17 sc = StandardScaler()
18 X = sc.fit_transform(X)
19
20 # Split training data and test data
21 indices = np.arange(len(X))
22 X_train, X_test, idx_train, idx_test = train_test_split(
23     X, indices, test_size=0.3, random_state=1
24 )
25 y_train = y[idx_train]
26 y_test = y[idx_test]
```

```

27
28 weights, params = [], []
29 for c in np.arange(-5, 5):
30     lr = LogisticRegression(
31         C=10.**c,
32         solver='lbfgs',
33         max_iter=1000
34     )
35     lr.fit(X_train, y_train)
36     weights.append(lr.coef_[1])
37     params.append(10.**c)
38 weights = np.array(weights)
39 plt.plot(params, weights[:, 0],
40           label='Petal length')
41 plt.plot(params, weights[:, 1], linestyle='--',
42           label='Petal width')
43 plt.ylabel('Weight coefficient')
44 plt.xlabel('C')
45 plt.legend(loc='upper left')
46 plt.xscale('log')
47 plt.show()

```

Code 3.10: Logistic Regression Regularized with different values of `c` on the Iris Dataset Python.

Executing this code fits 10 logistic regression models with different values for the inverse regularization parameter `c`. For illustration purposes, we only collected the weight coefficients of class 1 (the second class on the iris dataset, Iris-versicolor) versus all classifiers.

Observation 3.7.7

Here, we are using OvR technique for multiclass classification.

As shown in the resulting plot, the weight coefficients shrink when `c` decreases—that is, when the regularization strength increases:

We have to keep in mind that:

Increasing the regularization strength **can reduce overfitting**, but if *the strength is too high and the weight coefficients approach zero*, the model **can perform very poorly due to underfitting**.

3.8 MAXIMUM MARGIN CLASSIFICATION WITH SUPPORT VECTOR MACHINES

Another powerful and widely used learning algorithm is the **support vector machine (SVM)**, which can be considered an extension of the perceptron.

Observation 3.8.1 (Key Idea of Support Vector Machines)

Using the perceptron algorithm, we minimized misclassification errors. In SMVs, our optimization objective is to **maximize** the margin.

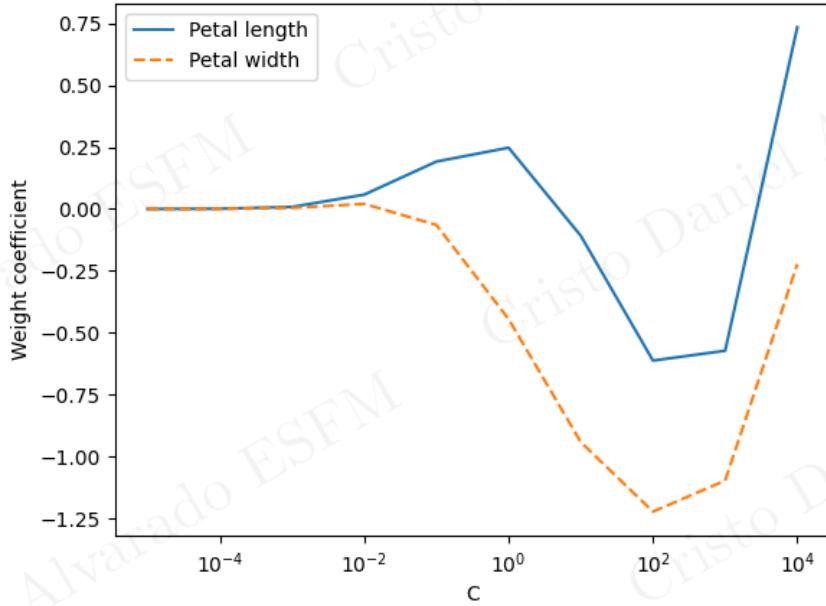


Figure 3.6: Weights for Different Values of C .

Definition 3.8.1 (Margin)

The **margin** is defined as *the distance between the separating hyperplane (called desicion boundary) and the training examples that are closest to this hyperplane*, which are the so called **support vectors**.

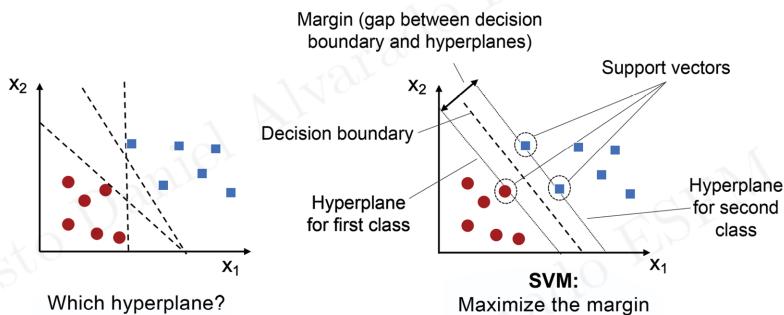


Figure 3.7: SVM and Perceptron Various Boundaries.

Idea 3.8.1 (Reduce Overfitting)

The key idea behind having decision boundaries with large margins is that they tend to have a lower generalization error, whereas models with small margins are more prone to overfitting.

While the main intuition behind SVMs is relatively simple, the mathematics behind them is advanced and requires sound knowledge of constrained optimization.

Observation 3.8.2

The following resources are recommended if interested in learning more:

- Chris J.C. Burges' explanation in *A Tutorial on Support Vector Machines for Pattern*

- Vladimir Vapnik's book *The Nature of Statistical Learning Theory*, Springer Science+Business Media, Vladimir Vapnik, 2000
- Andrew Ng's lecture notes.

3.8.1 DEALING WITH A NONLINEARLY SEPARABLE CASE USING SLACK VARIABLES

Let's mention the slack variable introduced by Vladimir Vapnik in 1995 and led to the so-called **soft-margin classification**.

Idea 3.8.2

The motivation for introducing the slack variable was that the *linear constraints in the SVM optimization are objective need to be relaxed for nonlinearly separable data to allow the convergence of the optimization in the presence of misclassifications*, under appropriate loss penalization.

The use of slack variable introduces a parameter commonly referred to as C in SVM. We can consider C as a hyperparameter controlling the penalty for misclassification.

Large values of C correspond to *large error penalties, whereas we are less strict about misclassification errors if we choose smaller values for C* . We can then use the C parameter to *control the width of the margin and therefore tune the bias-variance trade-off*.

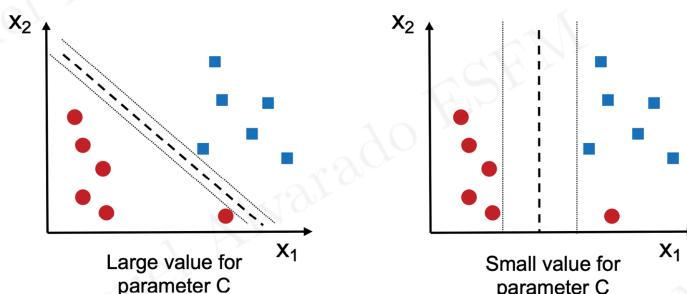


Figure 3.8: Parameter C in the Bias-Variance Trade-off.

Idea 3.8.3

This concept is related to regularization, which we discussed in the previous section in the context of regularized regression, where *decreasing the value of C increases the bias (underfitting) and lowers the variance (overfitting) of the model*.

3.8.2 TRAINING A MODEL WITH A SUPPORT VECTOR MACHINE

Now that we have learned the basic concepts behind a linear SVM, let's train an SVM model to classify the different flowers in our Iris dataset:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3 from matplotlib.colors import ListedColormap
4 from sklearn import datasets #To import Iris Dataset
5 from sklearn.preprocessing import StandardScaler
6 from sklearn.model_selection import train_test_split
```

```

7  from sklearn.svm import SVC
8
9  def plot_decision_regions(X, y, classifier, test_idx=None,
10    resolution=0.01):
11    markers = ('o', 's', '^', 'v', '<')
12    colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
13    cmap = ListedColormap(colors[:len(np.unique(y))])
14
15    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
16    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
17    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),
18                           np.arange(x2_min, x2_max, resolution))
19
20    lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
21                            .T)
22    lab = lab.reshape(xx1.shape)
23    plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
24    plt.xlim(xx1.min(), xx1.max())
25    plt.ylim(xx2.min(), xx2.max())
26
27    for idx, cl in enumerate(np.unique(y)):
28      plt.scatter(x=X[y == cl, 0],
29                  y=X[y == cl, 1],
30                  alpha=0.8,
31                  c=colors[idx],
32                  marker=markers[idx],
33                  label=f'Class {cl}',
34                  edgecolor='black')
35
36    if test_idx is not None:
37      X_test, y_test = X[test_idx, :], y[test_idx]
38      plt.scatter(X_test[:, 0], X_test[:, 1],
39                  c='none', edgecolor='black', alpha=1.0,
40                  linewidth=1, marker='o',
41                  s=100, label='Test set')
42
43 #Import Data
44
45 iris = datasets.load_iris()
46 X = iris.data[:,2:4]
47 y = iris.target
48
49 #Scale Standard Data
50
51 sc = StandardScaler()
52 X = sc.fit_transform(X)
53
54 # Split training data and test data
55 indices = np.arange(len(X))
56 X_train, X_test, idx_train, idx_test = train_test_split(
57   X, indices, test_size=0.3, random_state=1

```

```

55 )
56 y_train = y[idx_train]
57 y_test = y[idx_test]
58
59 svm = SVC(kernel='linear', C=1.0, random_state=1)
60 svm.fit(X_train, y_train)
61 plot_decision_regions(X,
62                         y,
63                         classifier=svm,
64                         test_idx=idx_test)
65 plt.xlabel('Petal length [standardized]')
66 plt.ylabel('Petal width [standardized]')
67 plt.legend(loc='upper left')
68 plt.tight_layout()
69 plt.show()

```

Code 3.11: Implementation of SVM on Python and Graph of Regions.

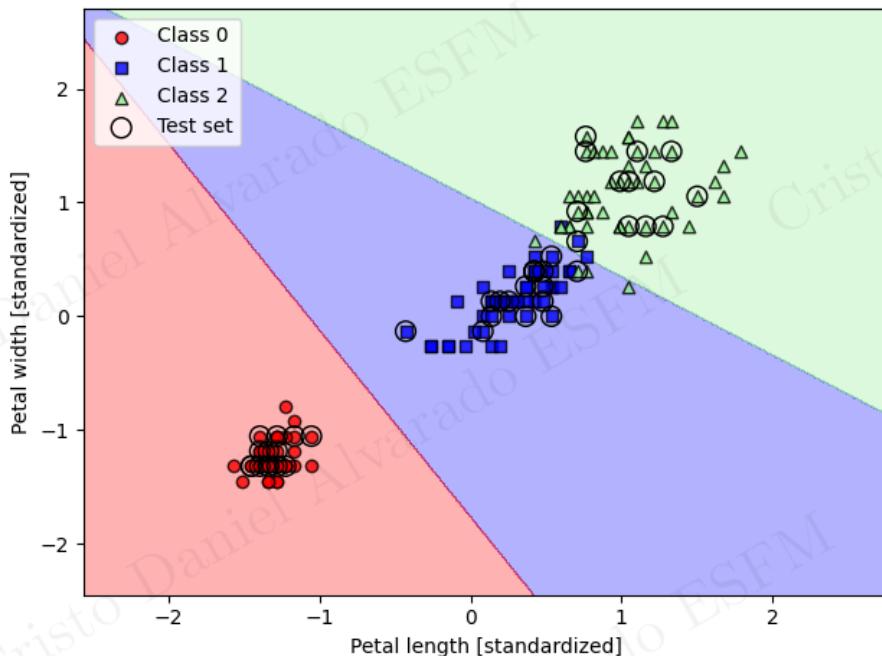


Figure 3.9: Implementation of SVM and Plot of Regions.

3.8.3 ALTERNATIVE IMPLEMENTATIONS IN SCIKIT-LEARN

The scikit-learn library's `LogisticRegression` class, which we used in the previous sections, can make use of the `LIBLINEAR` library by setting `solver='liblinear'`. `LIBLINEAR` is a highly optimized C/C++ library developed at the National Taiwan University ([Liblinear](#)).

Similarly, the `SVC` class that we used to train an `SVM` makes use of `LIBSVM`, which is an equivalent C/C++ library specialized for `SVMs` ([Libsvm](#)).

The advantage of using `LIBLINEAR` and `LIBSVM` over, for example, native Python implementations is that they allow the **extremely quick training of large amounts of linear classifiers**. However, *sometimes our datasets are too large to fit into computer memory*. Thus, scikit-learn also offers alternative implementations via the `SGDClassifier` class, which also supports online learning via the `partial_fit` method. The concept behind the `SGDClassifier` class is *similar to the stochastic gradient algorithm previously implemented for Adaline*.

Observation 3.8.3

We could initialize the SGD version of the perceptron (`loss='perceptron'`), logistic regression (`loss='log'`), and an SVM with default parameters (`loss='hinge'`) as follows:

```
1 from sklearn.linear_model import SGDClassifier  
2  
3 ppn = SGDClassifier(loss='perceptron')  
4 lr = SGDClassifier(loss='log')  
5 svm = SGDClassifier(loss='hinge')
```

Code 3.12: caption

Definition 3.8.2 (SGDClassifier)

SGDClassifier es una clase de scikit-learn que implementa clasificadores lineales entrenados usando Stochastic Gradient Descent (SGD).

3.9 SOLVING NONLINEAR PROBLEMS USING A KERNEL SVM

Another reason why SVMs enjoy high popularity among machine learning practitioners is that they can be easily **kernelized** to solve nonlinear classification problems.

Before we discuss the main concept behind the so-called kernel SVM, the most common variant of SVMs, let's first create a synthetic dataset to see what such a nonlinear classification problem may look like.

3.9.1 KERNEL METHODS FOR LINEARLY INSEPARABLE DATA

Using the following code, we will create a simple dataset that has the form of an **XOR** gate using the `logical_xor` function from NumPy, where 100 examples will be assigned the class label 1, and 100 examples will be assigned the class label -1:

```
1 import matplotlib.pyplot as plt  
2 import numpy as np  
3 np.random.seed(1)  
4 X_xor = np.random.randn(200, 2)  
5 y_xor = np.logical_xor(X_xor[:, 0] > 0,  
6                         X_xor[:, 1] > 0)  
7 y_xor = np.where(y_xor, 1, 0)  
8 plt.scatter(X_xor[y_xor == 1, 0],  
9             X_xor[y_xor == 1, 1],  
10            c='royalblue', marker='s',  
11            label='Class 1')  
12 plt.scatter(X_xor[y_xor == 0, 0],  
13             X_xor[y_xor == 0, 1],  
14             c='tomato', marker='o',  
15             label='Class 0')  
16 plt.xlim([-3, 3])  
17 plt.ylim([-3, 3])  
18 plt.xlabel('Feature 1')  
19 plt.ylabel('Feature 2')  
20 plt.legend(loc='best')
```

```

21 plt.tight_layout()
22 plt.show()

```

Code 3.13: `XOR` in Python.

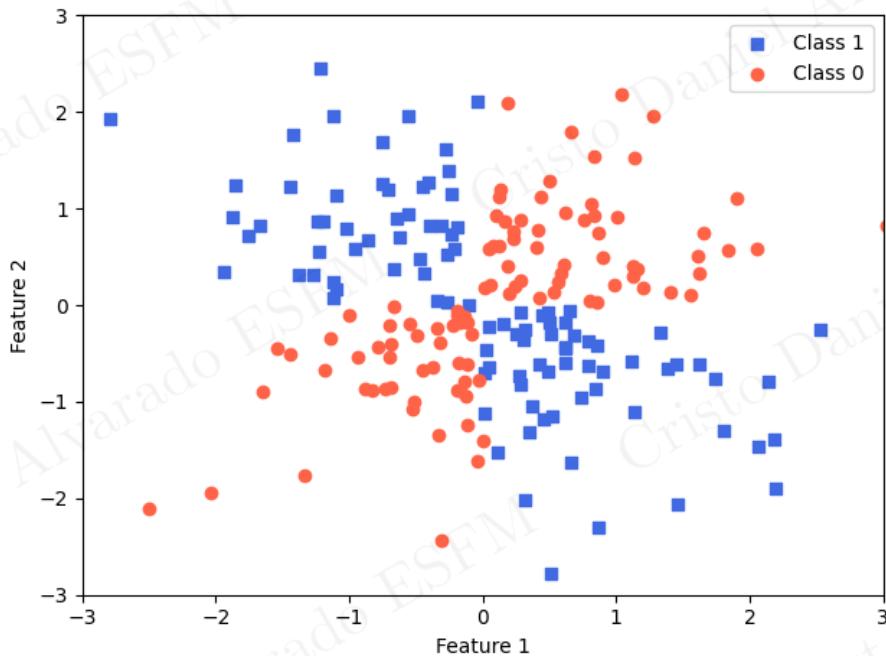


Figure 3.10: `XOR` Graph in Python.

Obviously, we would not be able to separate the examples from the positive and negative class very well using a linear hyperplane as a decision boundary via the linear logistic regression or linear SVM model that we discussed earlier.

Idea 3.9.1

The basic idea behind kernel methods for dealing with such linearly inseparable data is to create nonlinear combinations of the original features to project them onto a higher-dimensional space via a mapping function, φ , where the data becomes linearly separable. We can transform a two-dimensional dataset into a new three-dimensional feature space, where the classes become separable via the following projection:

$$\begin{aligned}\phi : \quad \mathbb{R}^2 &\rightarrow \mathbb{R}^3 \\ (x_1, x_2) &\mapsto (x_1, x_2, x_1^2 + x_2^2) = (z_1, z_2, z_3)\end{aligned}$$

This idea allows us to separate the two classes shown in the plot via a linear hyperplane that becomes a nonlinear decision boundary if we project it back onto the original feature space, as illustrated with this dataset:

3.9.2 USING THE KERNEL TRICK TO FIND SEPARATING HYPERPLANES IN A HIGH-DIMENSIONAL SPACE

To solve a nonlinear problem using an SVM, we would *transform the training data into a higher-dimensional feature space via a mapping function, ϕ , and train a linear SVM model to classify the data in this new feature space*. Then, we could use the same mapping function, ϕ , to transform new, unseen data to classify it using the linear SVM model.

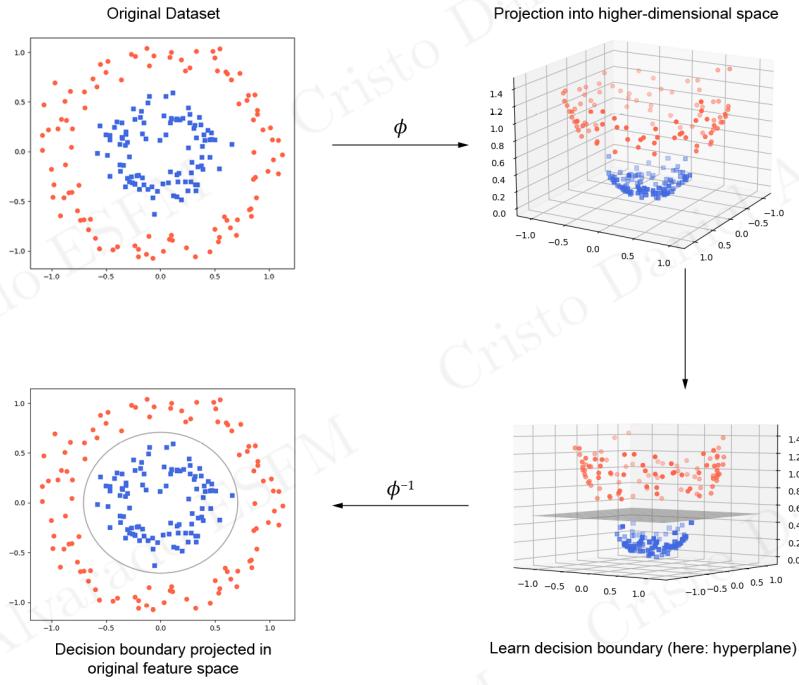


Figure 3.11: Dataset and Projection of Decision Boundary.

Observation 3.9.1 (Problem with Construction of New Features)

However, one problem with this mapping approach is that *the construction of the new features is computationally very expensive, especially if we are dealing with high-dimensional data.*

This is where the so-called **kernel trick** comes into play. Although we did not go into much detail about how to solve the quadratic programming task to train an SVM, in practice, we just need to replace the dot product $x^{(i)T}x^{(j)}$ by $\phi(x^{(i)})^T\phi(x^{(j)})$. To save the expensive step of calculating this dot product between two points explicitly, we define a so-called kernel function:

$$k(x^i, x^j) = \phi(x^i)^T\phi(x^j)$$

One of the most widely used kernels is the radial basis function (RBF) kernel, which can simply be called the Gaussian kernel:

$$k(x^i, x^j) = \exp\left(-\frac{\|x^i - x^j\|^2}{2\sigma^2}\right)$$

This is often simplified to:

$$k(x^i, x^j) = \exp(-\gamma\|x^i - x^j\|^2)$$

Here, $\gamma = \frac{1}{2\sigma^2}$ is a free parameter to be optimized.

Observation 3.9.2

Roughly speaking, the term "kernel" can be interpreted as a *similarity function between a pair of examples*.

The minus sign inverts the distance measure into a similarity score, and, due to the exponential term, *the resulting similarity score will fall into a range between 1 (for exactly similar examples)*

and 0 (for very dissimilar examples).

Now that we have covered the big picture behind the kernel trick, let's see if we can train a kernel SVM that is able to draw a nonlinear decision boundary that separates the XOR data well. Here, we simply use the SVC class from scikit-learn that we imported earlier and replace the `kernel='linear'` parameter with `kernel='rbf'`.

Note: Fit the SVM model and plot decision regions.

As we can see in the resulting plot, the kernel SVM separates the XOR data relatively well:

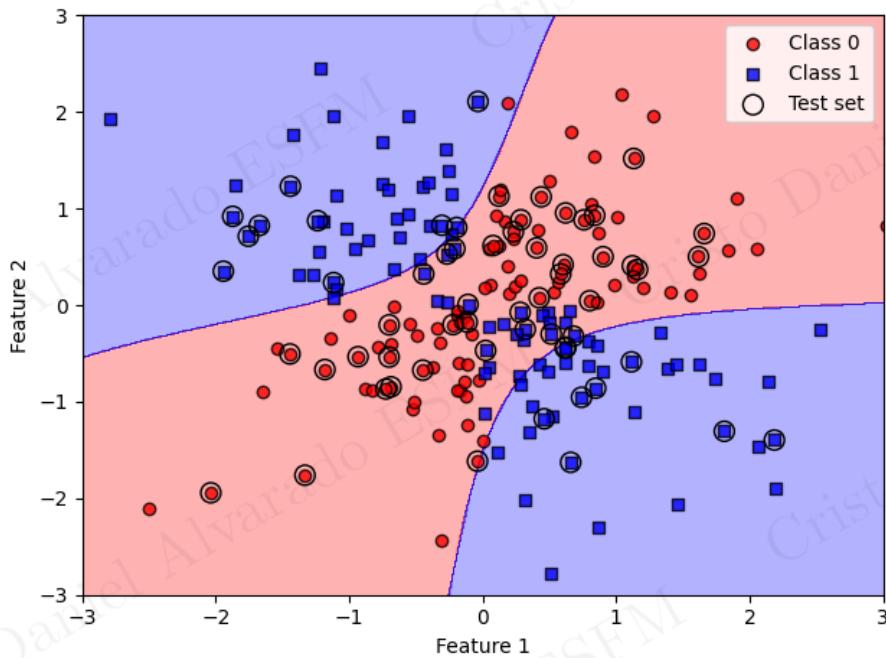


Figure 3.12: XOR Separation Using RBF.

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3 from sklearn.svm import SVC
4 from matplotlib.colors import ListedColormap
5 from sklearn.model_selection import train_test_split
6
7 def plot_decision_regions(X, y, classifier, test_idx=None,
8     resolution=0.01):
9     markers = ('o', 's', '^', 'v', '<')
10    colors = ('red', 'blue')
11    cmap = ListedColormap(colors[:len(np.unique(y))])
12
13    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
14    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
15    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),
16                           np.arange(x2_min, x2_max, resolution))
17
18    lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()]).T)
19    lab = lab.reshape(xx1.shape)
20    plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
```

```

19     plt.xlim(xx1.min(), xx1.max())
20     plt.ylim(xx2.min(), xx2.max())
21
22     for idx, cl in enumerate(np.unique(y)):
23         plt.scatter(x=X[y == cl, 0],
24                     y=X[y == cl, 1],
25                     alpha=0.8,
26                     c=colors[idx],
27                     marker=markers[idx],
28                     label=f'Class {cl}',
29                     edgecolor='black')
30
31     if test_idx is not None:
32         X_test, y_test = X[test_idx, :], y[test_idx]
33         plt.scatter(X_test[:, 0], X_test[:, 1],
34                     c='none', edgecolor='black', alpha=1.0,
35                     linewidth=1, marker='o',
36                     s=100, label='Test set')
37
38 np.random.seed(1)
39
40 X = np.random.randn(200, 2)
41 y = np.logical_xor(X[:, 0] > 0,
42                     X[:, 1] > 0)
43 y = np.where(y, 1, 0)
44
45 """
46
47 plt.scatter(X_xor[y_xor == 1, 0],
48             X_xor[y_xor == 1, 1],
49             c='royalblue', marker='s',
50             label='Class 1')
51
52 plt.scatter(X_xor[y_xor == 0, 0],
53             X_xor[y_xor == 0, 1],
54             c='tomato', marker='o',
55             label='Class 0')
56 """
57
58 # Split training data and test data
59 indices = np.arange(len(X))
60 X_train, X_test, idx_train, idx_test = train_test_split(
61     X, indices, test_size=0.3, random_state=1
62 )
63 y_train = y[idx_train]
64 y_test = y[idx_test]
65
66 #Train model
67 svm = SVC(kernel='rbf', random_state=1, gamma=0.2, C=1.0)
68 svm.fit(X_train, y_train)
69
70 #plot info

```

```

71 plot_decision_regions(X, y, classifier=svm, test_idx=idx_test)
72
73 plt.xlim([-3, 3])
74 plt.ylim([-3, 3])
75
76 plt.xlabel('Feature 1')
77 plt.ylabel('Feature 2')
78
79 plt.legend(loc='best')
80 plt.tight_layout()
81 plt.show()

```

Code 3.14: Decision Boundary using RBF Kernel on the XOR Dataset.

The γ parameter, which we set to $\gamma = 0.1$, can be understood as a *cut-off parameter for the Gaussian sphere*; increasing γ *increases the influence of the training examples*, leading to a tighter and bumpier decision boundary. To get a better understanding of γ , let's apply an RBF kernel SVM to our Iris flower dataset:

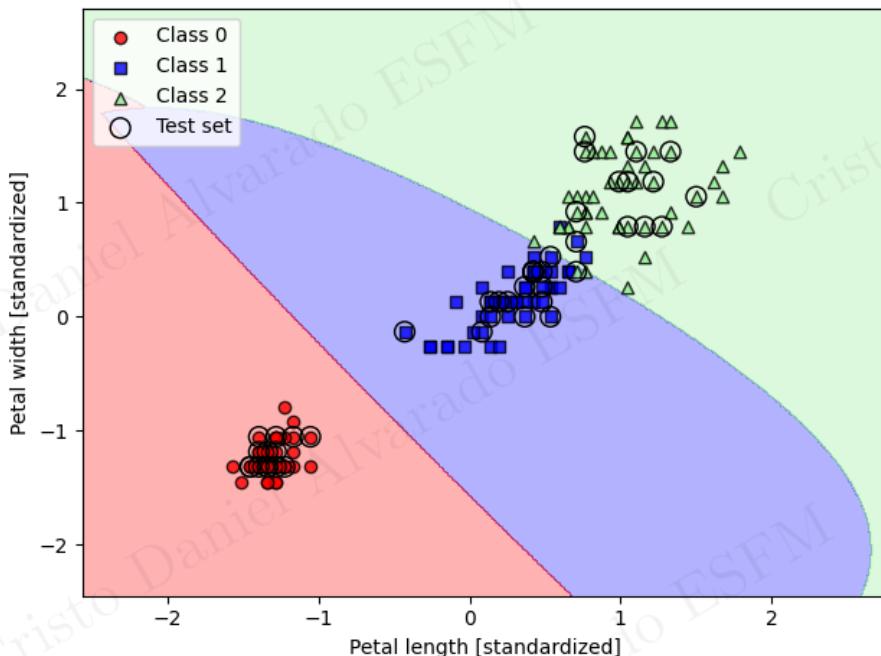


Figure 3.13: Iris SVC using RBF.

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3 from matplotlib.colors import ListedColormap
4 from sklearn import datasets #To import Iris Dataset
5 from sklearn.preprocessing import StandardScaler
6 from sklearn.model_selection import train_test_split
7 from sklearn.svm import SVC
8
9 def plot_decision_regions(X, y, classifier, test_idx=None,
   resolution=0.01):
10    markers = ('o', 's', '^', 'v', '<')
11    colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
12    cmap = ListedColormap(colors[:len(np.unique(y))])

```

```

13
14     x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
15     x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
16     xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution)
17                             ,
18                             np.arange(x2_min, x2_max, resolution))
19
20     lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
21                               .T)
22     lab = lab.reshape(xx1.shape)
23     plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
24     plt.xlim(xx1.min(), xx1.max())
25     plt.ylim(xx2.min(), xx2.max())
26
27
28     for idx, cl in enumerate(np.unique(y)):
29         plt.scatter(x=X[y == cl, 0],
30                     y=X[y == cl, 1],
31                     alpha=0.8,
32                     c=colors[idx],
33                     marker=markers[idx],
34                     label=f'Class {cl}',
35                     edgecolor='black')
36
37     if test_idx is not None:
38         X_test, y_test = X[test_idx, :], y[test_idx]
39         plt.scatter(X_test[:, 0], X_test[:, 1],
40                     c='none', edgecolor='black', alpha=1.0,
41                     linewidth=1, marker='o',
42                     s=100, label='Test set')
43
44
45 #Import Data
46
47
48 iris = datasets.load_iris()
49 X = iris.data[:,2:4]
50 y = iris.target
51
52 #Scale Standard Data
53
54 sc = StandardScaler()
55 X = sc.fit_transform(X)
56
57 # Split training data and test data
58 indices = np.arange(len(X))
59 X_train, X_test, idx_train, idx_test = train_test_split(
60     X, indices, test_size=0.3, random_state=1
61 )
62 y_train = y[idx_train]
63 y_test = y[idx_test]
64
65 svm = SVC(kernel='rbf', gamma=1, C=1.0, random_state=1)
66 svm.fit(X_train, y_train)
67 plot_decision_regions(X,

```

```

62
63
64
65
66
67
68
69

```

```

y ,
classifier=svm ,
test_idx=idx_test)
plt.xlabel('Petal length [standardized]')
plt.ylabel('Petal width [standardized]')
plt.legend(loc='upper left')
plt.tight_layout()
plt.show()

```

Code 3.15: Decision Boundary on the Iris Dataset using RBF.

If we increase the value of gamma, the following happens:

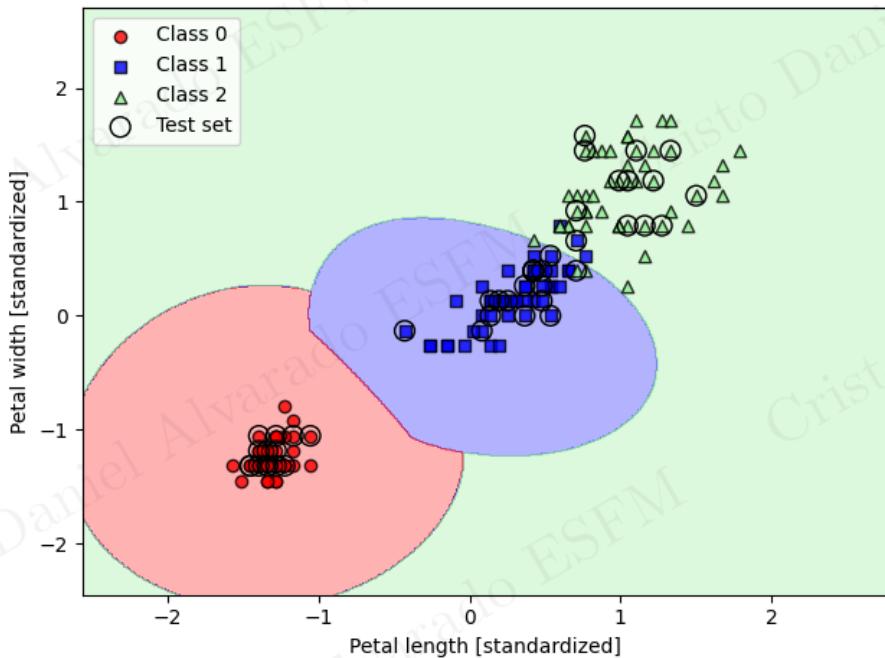


Figure 3.14: Iris Dataset using RBF Kernel SVM with $\gamma = 1$.

Observation 3.9.3

Although the model fits the training dataset very well, *such a classifier will likely have a high generalization error on unseen data*. This illustrates that the γ parameter also plays an important role in controlling overfitting or variance when the algorithm is too sensitive to fluctuations in the training dataset.

3.10 DECISION TREE LEARNING

Definition 3.10.1 (Decision Tree)

Decision tree classifiers are attractive models if we care about interpretability. As the name **decision tree** suggests, we can think of this model as *breaking down our data by making a decision based on asking a series of questions*.

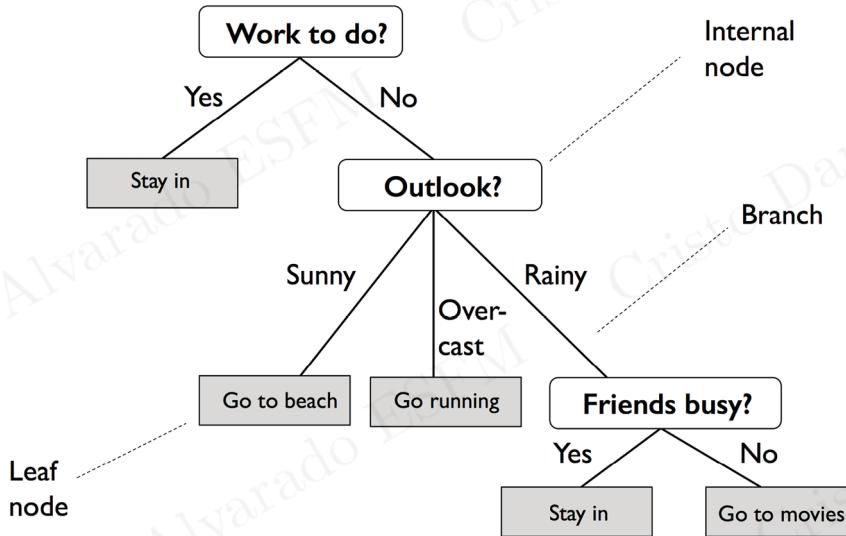


Figure 3.15: Decision Tree

Based on the features in our training dataset, the decision tree model learns a series of questions to infer the class labels of the examples.

Observation 3.10.1

Although the concept of a decision tree based on categorical variables is illustrated, the same concept applies if our features are real numbers, like in the Iris dataset.

For example, we could simply define a cut-off value along the sepal width feature axis and ask a binary question: *Is the sepal width ≥ 2.8 cm?*

Using the decision algorithm, we start at the tree root and split the data on the feature that results in the largest information gain (IG), which will be explained in more detail in the following section.

In an iterative process, we can then repeat this splitting procedure at each child node until the leaves are pure. This means that the training examples at each node all belong to the same class.

Idea 3.10.1

In practice, this can result in a very deep tree with many nodes, which can easily lead to overfitting. Thus, we typically want to prune the tree by setting a limit for the maximum depth of the tree.

3.10.1 MAXIMIZING IG GETTING THE MOST BANG FOR YOUR BUCK

To split the nodes at the most informative features, we need to define an objective function to optimize via the tree learning algorithm. Here, our objective function is to maximize the IG at each split, which we define as follows:

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^m \frac{N_j}{N_p} I(D_j)$$

Here:

- f is the feature to perform the split.
- D_p and D_j are the datasets of the parent and j -th child node.
- I is our **impurity measure**.
- N_p is the total number of training examples at the parent node.
- N_j is the number of examples in the j -th node.

The information gain (function IG) is simply the difference between the impurity of the parent node and the sum of the child node impurities — the lower the impurities of the child nodes, the larger the information gain.

Observation 3.10.2

However, for simplicity and to reduce the combinatorial search space, most libraries (including scikit-learn) implement binary decision trees. This means that each parent node is split into two child nodes: D_{left} and D_{right} .

$$IG(D_p, f) = I(D_p) - \frac{N_{left}}{N_p} I(D_{left}) - \frac{N_{right}}{N_p} I(D_{right})$$

The three impurity measures (I) or splitting criteria that are commonly used in binary decision trees are:

- **Gini impurity** (I_G).
- **Entropy** (I_H).
- **Classification error** (I_E).

Definition 3.10.2 (Entropy I_H)

The definition of **entropy** for all non-empty classes is given by:

$$I_H(t) = - \sum_{i=1}^c p(i|t) \log_2 p(i|t)$$

Here:

- $p(i|t)$ is the proportion of examples that belong to class i for a particular node t .

Observation 3.10.3

The entropy is therefore zero if all examples at a node belong to the same class, and the entropy is maximal if we have a uniform class distribution.

Example 3.10.1 (Objective of the Entropy Criterion)

In a binary class setting, the entropy is 0 if $p(i = 0|t) = 1$ or $p(i = 0|t) = 0$. If the classes are distributed uniformly with $p(i = 0|t) = \frac{1}{2}$ or $p(i = 1|t) = \frac{1}{2}$, the entropy is 1. Therefore, we can say that **the entropy criterion attempts to maximize the mutual information in the tree.**

To provide a visual intuition, let us visualize the entropy values for different class distributions via the following code:

```
1 import numpy as np
2 import matplotlib.pyplot as plt
3
4 def entropy(p):
5     return - p * np.log2(p) - (1 - p) * np.log2((1 - p))
6
7 x = np.arange(0.0, 1.0, 0.01)
8 ent = [entropy(p) if p != 0 else None for p in x]
9 plt.ylabel('Entropy')
10 plt.xlabel('Class-membership probability p(i=1)')
11 plt.plot(x, ent)
12 plt.show()
```

Code 3.16: Entropy Computation.

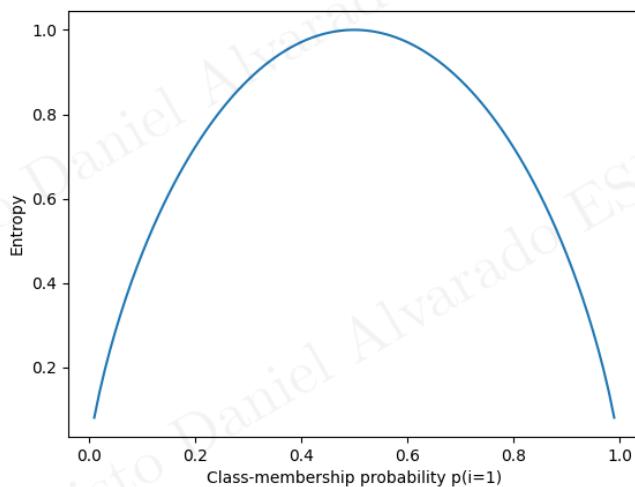


Figure 3.16: Entropy Visualization

Definition 3.10.3 (Gini Impurity)

The Gini impurity is given by:

$$I_G(t) = \sum_{i=1}^c p(i|t)(1 - p(i|t)) = 1 - \sum_{i=1}^c p(i|t)^2$$

Observation 3.10.4

That is, the Gini Impurity can be understood as a **criterion to minimize the probability of misclassification**.

Observation 3.10.5 (Relation Between Gini Impurity and Entropy)

Similar to entropy, Gini Impurity is maximal if the classes are perfectly mixed, for example, in a binary class setting (where $c = 2$):

$$I_G(t) = 1 - \sum_{i=1}^c \left(\frac{1}{2}\right)^2 = \frac{1}{2}$$

In practice, **both the Gini impurity and entropy typically yield very similar results**, and it is often not worth spending much time on evaluating trees using different impurity criteria rather than experimenting with different pruning cut-offs.

Definition 3.10.4 (Pruning)

Pruning means *removing parts of a model that are unnecessary or harmful, with the goal of making the model simpler, faster, and better at generalizing*.

Definition 3.10.5 (Classification Error)

Another impurity measure is the **classification error**:

$$I_E(t) = 1 - \max \{p(i|t)\}$$

This is a useful criterion for pruning, but not recommended for growing a decision tree, since it is less sensitive to changes in the class probabilities of the nodes. We can illustrate this by looking at the two possible splitting scenarios shown below:

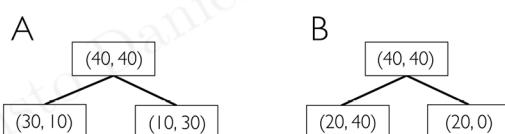


Figure 3.17: Decision Trees *A* and *B*.

Example 3.10.2

We start with a dataset D_p , at the parent node, which consists of 40 examples from class 1 and 40 examples from class 2, that we split into two datasets, D_{left} and D_{right} .

- The information gain using the **classification error** as a splitting criterion would be the

same (since $IG = 0.25$) in both scenarios, A and B :

$$\left\{ \begin{array}{l} I_E(D_p) = 1 - 0.5 = 0.5 \\ A : I_E(D_{\text{left}}) = 1 - \frac{3}{4} = 0.25 \\ A : I_E(D_{\text{right}}) = 1 - \frac{3}{4} = 0.25 \\ A : IG_E = 0.5 - \frac{4}{8} 0.25 - \frac{4}{8} 0.25 = 0.25 \\ B : I_E(D_{\text{left}}) = 1 - \frac{4}{6} = \frac{1}{3} \\ B : I_E(D_{\text{right}}) = 1 - 1 = 0 \\ B : IG_E = 0.5 - \frac{6}{8} \times \frac{1}{3} - 0 = 0.25 \end{array} \right.$$

- However, the **Gini impurity** would favor the split in scenario B ($IG_G \approx 0.1667$) over scenario A ($IG_G = 0.125$), which is indeed purer:

$$\left\{ \begin{array}{l} I_G(D_p) = 1 - (0.5^2 + 0.5^2) = 0.5 \\ A : I_G(D_{\text{left}}) = 1 - \left(\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2 \right) = \frac{3}{8} = 0.375 \\ A : I_G(D_{\text{right}}) = 1 - \left(\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2 \right) = \frac{3}{8} = 0.375 \\ A : IG_G = 0.5 - \frac{4}{8} 0.375 - \frac{4}{8} 0.375 = 0.125 \\ B : I_G(D_{\text{left}}) = 1 - \left(\left(\frac{2}{6}\right)^2 + \left(\frac{4}{6}\right)^2 \right) = \frac{4}{9} = 0.\bar{4} \\ B : I_G(D_{\text{right}}) = 1 - (1^2 + 0^2) = 0 \\ B : IG_G = 0.5 - \frac{6}{8} 0.\bar{4} - 0 = 0.1\bar{6} \end{array} \right.$$

- Similarly, the **entropy criterion** would also favor scenario B ($IG_H = 0.31$) over scenario A ($IG_H = 0.19$):

$$\left\{ \begin{array}{l} I_H(D_p) = -(0.5 \log_2(0.5) + 0.5 \log_2(0.5)) = 1 \\ A : I_H(D_{\text{left}}) = -\left(\frac{3}{4} \log_2\left(\frac{3}{4}\right) + \frac{1}{4} \log_2\left(\frac{1}{4}\right)\right) = 0.81 \\ A : I_H(D_{\text{right}}) = -\left(\frac{1}{4} \log_2\left(\frac{1}{4}\right) + \frac{3}{4} \log_2\left(\frac{3}{4}\right)\right) = 0.81 \\ A : IG_H = 1 - \frac{4}{8} 0.81 - \frac{4}{8} 0.81 = 0.19 \\ B : I_H(D_{\text{left}}) = -\left(\frac{2}{6} \log_2\left(\frac{2}{6}\right) + \frac{4}{6} \log_2\left(\frac{4}{6}\right)\right) = 0.92 \\ B : I_H(D_{\text{right}}) = 0 \\ B : IG_H = 1 - \frac{6}{8} 0.92 - 0 = 0.31 \end{array} \right.$$

For a more visual comparison of the three different impurity criteria that we discussed previously, let's plot the impurity indices for the probability range $[0, 1]$ for class 1. Note that we will also add a scaled version of the entropy (entropy / 2) to observe that the Gini impurity is an intermediate measure between entropy and the classification error. The code is as follows:

```

1 import matplotlib.pyplot as plt
2 import numpy as np
3
4 def gini(p):
5     return p*(1 - p) + (1 - p)*(1 - (1-p))
6
7 def entropy(p):
8     return - p*np.log2(p) - (1 - p)*np.log2((1 - p))
9
10 def error(p):
11     return 1 - np.max([p, 1 - p])

```

```

12
13 x = np.arange(0.0, 1.0, 0.01)
14 ent = [entropy(p) if p != 0 else None for p in x]
15 sc_ent = [e*0.5 if e else None for e in ent]
16 err = [error(i) for i in x]
17 fig = plt.figure()
18 ax = plt.subplot(111)
19 for i, lab, ls, c, in zip([ent, sc_ent, gini(x), err],
20                           ['Entropy', 'Entropy (scaled)',
21                            'Gini impurity',
22                            'Misclassification error'],
23                           [ '--', '--', '---', '-.'],
24                           ['black', 'lightgray',
25                            'red', 'green', 'cyan']):
26     line = ax.plot(x, i, label=lab,
27                      linestyle=ls, lw=2, color=c)
28 ax.legend(loc='upper center', bbox_to_anchor=(0.5, 1.15),
29            ncol=5, fancybox=True, shadow=False)
30 ax.axhline(y=0.5, linewidth=1, color='k', linestyle='--')
31 ax.axhline(y=1.0, linewidth=1, color='k', linestyle='--')
32 plt.ylim([0, 1.1])
33 plt.xlabel('p(i=1)')
34 plt.ylabel('impurity index')
35 plt.show()

```

Code 3.17: Entropy Computation.

Idea 3.10.2

This example illustrates that, although two splits may appear equivalent under a coarse criterion such as classification error, more sensitive criteria like Gini impurity and entropy better capture the true purity of the resulting nodes, favoring splits that produce highly homogeneous subsets. In summary:

Key Point	Criterion	Sensitivity	Interpretation
(1) Informative power	Classification Error	Low	Insensitive to changes in class distribution; only considers the majority class and may assign the same information gain to qualitatively different splits.
(1) Informative power	Gini Impurity	Medium	Sensitive to class proportions and able to distinguish between moderately and highly pure nodes.
(1) Informative power	Entropy	High	Highly sensitive to uncertainty; strongly rewards splits that produce nodes with high class concentration.

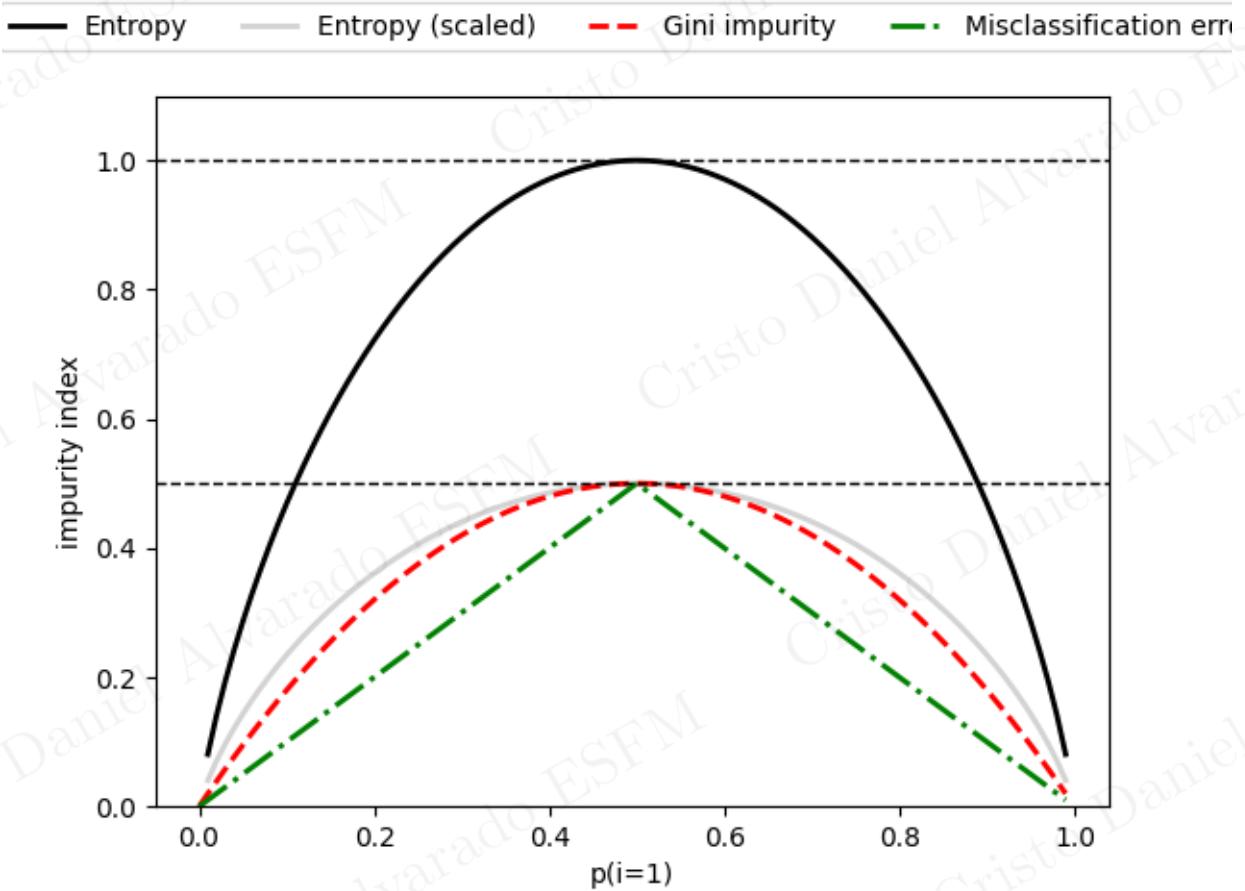


Figure 3.18: Caption

(2) Preference for pure nodes	Gini / Entropy	High	Both criteria favor splits that generate highly pure (or perfectly pure) child nodes, even if the remaining child node becomes less pure.
(3) Relative position	Classification Error	Very Low	Nearly flat behavior as a function of class probability; provides a coarse measure of impurity.
(3) Relative position	Gini Impurity	Medium	Acts as an intermediate measure between entropy and classification error, balancing sensitivity and stability.
(3) Relative position	Entropy	Very High	Exhibits strong curvature near probabilities close to 0 or 1, heavily penalizing uncertainty and strongly favoring purity.

3.11 BUILDING A DECISION TREE

Decision trees can build complex decision boundaries by dividing the feature space into rectangles.

Observation 3.11.1

However, we have to be careful since the deeper the decision tree, the more complex the decision

boundary becomes, which can easily result in overfitting.

Using scikit-learn, we will now train a decision tree with a maximum depth of 4, using the Gini impurity as a criterion for impurity.

Although *feature scaling may be desired for visualization purposes, note that feature scaling is not a requirement for decision tree algorithms*. The code is as follows:

```
1 import matplotlib.pyplot as plt
2 import numpy as np
3 from sklearn.tree import DecisionTreeClassifier
4 from matplotlib.colors import ListedColormap
5 from sklearn.model_selection import train_test_split
6 from sklearn import datasets
7
8 def plot_decision_regions(X, y, classifier, test_idx=None,
9     resolution=0.2):
10    markers = ('o', 's', '^', 'v', '<')
11    colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
12    cmap = ListedColormap(colors[:len(np.unique(y))])
13
14    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
15    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
16    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),
17                           np.arange(x2_min, x2_max, resolution))
18
19    lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
20                            .T)
21    lab = lab.reshape(xx1.shape)
22    plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
23    plt.xlim(xx1.min(), xx1.max())
24    plt.ylim(xx2.min(), xx2.max())
25
26    for idx, cl in enumerate(np.unique(y)):
27        plt.scatter(x=X[y == cl, 0],
28                    y=X[y == cl, 1],
29                    alpha=0.8,
30                    c=colors[idx],
31                    marker=markers[idx],
32                    label=f'Class {cl}',
33                    edgecolor='black')
34
35    if test_idx is not None:
36        X_test, y_test = X[test_idx, :], y[test_idx]
37        plt.scatter(X_test[:, 0], X_test[:, 1],
38                    c='none', edgecolor='black', alpha=1.0,
39                    linewidth=1, marker='o',
40                    s=100, label='Test set')
41
42 #Gather data
43
44 iris = datasets.load_iris()
```

```

42 X = iris.data[:, 2:4]
43 y = iris.target
44
45 # Split training data and test data
46 indices = np.arange(len(X))
47 X_train, X_test, idx_train, idx_test = train_test_split(
48     X, indices, test_size=0.3, random_state=1
49 )
50 y_train = y[idx_train]
51 y_test = y[idx_test]
52
53 #Train model
54
55 tree_model = DecisionTreeClassifier(criterion='gini',
56                                     max_depth=4,
57                                     random_state=1)
58 tree_model.fit(X_train, y_train)
59 X_combined = np.vstack((X_train, X_test))
60 y_combined = np.hstack((y_train, y_test))
61 plot_decision_regions(X_combined,
62                       y_combined,
63                       classifier=tree_model,
64                       test_idx=idx_test)
65 plt.xlabel('Petal length [cm]')
66 plt.ylabel('Petal width [cm]')
67 plt.legend(loc='upper left')
68 plt.tight_layout()
69 plt.show()

```

Code 3.18: Code for Decision Tree Implementation.

Observation 3.11.2

A nice feature in scikit-learn is that it allows us to readily visualize the decision tree model after training via the following code:

```

1 from sklearn import tree
2 feature_names = ['Sepal length', 'Sepal width',
3                  'Petal length', 'Petal width']
4 tree.plot_tree(tree_model,
5                feature_names=feature_names,
6                filled=True)
7 plt.show()

```

Code 3.19: caption

Setting `filled=True` in the `plot_tree` function we called colors the nodes by the majority class label at that node.

Observation 3.11.3

There are many additional options available, which you can find in the documentation at [Plot Tree](#).

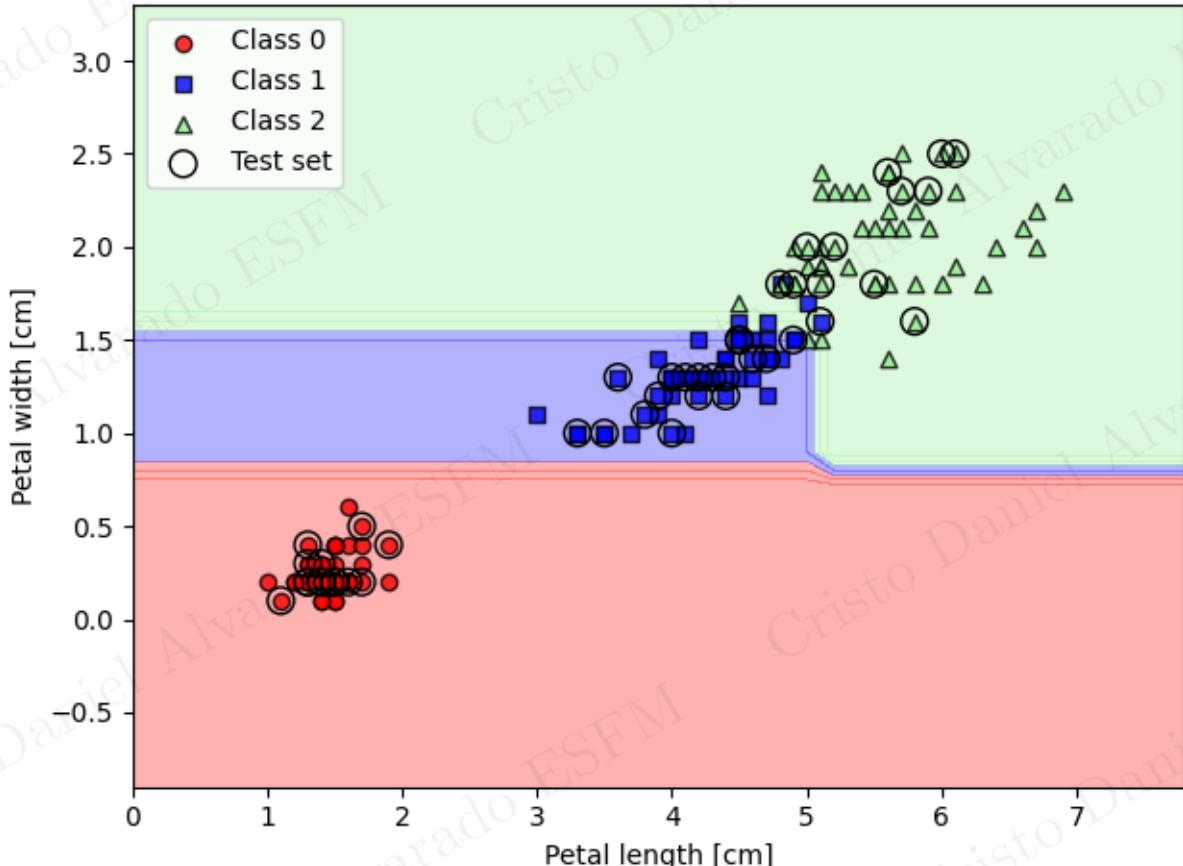


Figure 3.19: Decision Tree Boundary Visualized.

Looking at the root node, it starts with 105 examples at the top. The first split uses a sepal width $cut-off \leq 0.75$ cm for splitting the root node into two child nodes with 35 examples (left child node) and 70 examples (right child node). After the first split, we can see that the left child node is already pure and only contains examples from the Iris-setosa class (`Gini impurity = 0`). The further splits on the right are then used to separate the examples from the Iris-versicolor and Iris-virginica class.

Idea 3.11.1

The decision tree does a very good job of separating the flower classes. Unfortunately, scikit-learn currently does not implement functionality to manually post-prune a decision tree. However, we could go back to our previous code example, change the `max_depth` of our decision tree to 3, and compare it to our current model, but we leave this as an exercise for the interested reader.

Observation 3.11.4 (Cost Complexity Post-Pruning)

Alternatively, scikit-learn provides an automatic cost complexity post-pruning procedure for decision trees. Interested readers can find more information about this more advanced topic in the following tutorial: [Plot Cost Complexity Pruning](#).

3.11.1 COMBINING MULTIPLE DECISION TREES VIA RANDOM FORESTS

Ensemble methods have gained huge popularity in applications of machine learning during the last decade due to their good classification performance and robustness toward overfitting.

Let's discuss the decision tree-based random forest algorithm, which is known for its good scalability and ease of use.

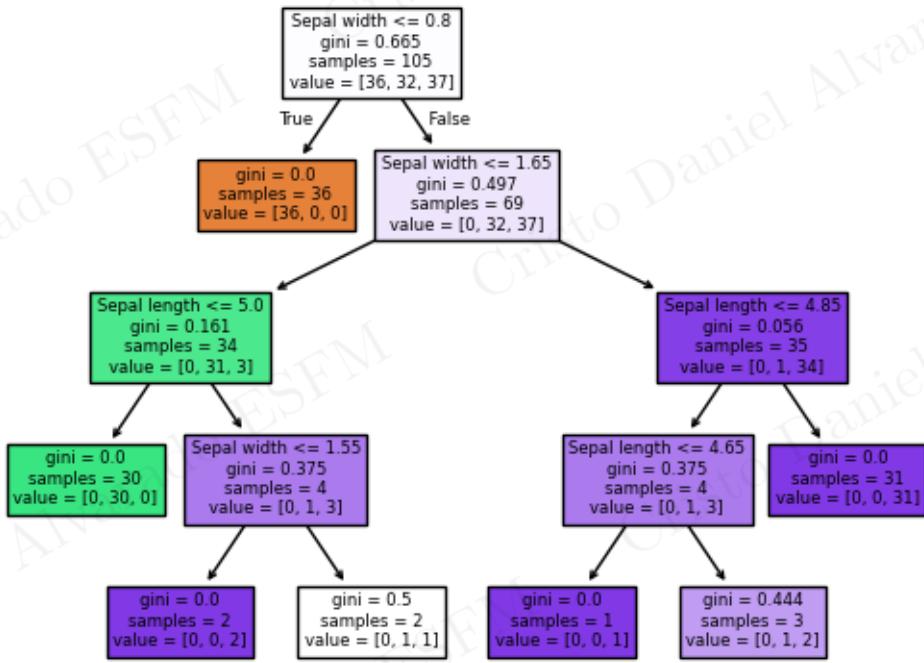


Figure 3.20: Tree Visualized.

Definition 3.11.1 (Random Forest)

A **random forest** can be considered as *an ensemble of decision trees*.

The idea behind a random forest is to *average multiple (deep) decision trees that individually suffer from high variance to build a more robust model that has a better generalization performance and is less susceptible to overfitting*.

The random forest algorithm can be summarized in four simple steps:

- 1) Draw a random bootstrap sample of size n (randomly choose n examples from the training dataset with replacement).
- 2) Grow a decision tree from the bootstrap sample. At each node:
 - Randomly select d features without replacement.
 - Split the node using the feature that provides the best split according to the objective function, for instance, maximizing the information gain.
- 3) Repeat steps 1)-2) k times.
- 4) Aggregate the prediction by each tree to assign the class label by majority vote.

Observation 3.11.5

We should note one slight modification in step 2) when we are training the individual decision

trees: instead of evaluating all features to determine the best split at each node, we only consider a random subset of those.

3.12 K-NEAREST NEIGHBOURS

The last supervised learning algorithm that we want to discuss in this section is the k-nearest neighbor (KNN) classifier, which is particularly interesting because it is fundamentally different from the learning algorithms discussed so far.

Observation 3.12.1

KNN is a typical example of a lazy learner. It is called *lazy* not because of its apparent simplicity, but because *it doesn't learn a discriminative function from the training data but memorizes the training dataset instead*.

KNN belongs to a subcategory of non-parametric models described as instance-based learning. *Models based on instance-based learning memorize the training dataset, and lazy learning is a special case of instance-based learning that incurs no cost during the learning process.*

- **Parametric** models estimate parameters.
- **Non-parametric** models adapt to data.

The KNN algorithm itself is straightforward and can be summarized by the following steps:

- 1) Choose the number of k and a distance metric
- 2) Find the k -nearest neighbors of the data record to classify
- 3) Assign the class label by majority vote

Based on the chosen distance metric, the KNN algorithm finds the k examples in the training dataset that are closest (most similar) to the point that we want to classify.

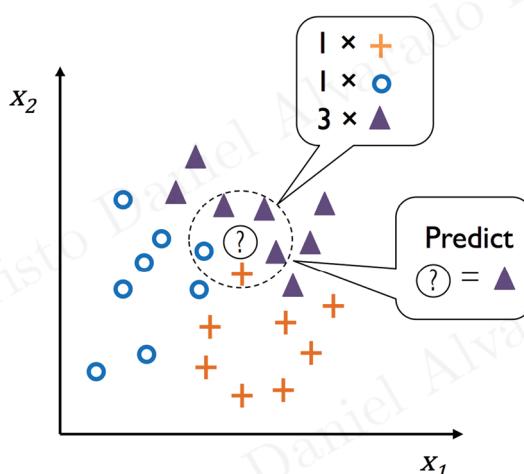


Figure 3.21: k -Nearest Neighbours Algorithm

Observation 3.12.2

The class label of the data point is then determined by a majority vote among its k nearest neighbors.

3.12.1 ADVANTAGES AND DISADVANTAGES OF MEMORY-BASED APPROACHES

- **Advantage** of a memory-based approach is that the classifier immediately adapts as new training data are collected.
- **Downside** is that the computational complexity for classifying new examples grows linearly with the number of examples in the training dataset in the worst-case scenario—unless the dataset has very few dimensions and the algorithm uses efficient data structures for querying the training data.

Such data structures include $k - d$ tree and ball tree, which are both supported in scikit-learn.

In addition to computational costs, large datasets can be problematic in terms of storage requirements.

Idea 3.12.1

For relatively small to medium-sized datasets, memory-based methods can provide good predictive and computational performance and are therefore suitable for many real-world problems.

By executing the following code, we implement a KNN model in scikit-learn using a Euclidean distance metric:

```
1 from sklearn.neighbors import KNeighborsClassifier
2 import matplotlib.pyplot as plt
3 import numpy as np
4 from sklearn.tree import DecisionTreeClassifier
5 from matplotlib.colors import ListedColormap
6 from sklearn.model_selection import train_test_split
7 from sklearn import datasets
8
9 def plot_decision_regions(X, y, classifier, test_idx=None,
   resolution=0.2):
10    markers = ('o', 's', '^', 'v', '<')
11    colors = ('red', 'blue', 'lightgreen', 'gray', 'cyan')
12    cmap = ListedColormap(colors[:len(np.unique(y))])
13
14    x1_min, x1_max = X[:, 0].min() - 1, X[:, 0].max() + 1
15    x2_min, x2_max = X[:, 1].min() - 1, X[:, 1].max() + 1
16    xx1, xx2 = np.meshgrid(np.arange(x1_min, x1_max, resolution),
17                           np.arange(x2_min, x2_max, resolution))
18    lab = classifier.predict(np.array([xx1.ravel(), xx2.ravel()])
19                            .T)
20    lab = lab.reshape(xx1.shape)
21    plt.contourf(xx1, xx2, lab, alpha=0.3, cmap=cmap)
22    plt.xlim(xx1.min(), xx1.max())
23    plt.ylim(xx2.min(), xx2.max())
```

```

23
24     for idx, cl in enumerate(np.unique(y)):
25         plt.scatter(x=X[y == cl, 0],
26                      y=X[y == cl, 1],
27                      alpha=0.8,
28                      c=colors[idx],
29                      marker=markers[idx],
30                      label=f'Class {cl}',
31                      edgecolor='black')
32
33     if test_idx is not None:
34         X_test, y_test = X[test_idx, :], y[test_idx]
35         plt.scatter(X_test[:, 0], X_test[:, 1],
36                     c='none', edgecolor='black', alpha=1.0,
37                     linewidth=1, marker='o',
38                     s=100, label='Test set')
39
40 #Gather data
41
42 iris = datasets.load_iris()
43 X = iris.data[:, 2:4]
44 y = iris.target
45
46 # Split training data and test data
47 indices = np.arange(len(X))
48 X_train, X_test, idx_train, idx_test = train_test_split(
49     X, indices, test_size=0.3, random_state=1
50 )
51 y_train = y[idx_train]
52 y_test = y[idx_test]
53
54 #Train model
55
56 knn = KNeighborsClassifier(n_neighbors=5, p=2,
57                             metric='minkowski')
58 knn.fit(X_train, y_train)
59 X_combined = np.vstack((X_train, X_test))
60 y_combined = np.hstack((y_train, y_test))
61 plot_decision_regions(X_combined,
62                       y_combined,
63                       classifier=knn,
64                       test_idx=idx_test)
65 plt.xlabel('Petal length [standardized]')
66 plt.ylabel('Petal width [standardized]')
67 plt.legend(loc='upper left')
68 plt.tight_layout()
69 plt.show()

```

Code 3.20: Code for k -Nearest Neighbours.

By specifying five neighbors, we obtain a relatively smooth decision boundary:

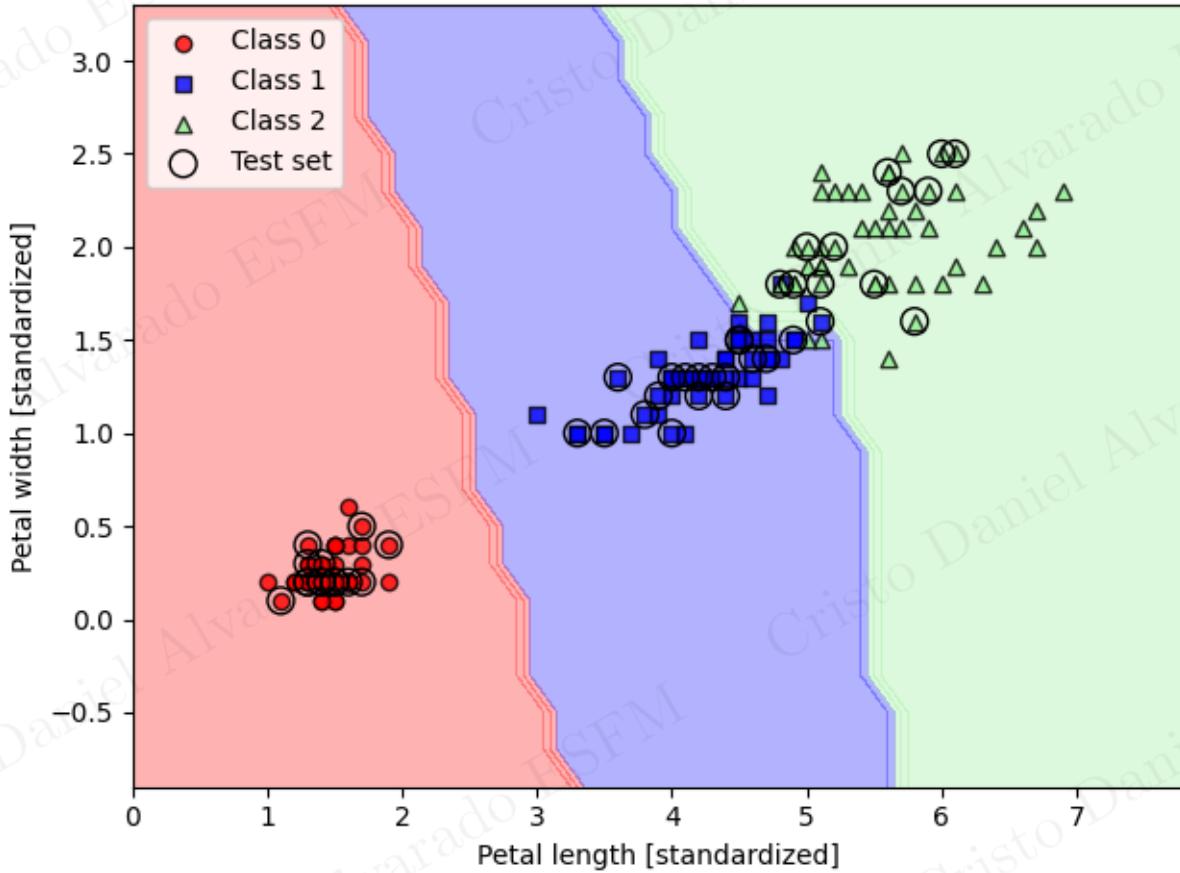


Figure 3.22: Caption

Observation 3.12.3

In the case of a tie, scikit-learn prefers neighbors with a closer distance. If the neighbors have identical distances, the algorithm chooses the class label that appears first in the training dataset.

Idea 3.12.2

Choosing k is crucial to finding a good balance between overfitting and underfitting. A suitable distance metric is also essential. A simple Euclidean distance measure is often used for real-valued features, such as the centimeters-based measurements in the Iris dataset.

When using Euclidean distance, standardizing the data is important so that each feature contributes equally to the distance. The minkowski distance used in the preceding code is a generalization of the Euclidean and Manhattan distances and can be written as follows:

$$d_p(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p}, p \in \mathbb{N}$$

It becomes the Euclidean distance when $p=2$ and the Manhattan distance when $p=1$. Many other distance metrics are available in scikit-learn and can be provided to the metric parameter.

Observation 3.12.4

KNN is susceptible to overfitting due to the curse of dimensionality, where the feature space becomes increasingly sparse as dimensionality grows. Even the closest neighbors can be too far

away in a high-dimensional space to provide a good estimate.

Regularization helps avoid overfitting in models such as logistic regression, but it is not applicable to decision trees and KNN. Instead, feature selection and dimensionality reduction techniques help mitigate the curse of dimensionality.

Observation 3.12.5

Note: Use RAPIDS' cuML for GPU-accelerated machine learning with large datasets.

CHAPTER 4

BUILDING GOOD TRAINIGN DATASETS: DATA PREPROCESSING

4.1 INTRODUCTION

The quality of the data and the amount of useful information that it contains are key factors that determine how well a machine learning algorithm can learn.

It is absolutely critical to ensure that we examine and preprocess a dataset before we feed it to a machine learning algorithm.

Observation 4.1.1

The topics that we will cover in this chapter are as follows:

- Removing and imputing missing values from the dataset
- Getting categorical data into shape for machine learning algorithms
- Selecting relevant features for the model construction

4.2 DEALING WITH MISSING DATA

Sometimes data may be missing due to various reasons.

Observation 4.2.1

There could have been an error in the data collection process, certain measurements may not be applicable, or particular fields could have been simply left blank in a survey, for example. We typically see missing values as blank spaces in our data table or as placeholder strings such as `NaN`, which stands for *not a number* or `NULL`.

We will work through several practical techniques for dealing with missing values by removing entries from our dataset or imputing missing values from other training examples and features.

4.3 IDENTIFYING MISSING VALUES IN TABULAR DATA

Before we discuss several techniques for dealing with missing values, let's create a simple example DataFrame from a comma-separated values ([CSV](#)) file to get a better grasp of the problem:

```
1 import pandas as pd
2 from io import StringIO
3 csv_data = \
4     'A,B,C,D'
```

```

5 1.0,2.0,3.0,4.0
6 5.0,6.0,,8.0
7 10.0,11.0,12.0,''
8 # If you are using Python 2.7, you need
9 # to convert the string to unicode:
10 # csv_data = unicode(csv_data)
11 df = pd.read_csv(StringIO(csv_data))
12 print(df)

```

Code 4.1: CSV with Missing Data.

The output is the following:

	A	B	C	D
1				
2	0	1.0	2.0	3.0
3	1	5.0	6.0	NaN
4	2	10.0	11.0	12.0

Code 4.2: CSV with Missing Data Output

Using the preceding code, we read **CSV**-formatted data into a pandas **DataFrame** via the **read_csv** function and noticed that the two missing cells were replaced by **NaN**. The **StringIO** function in the preceding code example was simply used for the purposes of illustration. It allowed us to read the string assigned to **csv_data** into a pandas **DataFrame** as if it was a regular **CSV** file on our hard drive.

Observation 4.3.1

For a larger **DataFrame**, it can be tedious to look for missing values manually; in this case, we can use the **isnull** method to return a **DataFrame** with **Boolean** values that indicate whether a cell contains a numeric value (**False**) or if data is missing (**True**).

Using the **sum** method, we can then return the number of missing values per column as follows:

```
1 print(df.isnull().sum())
```

Code 4.3: Check Missing Data per Row

The output is this:

	A	B	C	D
1	0	0	1	1
2				
3				
4				
5	dtype: int64			

Code 4.4: Output Missing Data Per Row

Although scikit-learn was originally developed for working with **NumPy** arrays only, it can sometimes be more convenient to preprocess data using pandas' **DataFrame**.

Nowadays, most scikit-learn functions support **DataFrame** objects as inputs, but since **NumPy** array handling is more mature in the scikit-learn API, it is recommended to use **NumPy** arrays when possible. Note that you can always access the underlying **NumPy** array of a **DataFrame** via the **values** attribute before you feed it into a scikit-learn estimator:

```
1 print(df.values)
```

Code 4.5: Dataframe `df`

```
1 array([[ 1.,  2.,  3.,  4.],
2       [ 5.,  6., nan,  8.],
3       [10., 11., 12., nan]])
```

Code 4.6: Values of Dataframe `df`

4.3.1 ELIMINATING TRAINING EXAMPLES OR FEATURES WITH MISSING VALUES

One of the easiest ways to deal with missing data is simply to remove the corresponding features (columns) or training examples (rows) from the dataset entirely; rows with missing values can easily be dropped via the `dropna` method:

```
1 print(df.dropna(axis=0))
2
3 output:
4
5      A      B      C      D
6 0    1.0    2.0    3.0    4.0
```

Code 4.7: Droping Rows from `df`

Similarly, we can drop columns that have at least one NaN in any row by setting the axis argument to 1:

```
1 print(df.dropna(axis=1))
2
3 output:
4
5      A      B
6 0    1.0    2.0
7 1    5.0    6.0
8 2   10.0   11.0
```

Code 4.8: Dropping Columns from a `DataFrame df`

APPENDIX A

USING PYTHON FOR MACHINE LEARNING

A.1 BASIC CONFIGURATION

For machine learning tasks, we will mostly refer to the *Scikit-Learn* library, which is one of the most popular and accessible open-source machine learning libraries for python.

When we focus on a subfield of machine learning called: **deep learning**, we will use the latest version of the *PyTorch library*, which specializes in the training of the so called **neural network models**.

Definition A.1.1 (Scikit-learn)

Scikit-learn is a *classical machine learning library for Python*. It is built on top of NumPy and SciPy and provides a clean, uniform, and simple API for a wide variety of traditional ML algorithms.

PyTorch is an *open-source deep learning framework developed primarily by Facebook's AI Research lab (FAIR)*. It provides the foundational building blocks for building and training neural networks, with a strong focus on flexibility and speed.

We'll use version 3.9 of Python. To check the version of python we use the following code:

```
1 python --version
```

Code A.1: Check Python Version

Observation A.1.1 (Use of pip)

To install additional packages used throughout the course, we can install them via the **pip** program.

Definition A.1.2 (Pip)

pip is the *standard package installer for Python*. It is a *command-line utility that allows users to install, manage, and uninstall Python packages and libraries that are not part of the Python standard library*.

After installing python, it's possible to execute the following command in order to install some package:

```
1 pip install SomePackage
```

Code A.2: **pip** Package Installation.

A.2 ANACONDA PYTHON DISTRIBUTION AND PACKAGE MANAGER

Idea A.2.1

A recommended open-source package management system for installing python for scientific computing contexts is conda by Continuum Analytics.

Definition A.2.1 (Conda)

Conda is a *free and licensed under a permissive open-source licence*. Its goal is to help with the *installation and version management of Python packages for data science, math and engineering across different operating systems*.

Conda comes in different flavours, like a Linux installation. Some of them are **Anaconda**, **Miniconda** and **Miniforge**.

- **Anaconda** comes with many scientific computing packages pre-installed. The Anaconda installer can be downloaded [here](#) and an Anaconda quick start guide is available [here](#).
- **Miniconda** is a leaner alternative to Anaconda ([here](#)). Essentially, it is similar to Anaconda but without any packages pre-installed, which many people (including the authors) prefer.
- **Miniforge** is similar to Miniconda but community-maintained and uses a different package repository (conda-forge) from Miniconda and Anaconda. We found that Miniforge is a great alternative to Miniconda. Download and installation instructions can be found in the GitHub repository [here](#).

After successfully installing conda through either Anaconda, Miniconda, or Miniforge, we can install and update, respectively, new Python packages using the following command:

```
1 conda install SomePackage  
2 conda update SomePackage
```

Code A.3: Conda Package Installation.

Packages that are not available through the official conda channel might be available via the community-supported conda-forge project ([conda-forge](#)), which can be specified via the `--channel conda-forge` flag. For example:

```
1 conda install SomePackage --channel conda-forge
```

Code A.4: caption

Packages not available through the default conda channel or conda-forge can be installed via pip as explained earlier in Code A.2.

A.3 PACKAGES FOR SCIENTIFIC COMPUTING, DATA SCIENCE, AND MACHINE LEARNING

We will use mainly NumPy's arrays and sometimes Pandas (higher level data manipulation tools). Also the Matplotlib library will be useful to visualize quantitative data. The version of the packages to install are the following:

- NumPy 1.21.2
- SciPy 1.7.0
- Scikit-learn 1.0
- Matplotlib 3.4.3
- Pandas 1.3.2

After installing these packages, you can double-check the installed version by importing the package in Python and accessing its `__version__` attribute:

```
1 > Use 'numpy.__version__' to check the installed version.  
2 '1.21.2'
```

Code A.5: Check Version of Packages

APPENDIX B

A NOTE ON DISTRIBUTIONS

B.1 NORMAL DISTRUBUTION

B.2 BERNOUlli'S DISTRIBUTION

Definition B.2.1 (Bernoulli distribution)

The **Bernoulli distribution**, named after Swiss mathematician Jacob Bernoulli, *is the discrete probability distribution of a random variable which takes the value 1 with probability p and the value 0 with probability q = 1 - p.*

The Bernoulli distribution is a special case of the binomial distribution where a single trial is conducted (so n would be 1 for such a binomial distribution). It is *also a special case of the two-point distribution, for which the possible outcomes need not be 0 and 1.*

If X is a random variable with a Bernoulli distribution, then:

$$\begin{aligned} P(X = 1) &= p \\ P(X = 0) &= q = 1 - p \end{aligned}$$

The Bernoulli distrubution is simply $B(1, p)$, also written as $\text{Bernoulli}(p)$.