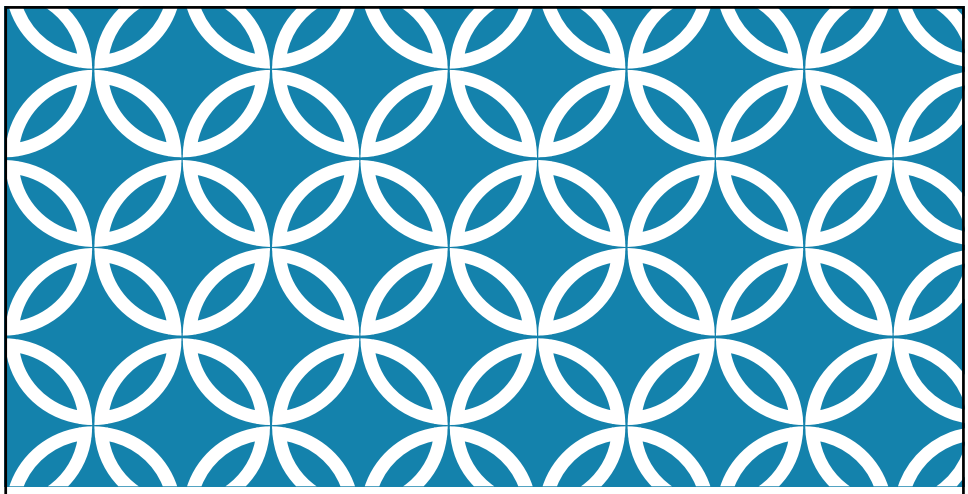


MACHINE LEARNING FOR DUMMIES

Prashant Sahu

OUTLINE

1. Background on machine learning algorithms.
2. Linear machine learning algorithms.
3. Nonlinear machine learning algorithms.
4. Ensemble machine learning algorithms.

WHAT IS LEARNING ?

Definition: A computer program is said to **learn** from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.

For example, a computer program that learns to play chess might improve its performance as measured by its ability to win at the class of tasks involving playing chess games, through experience obtained by playing games against itself.

WHAT IS LEARNING ?

In general, to have a well-defined learning problem, we must identity these three features:

- the class of tasks,
- the measure of performance to be improved, and
- the source of experience.

EXAMPLE LEARNING PROBLEMS

1. A checkers learning problem:

- **Task T:** playing checkers
- **Performance measure P:** percent of games won against opponents
- **Training experience E:** playing practice games against itself

2. A handwriting recognition learning problem:

- **Task T:** recognizing and classifying handwritten words within images
- **Performance measure P:** percent of words correctly classified
- **Training experience E:** a database of handwritten words with given classifications

EXAMPLE LEARNING PROBLEMS

3. A robot driving learning problem:

- **Task T:** driving on public four-lane highways using vision sensors
- **Performance measure P:** average distance travelled before an error (as judged by human overseer)
- **Training experience E:** a sequence of images and steering commands recorded while observing a human driver

DESIGNING A LEARNING SYSTEM

- Choosing the Training Experience
- Choosing the Target Function
- Choosing a Representation for the Target Function
- Choosing a Function Approximation Algorithm
  - Estimating Training Values
  - Adjusting The Weights

Task: F**or checkers learning problem,** design the above.

SO WHAT EXACTLY IS MACHINE LEARNING?

Machine learning teaches computers to do what comes naturally to humans: **learn from experience.**

Machine learning algorithms use computational methods to "learn" information directly from data without relying on a predetermined equation as a model.

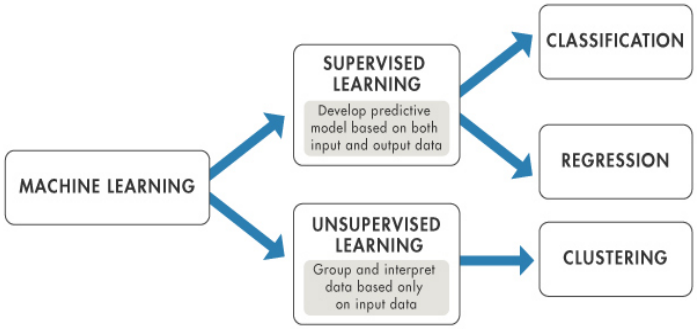
The algorithms adaptively improve their performance as the number of samples available for learning increases.

MACHINE LEARNING TYPES

Machine learning uses two types of techniques:

supervised learning, which trains a model on known input and output data so that it can predict future outputs, and

unsupervised learning, which finds hidden patterns or intrinsic structures in input data.



SUPERVISED MACHINE LEARNING

The aim of supervised machine learning is to build a model that makes predictions based on evidence in the presence of uncertainty. A supervised learning algorithm takes a known set of input data and known responses to the data (output) and trains a model to generate reasonable predictions for the response to new data.

Classification techniques predict categorical responses, for example, whether an email is genuine or spam, or whether a tumor is cancerous or benign. Classification models classify input data into categories.

Typical applications include medical imaging, image and speech recognition, and credit scoring.

Regression techniques predict continuous responses, for example, changes in temperature or fluctuations in power demand.

Typical applications include electricity load forecasting and algorithmic trading.

UNSUPERVISED LEARNING

Unsupervised learning finds hidden patterns or intrinsic structures in data.

It is used to draw inferences from datasets consisting of input data without labelled responses.

Clustering is the most common unsupervised learning technique.

It is used for exploratory data analysis to find hidden patterns or groupings in data.

Applications for clustering include gene sequence analysis, market research, and object recognition.

SELECTING THE RIGHT ALGORITHM

There are dozens of supervised and unsupervised machine learning algorithms, and each takes a different approach to learning.

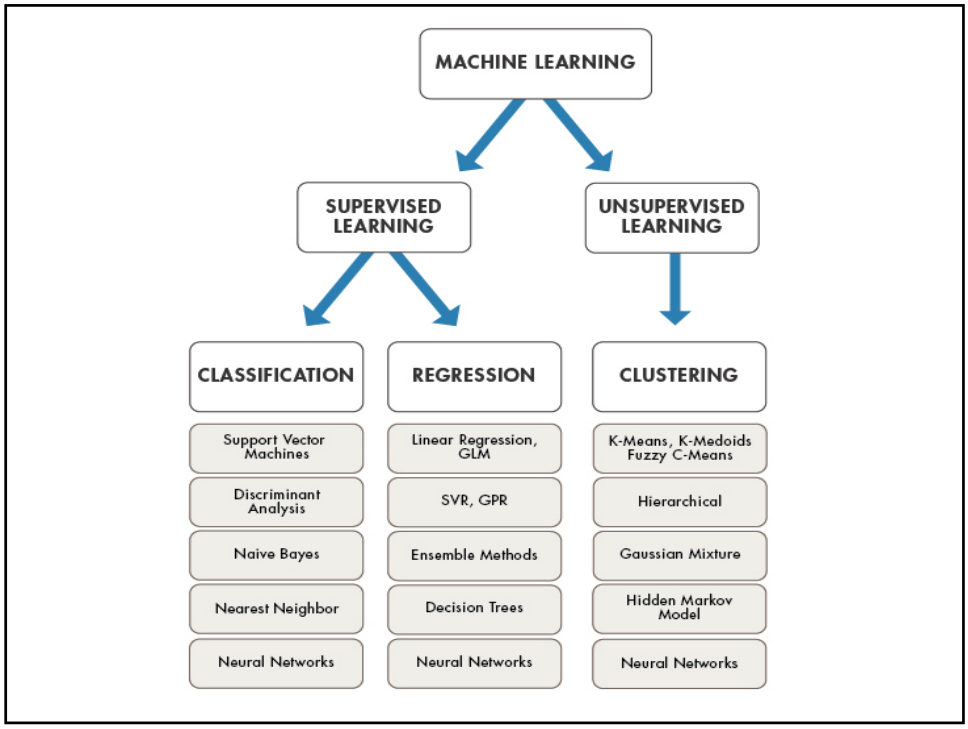
There is no best method or one size fits all. Finding the right algorithm is partly based on trial and error—even highly experienced data scientists cannot tell whether an algorithm will work without trying it out.

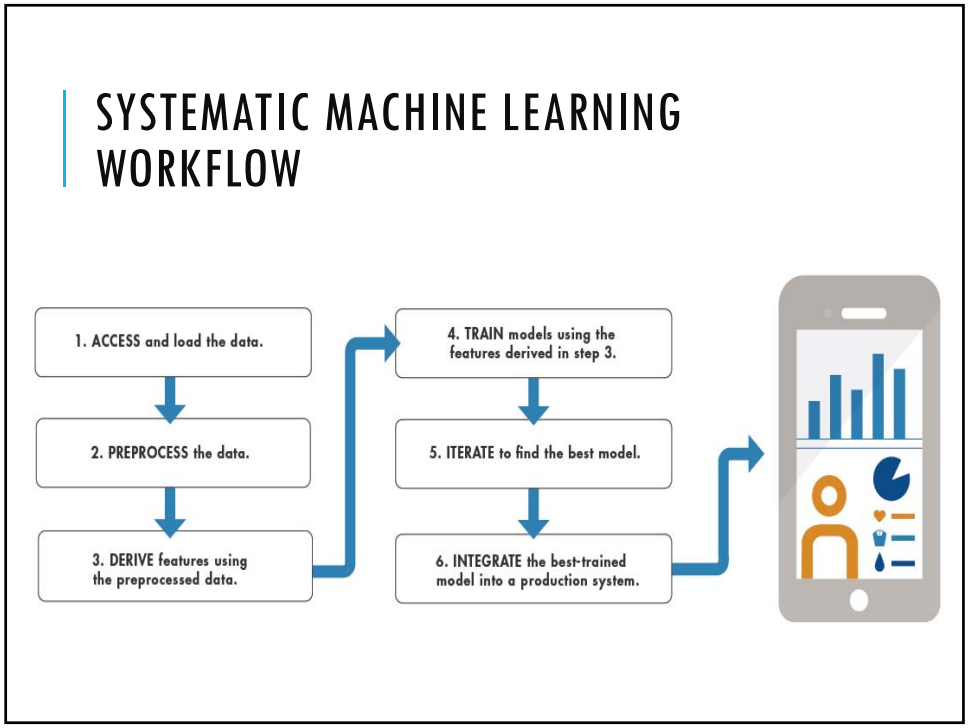
Highly flexible models tend to overfit data by modeling minor variations that could be noise.

Simple models are easier to interpret but might have lower accuracy.

Therefore, choosing the right algorithm requires trading off one benefit against another, including model speed, accuracy, and complexity.

Trial and error is at the core of machine learning—if one approach or algorithm does not work, you try another.


MACHINE LEARNING
SUPERVISED LEARNING
UNSUPERVISED LEARNING
CLASSIFICATION
REGRESSION
CLUSTERING
Support Vector Machines
Discriminant Analysis
Naive Bayes
Nearest Neighbor
Neural Networks
Linear Regression, GLM
SVR, GPR
Ensemble Methods
Decision Trees
Neural Networks
K-Means, K-Medoids Fuzzy C-Means
Hierarchical
Gaussian Mixture
Hidden Markov Model
Neural Networks


SYSTEMATIC MACHINE LEARNING WORKFLOW
1. ACCESS and load the data.
2. PREPROCESS the data.
3. DERIVE features using the preprocessed data.
4. TRAIN models using the features derived in step 3.
5. ITERATE to find the best model.
6. INTEGRATE the best-trained model into a production system.

OUTLINE

1. Background on machine learning algorithms.
2. Linear machine learning algorithms.
3. Nonlinear machine learning algorithms.
4. Ensemble machine learning algorithms.

ALGORITHMS BACKGROUND

This part will give you a foundation in machine learning algorithms.

This will give you the context to be able to understand any machine learning algorithm. You will discover:

- Terminology used in machine learning when describing data.
- The framework for understanding the problem solved by all machine learning algorithms.
- Important differences between parametric and nonparametric algorithms

LINEAR ALGORITHMS

- Gradient descent optimization procedure that may be used in the heart of many machine learning algorithms.
- Linear Regression for predicting real values with two tutorials to make sure it really sinks in.
- Logistic regression for classification on problems with two categories.
- Linear discriminant analysis for classification on problems with more than two categories.

NONLINEAR ALGORITHMS

These are techniques that make fewer assumptions about your problem and are able to learn a large variety of problem types. But this power needs to be used carefully because they can learn too well and overfit your training data.

You will discover the following nonlinear algorithms:

- Classification and regression trees the staple decision tree algorithm.
- Naive Bayes using probability for classification with two tutorials showing you useful ways this technique can be used.
- K-Nearest Neighbors that do not require any model at all other than your dataset.
- Learning Vector Quantization which extends K-Nearest Neighbors by learning to compress your training dataset down in size.
- Support vector machines which are perhaps one of the most popular and powerful out of the box algorithms.

PARAMETRIC AND NONPARAMETRIC MACHINE LEARNING ALGORITHMS

- Parametric machine learning algorithms simply the mapping to a know functional form.
- Nonparametric algorithms can learn any mapping from inputs to outputs.
- All algorithms can be organized into parametric or nonparametric groups.

PARAMETRIC MACHINE LEARNING ALGORITHMS

A learning model that summarizes data with a set of parameters of fixed size (independent of the number of training examples) is called a parametric model.

No matter how much data you throw at a parametric model, it won't change its mind about how many parameters it needs.

The algorithms involve two steps:
1. Select a form for the function.
2. Learn the coefficients for the function from the training data.

PARAMETRIC MACHINE LEARNING ALGORITHMS

Some more examples of parametric machine learning algorithms include:

- Logistic Regression
- Linear Discriminant Analysis
- Perceptron

PARAMETRIC MACHINE LEARNING ALGORITHMS

Benefits of Parametric Machine Learning Algorithms:

- **Simpler**
- **Speed**
- **Less Data**

Limitations of Parametric Machine Learning Algorithms:

- ✓**Constrained**
- ✓**Limited Complexity**
- ✓**Poor Fit**

NONPARAMETRIC MACHINE LEARNING ALGORITHMS

Nonparametric methods are good when you have a lot of data and no prior knowledge, and when you don't want to worry too much about choosing just the right features.

Some more examples of popular nonparametric machine learning algorithms are:

- Decision Trees like CART and C4.5
- Naive Bayes
- Support Vector Machines
- Neural Networks

NONPARAMETRIC MACHINE LEARNING ALGORITHMS

Benefits of Nonparametric Machine Learning Algorithms:

- Flexibility
- Power
- Performance

Limitations of Nonparametric Machine Learning Algorithms:

- More data
- Slower
- Overfitting

SUPERVISED MACHINE LEARNING

Supervised learning is where you have input variables (*X*) and an output variable (*Y*) and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

Classification: A classification problem is when the output variable is a category, such as *red* or *blue* or *disease* and *no disease*.

Regression: A regression problem is when the output variable is a real value, such as *dollars* or *weight*.

SUPERVISED MACHINE LEARNING

Some popular examples of supervised machine learning algorithms are:

- Linear regression for regression problems.
- Random forest for classification and regression problems.
- Support vector machines for classification problems.

UNSUPERVISED MACHINE LEARNING

Unsupervised learning is where you only have input data (X) and no corresponding output variables.

The goal for unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data.

Types of Unsupervised learning problems:

- **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
- **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy A also tend to buy B.

UNSUPERVISED MACHINE LEARNING

Some popular examples of unsupervised learning algorithms are:

- k-means for clustering problems.
- Apriori algorithm for association rule learning problems

SEMI-SUPERVISED MACHINE LEARNING

Problems where you have a large amount of input data (*X*) and only some of the data is labelled (*Y*) are called semi-supervised learning problems.

- These problems sit in between both supervised and unsupervised learning.
- A good example is a photo archive where only some of the images are labeled, (e.g. dog, cat, person) and the majority are unlabelled.