## Introduction

The field of **pattern recognition** is concerned with the automatic discovery of regularities in data through the use of computer algorithms and with the use of these regularities to take actions such as classifying the data into different categories. **Generalization,** the ability to categorize correctly new examples that differ from those used for training is a central goal in **pattern recognition.**

Applications in which the training data comprises examples of the input vectors with their corresponding target vectors are known as **supervised learning** problems. The cases such as the digit recognition, in which the aim is to assign each input vector to one of a finite number of discrete categories, are called **classification.** If the desired output consists of one or more continuous variables, then the task is called **regression.**

The pattern recognition problems in which the training data consists of a set of input vectors x without any corresponding target values are called **unsupervised learning** problems. The goal in such **unsupervised learning** problems may be to discover groups of similar examples within the data, where it is called **clustering**, or to determine the distribution of data within the input space, known as **density estimation,** or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization.

The technique of **reinforcement learning** is concerned with the problem of finding suitable actions to take in a given situation in order to maximize a reward. A general feature of **reinforcement learning** is the trade-off between **exploration,** in which the system tries out new kinds of actions to see how effective they are, and **exploitation,** in which the system makes use of actions that are known to yield a high reward.

## Probability Theory

Probability theory provides a consistent framework for the quantification and manipulation of uncertainty. When combined with decision theory, it allows us to make optimal predictions given all the information available to us, even though that information may be incomplete or ambiguous.

## The Rules of Probability

* Sum Rule:  *p(X) =*
* Product Rule: *p(X, Y) = p(Y / X)p(X)*

Here, *p(X, Y)* is a **joint probability** and is verbalized as “the probability of X and Y”. Similarly, the quantity *p(Y/X)* is a **conditional probability** and is verbalized as “the probability of Y given X”, where the quantity *p(X)* is a **marginal probability** and is verbalized as “the probability of X”.

## Bayes’ Theorem

From the product rule, together with the symmetry property *p(X, Y) = p(Y, X)*, we obtain the following relationship between conditional probabilities which is called **Bayes’ theorem** which plays a central role in pattern recognition and machine learning. Using the sum rule, the denominator in Bayes’ theorem can be expressed in terms of the quantities appearing in the numerator as being the normalization constant required ensuring that the sum of the conditional probability on the left-hand side over all values of *Y* equals one.

* *p(Y / X) =*
* *p(X) =*

## Model Selection

In the maximum likelihood approach, the performance on the training set is not a good indicator of predictive performance on unseen data due to the problem of over-fitting. If data is plentiful, the one approach is simply to use some of the available data to train a range of models, or a given model with a range of values for its complexity parameters, and then to compare them on independent data, sometimes called a **validation set,** and select the one having the best predictive performance. However, in many applications, the data for training and testing will be limited, and in order to build good models, we wish to use as much of the available data as possible for training. If the validation set is small, it will give a relatively noisy estimate of predictive performance. One solution to this problem is to use **cross-validation.** This allows a proportion *(S – 1) / S* of the available data to be used for training while making use of all of the data to assess performance.

## Decision Theory

If we have an input vector **x** together with a corresponding vector **t** of target variables, our goal is to predict **t** given a new value for **x.** For regression problems, **t** will comprise continuous variables, whereas for classification problems, **t** will represent class labels.

## Minimizing Loss Function

Loss function is also called a cost function, which is a single, overall measure of loss incurred in taking any of the available decisions or actions. Our goal is then to minimize the total loss incurred.