A Brief Introduction to Machine Learning for Engineers
Other titles in Foundations and Trends® in Signal Processing

**Massive MIMO Networks: Spectral, Energy, and Hardware Efficiency**
Emil Bjornson, Jakob Hoydis and Luca Sanguinetti

**Using Inertial Sensors for Position and Orientation Estimation**
Manon Kok, Jeroen D. Hol and Thomas B. Schon
ISBN: 978-1-68083-356-0

**Computational Visual Attention Models**
Milind S. Gide and Lina J. Karam
ISBN: 978-1-68083-280-8

**Video Coding: Part II of Fundamentals of Source and Video Coding**
Thomas Wiegand and Heiko Schwarz
ISBN: 978-1-68083-178-8
A Brief Introduction to Machine Learning for Engineers

Osvaldo Simeone
Department of Informatics
King’s College London
osvaldo.simeone@kcl.ac.uk
Editorial Scope

Topics

Foundations and Trends® in Signal Processing publishes survey and tutorial articles in the following topics:

- Adaptive signal processing
- Audio signal processing
- Biological and biomedical signal processing
- Complexity in signal processing
- Digital signal processing
- Distributed and network signal processing
- Image and video processing
- Linear and nonlinear filtering
- Multidimensional signal processing
- Multimodal signal processing
- Multirate signal processing
- Multiresolution signal processing
- Nonlinear signal processing
- Randomized algorithms in signal processing
- Sensor and multiple source signal processing, source separation
- Signal decompositions, subband and transform methods, sparse representations
- Signal processing for communications
- Signal processing for security and forensic analysis, biometric signal processing
- Signal quantization, sampling, analog-to-digital conversion, coding and compression
- Signal reconstruction, digital-to-analog conversion, enhancement, decoding and inverse problems
- Speech/audio/image/video compression
- Speech and spoken language processing
- Statistical/machine learning
- Statistical signal processing
  - Classification and detection
  - Estimation and regression
  - Tree-structured methods

Information for Librarians

Foundations and Trends® in Signal Processing, 2018, Volume 12, 4 issues. ISSN paper version 1932-8346. ISSN online version 1932-8354. Also available as a combined paper and online subscription.
## Contents

1 Basics 5  

1 Introduction 6  
1.1 What is Machine Learning? 6  
1.2 When to Use Machine Learning? 8  
1.3 Goals and Outline 11  

2 A Gentle Introduction through Linear Regression 15  
2.1 Supervised Learning 15  
2.2 Inference 17  
2.3 Frequentist Approach 19  
2.4 Bayesian Approach 35  
2.5 Minimum Description Length (MDL)* 42  
2.6 Information-Theoretic Metrics 44  
2.7 Interpretation and Causality* 46  
2.8 Summary 48  

3 Probabilistic Models for Learning 51  
3.1 Preliminaries 52  
3.2 The Exponential Family 53  
3.3 Frequentist Learning 59  
3.4 Bayesian Learning 62
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>Supervised Learning via Generalized Linear Models (GLM)</td>
<td>69</td>
</tr>
<tr>
<td>3.6</td>
<td>Maximum Entropy Property*</td>
<td>71</td>
</tr>
<tr>
<td>3.7</td>
<td>Energy-based Models*</td>
<td>72</td>
</tr>
<tr>
<td>3.8</td>
<td>Some Advanced Topics*</td>
<td>73</td>
</tr>
<tr>
<td>3.9</td>
<td>Summary</td>
<td>74</td>
</tr>
<tr>
<td>II</td>
<td>Supervised Learning</td>
<td>75</td>
</tr>
<tr>
<td>4</td>
<td>Classification</td>
<td>76</td>
</tr>
<tr>
<td>4.1</td>
<td>Preliminaries: Stochastic Gradient Descent</td>
<td>77</td>
</tr>
<tr>
<td>4.2</td>
<td>Classification as a Supervised Learning Problem</td>
<td>78</td>
</tr>
<tr>
<td>4.3</td>
<td>Discriminative Deterministic Models</td>
<td>80</td>
</tr>
<tr>
<td>4.4</td>
<td>Discriminative Probabilistic Models: Generalized Linear Models</td>
<td>90</td>
</tr>
<tr>
<td>4.5</td>
<td>Discriminative Probabilistic Models: Beyond GLM</td>
<td>96</td>
</tr>
<tr>
<td>4.6</td>
<td>Generative Probabilistic Models</td>
<td>101</td>
</tr>
<tr>
<td>4.7</td>
<td>Boosting*</td>
<td>105</td>
</tr>
<tr>
<td>4.8</td>
<td>Summary</td>
<td>106</td>
</tr>
<tr>
<td>5</td>
<td>Statistical Learning Theory*</td>
<td>112</td>
</tr>
<tr>
<td>5.1</td>
<td>A Formal Framework for Supervised Learning</td>
<td>113</td>
</tr>
<tr>
<td>5.2</td>
<td>PAC Learnability and Sample Complexity</td>
<td>118</td>
</tr>
<tr>
<td>5.3</td>
<td>PAC Learnability for Finite Hypothesis Classes</td>
<td>119</td>
</tr>
<tr>
<td>5.4</td>
<td>VC Dimension and Fundamental Theorem of PAC Learning</td>
<td>123</td>
</tr>
<tr>
<td>5.5</td>
<td>Summary</td>
<td>125</td>
</tr>
<tr>
<td>III</td>
<td>Unsupervised Learning</td>
<td>128</td>
</tr>
<tr>
<td>6</td>
<td>Unsupervised Learning</td>
<td>129</td>
</tr>
<tr>
<td>6.1</td>
<td>Unsupervised Learning</td>
<td>130</td>
</tr>
<tr>
<td>6.2</td>
<td>$K$-Means Clustering</td>
<td>133</td>
</tr>
<tr>
<td>6.3</td>
<td>ML, ELBO and EM</td>
<td>135</td>
</tr>
<tr>
<td>6.4</td>
<td>Directed Generative Models</td>
<td>146</td>
</tr>
<tr>
<td>6.5</td>
<td>Undirected Generative Models</td>
<td>154</td>
</tr>
<tr>
<td>6.6</td>
<td>Discriminative Models</td>
<td>157</td>
</tr>
<tr>
<td>6.7</td>
<td>Autoencoders</td>
<td>159</td>
</tr>
</tbody>
</table>
ABSTRACT
This monograph aims at providing an introduction to key concepts, algorithms, and theoretical results in machine learning. The treatment concentrates on probabilistic models for supervised and unsupervised learning problems. It introduces fundamental concepts and algorithms by building on first principles, while also exposing the reader to more advanced topics with extensive pointers to the literature, within a unified notation and mathematical framework. The material is organized according to clearly defined categories, such as discriminative and generative models, frequentist and Bayesian approaches, exact and approximate inference, as well as directed and undirected models. This monograph is meant as an entry point for researchers with an engineering background in probability and linear algebra.
Notation

- Random variables or random vectors – both abbreviated as rvs – are represented using roman typeface, while their values and realizations are indicated by the corresponding standard font. For instance, the equality \( x = x \) indicates that rv \( x \) takes value \( x \).
  - Matrices are indicated using uppercase fonts, with roman typeface used for random matrices.
  - Vectors will be taken to be in column form.
  - \( X^T \) and \( X^\dagger \) are the transpose and the pseudoinverse of matrix \( X \), respectively.
  - The distribution of a rv \( x \), either probability mass function (pmf) for a discrete rv or probability density function (pdf) for continuous rvs, is denoted as \( p_x \), \( p_x(x) \), or \( p(x) \).
  - The notation \( x \sim p_x \) indicates that rv \( x \) is distributed according to \( p_x \).
  - For jointly distributed rvs \( (x, y) \sim p_{xy} \), the conditional distribution of \( x \) given the observation \( y = y \) is indicated as \( p_{x|y=y} \), \( p_{x|y}(x|y) \) or \( p(x|y) \).
  - The notation \( x|y = y \sim p_{x|y=y} \) indicates that rv \( x \) is drawn according to the conditional distribution \( p_{x|y=y} \).
  - The notation \( E_{x \sim p_x} [\cdot] \) indicates the expectation of the argument with respect to the distribution of the rv \( x \sim p_x \). Accordingly, we will also write \( E_{x \sim p_{x|y}} [\cdot|y] \) for the conditional expectation with respect to
the distribution $p_{x|y=y}$. When clear from the context, the distribution over which the expectation is computed may be omitted.

- The notation $\Pr_{x \sim p_x}[\cdot]$ indicates the probability of the argument event with respect to the distribution of the rv $x \sim p_x$. When clear from the context, the subscript is dropped.

- The notation $\log$ represents the logarithm in base two, while $\ln$ represents the natural logarithm.

- $x \sim \mathcal{N}(\mu, \Sigma)$ indicates that random vector $x$ is distributed according to a multivariate Gaussian pdf with mean vector $\mu$ and covariance matrix $\Sigma$. The multivariate Gaussian pdf is denoted as $\mathcal{N}(x|\mu, \Sigma)$ as a function of $x$.

- $x \sim \mathcal{U}(a, b)$ indicates that rv $x$ is distributed according to a uniform distribution in the interval $[a, b]$. The corresponding uniform pdf is denoted as $\mathcal{U}(x|a, b)$.

- $\delta(x)$ denotes the Dirac delta function or the Kronecker delta function, as clear from the context.

- $||a||^2 = \sum_{i=1}^{N} a_i^2$ is the quadratic, or $l_2$, norm of a vector $a = [a_1, \ldots, a_N]^T$. We similarly define the $l_1$ norm as $||a||_1 = \sum_{i=1}^{N} |a_i|$, and the $l_0$ pseudo-norm $||a||_0$ as the number of non-zero entries of vector $a$.

- $I$ denotes the identity matrix, whose dimensions will be clear from the context. Similarly, $1$ represents a vector of all ones.

- $\mathbb{R}$ is the set of real numbers; $\mathbb{R}^+$ the set of non-negative real numbers; $\mathbb{R}^-$ the set of non-positive real numbers; and $\mathbb{R}^N$ is the set of all vectors of $N$ real numbers.

- $1(\cdot)$ is the indicator function: $1(x) = 1$ if $x$ is true, and $1(x) = 0$ otherwise.

- $|S|$ represents the cardinality of a set $S$.

- $x_S$ represents a set of rvs $x_k$ indexed by the integers $k \in S$. 

Full text available at: http://dx.doi.org/10.1561/2000000102
**Acronyms**

AI: Artificial Intelligence  
AMP: Approximate Message Passing  
BN: Bayesian Network  
DAG: Directed Acyclic Graph  
ELBO: Evidence Lower BOund  
EM: Expectation Maximization  
ERM: Empirical Risk Minimization  
GAN: Generative Adversarial Network  
GLM: Generalized Linear Model  
HMM: Hidden Markov Model  
i.i.d.: independent identically distributed  
KL: Kullback-Leibler  
LASSO: Least Absolute Shrinkage and Selection Operator  
LBP: Loopy Belief Propagation  
LL: Log-Likelihood  
LLR: Log-Likelihood Ratio  
LS: Least Squares  
MC: Monte Carlo  
MCMC: Markov Chain Monte Carlo  
MDL: Minimum Description Length  
MFVI: Mean Field Variational Inference  
ML: Maximum Likelihood
MRF: Markov Random Field
NLL: Negative Log-Likelihood
PAC: Probably Approximately Correct
pdf: probability density function
pmf: probability mass function
PCA: Principal Component Analysis
PPCA: Probabilistic Principal Component Analysis
QDA: Quadratic Discriminant Analysis
RBM: Restricted Boltzmann Machine
SGD: Stochastic Gradient Descent
SVM: Support Vector Machine
rv: random variable or random vector (depending on the context)
s.t.: subject to
VAE: Variational AutoEncoder
VC: Vapnik–Chervonenkis
VI: Variational Inference
Part I

Basics
Introduction

Having taught courses on machine learning, I am often asked by colleagues and students with a background in engineering to suggest “the best place to start” to get into this subject. I typically respond with a list of books – for a general, but slightly outdated introduction, read this book; for a detailed survey of methods based on probabilistic models, check this other reference; to learn about statistical learning, I found this text useful; and so on. This answer strikes me, and most likely also my interlocutors, as quite unsatisfactory. This is especially so since the size of many of these books may be discouraging for busy professionals and students working on other projects. This monograph is an attempt to offer a basic and compact reference that describes key ideas and principles in simple terms and within a unified treatment, encompassing also more recent developments and pointers to the literature for further study.

1.1 What is Machine Learning?

A useful way to introduce the machine learning methodology is by means of a comparison with the conventional engineering design flow. This
1.1. What is Machine Learning?

starts with a in-depth analysis of the problem domain, which culminates with the definition of a mathematical model. The mathematical model is meant to capture the key features of the problem under study, and is typically the result of the work of a number of experts. The mathematical model is finally leveraged to derive hand-crafted solutions to the problem that offer given optimality guarantees.

For instance, consider the problem of defining a chemical process to produce a given molecule. The conventional flow requires chemists to leverage their knowledge of models that predict the outcome of individual chemical reactions, in order to craft a sequence of suitable steps that synthesize the desired molecule. Another example is the design of speech translation or image/video compression algorithms. Both of these tasks involve the definition of models and algorithms by teams of experts, such as linguists, psychologists, and signal processing practitioners, not infrequently during the course of long standardization meetings.

The engineering design flow outlined above may be too costly and inefficient for problems in which faster or less expensive solutions are desirable. The machine learning alternative is to collect large data sets, e.g., of labelled speech, images or videos, and to use this information to train general-purpose learning machines to carry out the desired task. While the standard engineering flow relies on domain knowledge and on design optimized for the problem at hand, machine learning lets large amounts of data dictate algorithms and solutions. To this end, rather than requiring a precise model of the set-up under study, machine learning requires the specification of an objective, of a generic model to be trained, and of an optimization technique.

Returning to the first example above, a machine learning approach would proceed by training a general-purpose machine to predict the outcome of known chemical reactions based on a large data set, and by then using the trained algorithm to explore ways to produce more complex molecules. In a similar manner, large data sets of images or videos would be used to train a general-purpose algorithm with the aim of obtaining compressed representations from which the original input can be recovered with some distortion.
1.2 When to Use Machine Learning?

Based on the discussion above, machine learning can offer an efficient alternative to the conventional engineering flow when development cost and time are the main concerns, or when the problem appears to be too complex to afford the development of solutions with optimality guarantees. On the flip side, the approach has the key disadvantages of providing generally suboptimal performance, of producing black-box, and hence non-interpretable, solutions, and of applying only to a limited set of problems.

In order to identify tasks for which machine learning methods may be useful, reference [31] suggests the following criteria:

1. the task involves a function that maps well-defined inputs to well-defined outputs;
2. large data sets exist or can be created containing input-output pairs;
3. the task provides clear feedback with clearly definable goals and metrics;
4. the task does not involve long chains of logic or reasoning that depend on diverse background knowledge or common sense;
5. the task does not require detailed explanations for how the decision was made;
6. the task has a tolerance for error and no need for provably correct or optimal solutions;
7. the phenomenon or function being learned should not change rapidly over time; and
8. no specialized dexterity, physical skills, or mobility is required.

These criteria are useful guidelines for the decision of whether machine learning methods are suitable for a given task of interest. They also offer a convenient demarcation line between machine learning as is intended
1.2. When to Use Machine Learning?

Today, with its focus on training and computational statistics tools, and more general notions of Artificial Intelligence (AI) based on knowledge and common sense [86] (see [126] for an overview on AI research).

1.2.1 Learning Tasks

We can distinguish among three different main types of machine learning problems, which are briefly introduced below. The discussion reflects the focus of this monograph on parametric probabilistic models, as further elaborated on in the next section.

1. **Supervised learning:** We have \( N \) labelled training examples \( D = \{(x_n, t_n)\}_{n=1}^N \), where \( x_n \) represents a covariate, or explanatory variable, while \( t_n \) is the corresponding target label, or response. For instance, variable \( x_n \) may represent the text of an email, while the label \( t_n \) may be a binary variable indicating whether the email is spam or not. The goal of supervised learning is to predict the value of the label \( t \) for an input \( x \) that is not in the training set. In other words, supervised learning aims at generalizing the observations in the data set \( D \) to new inputs. For example, an algorithm trained on a set of emails should be able to classify a new email not present in the data set \( D \).

We can generally distinguish between **classification** problems, in which the label \( t \) is discrete, as in the example above, and **regression** problems, in which variable \( t \) is continuous. An example of a regression task is the prediction of tomorrow’s temperature \( t \) based on today’s meteorological observations \( x \).

An effective way to learn a predictor is to identify from the data set \( D \) a predictive distribution \( p(t|x) \) from a set of parametrized distributions. The conditional distribution \( p(t|x) \) defines a profile of beliefs over all possible of the label \( t \) given the input \( x \). For instance, for temperature prediction, one could learn mean and variance of a Gaussian distribution \( p(t|x) \) as a function of the input \( x \). As a special case, the output of a supervised learning algorithm may be in the form of a deterministic predictive function \( t = \hat{t}(x) \).

2. **Unsupervised learning:** Suppose now that we have an unlabelled set of training examples \( D = \{x_n\}_{n=1}^N \). Less well defined than...
supervised learning, unsupervised learning generally refers to the task of learning properties of the mechanism that generates this data set. Specific tasks and applications include clustering, which is the problem of grouping similar examples $x_n$; dimensionality reduction, feature extraction, and representation learning, all related to the problem of representing the data in a smaller or more convenient space; and generative modelling, which is the problem of learning a generating mechanism to produce artificial examples that are similar to available data in the data set $D$.

As a generalization of both supervised and unsupervised learning, *semi-supervised learning* refers to scenarios in which not all examples are labelled, with the unlabelled examples providing information about the distribution of the covariates $x$.

3. Reinforcement learning: Reinforcement learning refers to the problem of inferring optimal sequential decisions based on rewards or punishments received as a result of previous actions. Under supervised learning, the “label” $t$ refers to an action to be taken when the learner is in an informational state about the environment given by a variable $x$. Upon taking an action $t$ in a state $x$, the learner is provided with feedback on the immediate reward accrued via this decision, and the environment moves on to a different state. As an example, an agent can be trained to navigate a given environment in the presence of obstacles by penalizing decisions that result in collisions.

Reinforcement learning is hence neither supervised, since the learner is not provided with the optimal actions $t$ to select in a given state $x$; nor is it fully unsupervised, given the availability of feedback on the quality of the chosen action. Reinforcement learning is also distinguished from supervised and unsupervised learning due to the influence of previous actions on future states and rewards.

This monograph focuses on supervised and unsupervised learning. These general tasks can be further classified along the following dimensions.

- *Passive vs. active learning*: A passive learner is given the training examples, while an active learner can affect the choice of training examples on the basis of prior observations.
1.3 Goals and Outline

This monograph aims at providing an introduction to key concepts, algorithms, and theoretical results in machine learning. The treatment concentrates on probabilistic models for supervised and unsupervised learning problems. It introduces fundamental concepts and algorithms by building on first principles, while also exposing the reader to more advanced topics with extensive pointers to the literature, within a unified notation and mathematical framework. Unlike other texts that are focused on one particular aspect of the field, an effort has been made here to provide a broad but concise overview in which the main ideas and techniques are systematically presented. Specifically, the material is organized according to clearly defined categories, such as discriminative and generative models, frequentist and Bayesian approaches, exact and approximate inference, as well as directed and undirected models. This monograph is meant as an entry point for researchers with a background in probability and linear algebra. A prior exposure to information theory is useful but not required.

Detailed discussions are provided on basic concepts and ideas, including overfitting and generalization, Maximum Likelihood and regularization, and Bayesian inference. The text also endeavors to provide intuitive explanations and pointers to advanced topics and research directions. Sections and subsections containing more advanced material that may be skipped at a first reading are marked with a star (*).

The reader will find here neither discussions on computing platform or programming frameworks, such as map-reduce, nor details on specific applications involving large data sets. These can be easily found in a vast and growing body of work. Furthermore, rather than providing
exhaustive details on the existing myriad solutions in each specific category, techniques have been selected that are useful to illustrate the most salient aspects. Historical notes have also been provided only for a few selected milestone events.

Finally, the monograph attempts to strike a balance between the algorithmic and theoretical viewpoints. In particular, all learning algorithms are introduced on the basis of theoretical arguments, often based on information-theoretic measures. Moreover, a chapter is devoted to statistical learning theory, demonstrating how to set the field of supervised learning on solid theoretical foundations. This chapter is more theoretically involved than the others, and proofs of some key results are included in order to illustrate the theoretical underpinnings of learning. This contrasts with other chapters, in which proofs of the few theoretical results are kept at a minimum in order to focus on the main ideas.

The rest of the monograph is organized into five parts. The first part covers introductory material. Specifically, Chapter 2 introduces the frequentist, Bayesian and Minimum Description Length (MDL) learning frameworks; the discriminative and generative categories of probabilistic models; as well as key concepts such as training loss, generalization, and overfitting – all in the context of a simple linear regression problem. Information-theoretic metrics are also briefly introduced, as well as the advanced topics of interpretation and causality. Chapter 3 then provides an introduction to the exponential family of probabilistic models, to Generalized Linear Models (GLMs), and to energy-based models, emphasizing main properties that will be invoked in later chapters.

The second part concerns supervised learning. Chapter 4 covers linear and non-linear classification methods via discriminative and generative models, including Support Vector Machines (SVMs), kernel methods, logistic regression, multi-layer neural networks and boosting. Chapter 5 is a brief introduction to the statistical learning framework of the Probably Approximately Correct (PAC) theory, covering the Vapnik–Chervonenkis (VC) dimension and the fundamental theorem of PAC learning.
1.3. Goals and Outline

The third part, consisting of a single chapter, introduced unsupervised learning. In particular, in Chapter 6, unsupervised learning models are described by distinguishing among directed models, for which Expectation Maximization (EM) is derived as the iterative maximization of the Evidence Lower BOUND (ELBO); undirected models, for which Restricted Boltzmann Machines (RBMs) are discussed as a representative example; discriminative models trained using the InfoMax principle; and autoencoders. Generative Adversarial Networks (GANs) are also introduced.

The fourth part covers more advanced modelling and inference approaches. Chapter 7 provides an introduction to probabilistic graphical models, namely Bayesian Networks (BNs) and Markov Random Fields (MRFs), as means to encode more complex probabilistic dependencies than the models studied in previous chapters. Approximate inference and learning methods are introduced in Chapter 8 by focusing on Monte Carlo (MC) and Variational Inference (VI) techniques. The chapter briefly introduces in a unified way techniques such as variational EM, Variational AutoEncoders (VAE), and black-box inference. Some concluding remarks are provided in the last part, consisting of Chapter 9.

We conclude this chapter by emphasizing the importance of probability as a common language for the definition of learning algorithms [34]. The centrality of the probabilistic viewpoint was not always recognized, but has deep historical roots. This is demonstrated by the following two quotes, the first from the first AI textbook published by P. H. Winston in 1977, and the second from an unfinished manuscript by J. von Neumann (see [126, 63] for more information):

“Many ancient Greeks supported Socrates opinion that deep, inexplicable thoughts came from the gods. Today’s equivalent to those gods is the erratic, even probabilistic neuron. It is more likely that increased randomness of neural behavior is the problem of the epileptic and the drunk, not the advantage of the brilliant.”

from Artificial Intelligence, 1977;
“All of this will lead to theories of computation which are much less rigidly of an all-or-none nature than past and present formal logic... There are numerous indications to make us believe that this new system of formal logic will move closer to another discipline which has been little linked in the past with logic. This is thermodynamics primarily in the form it was received from Boltzmann.”

References


[63] Hinton, G. 2016. “Neural Networks for Machine Learning (online course)”.


[89] Li, Y. “Topics in Approximate Inference”. URL: http://yingzhenli.net/home/pdf/topics_approx_infer.pdf.


References

variational inference”. In: Artificial Intelligence and Statistics.
814–822.

exponential families”. In: Artificial Intelligence and Statistics.
762–771.

backpropagation and approximate inference in deep generative

“Stabilizing Training of Generative Adversarial Networks through

arXiv: 1709.10367 [stat.ML].

“Learning representations by back-propagating errors”. Cognitive
modeling. 5(3): 1.


Boltzmann machines for collaborative filtering”. In: Proceedings
of the 24th international conference on Machine learning. ACM.
791–798.

strategies as a scalable alternative to reinforcement learning”.

Learning with Dynamic Spiking Neurons and Fixed Feedback

“Distributed methods for constrained nonconvex multi-agent


References


