Machine Learning

A Bayesian and Optimization Perspective

Academic Press, 2015

Sergios Theodoridis[1]

[1] Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece.

Spring, 2016

Chapter 1: Some Introductory Remarks to Machine Learning

Version I

What Machine Learning Is About?

- Learning through personal experience and knowledge, that propagates from generation to generation, is at the heart of human intelligence.
- Also, at the heart of any scientific field lies the development of models in order to explain the available experimental evidence. In other words, we always **learn from data**.
- Learning from data encompasses techniques that attempt to detect and unveil a possible hidden structure and regularity patterns associated with their generation mechanism. This information will in turn helps the analysis and our understanding the nature of the data, which in the sequel can be used in order to make predictions for the future.
- Besides modeling the underlying structure, a major direction is to develop **efficient** algorithms for designing the models and also for the analysis and prediction.
- Designing efficient algorithms is gaining in importance in the dawn of what we call big data era, where one has to deal with massive number of data, which may be represented in spaces of very large **dimensionality**.
- Computational efficiency and at the same time **robustness** in noise are key issues in designing algorithms.

What Machine Learning Is About?

- Learning through personal experience and knowledge, that propagates from generation to generation, is at the heart of human intelligence.
- Also, at the heart of any scientific field lies the development of models in order to explain the available experimental evidence. In other words, we always **learn from data**.
- Learning from data encompasses techniques that attempt to detect and unveil a possible hidden structure and regularity patterns associated with their generation mechanism. This information will in turn helps the analysis and our understanding the nature of the data, which in the sequel can be used in order to make predictions for the future.
- Besides modeling the underlying structure, a major direction is to develop **efficient** algorithms for designing the models and also for the analysis and prediction.
- Designing efficient algorithms is gaining in importance in the dawn of what we call big data era, where one has to deal with massive number of data, which may be represented in spaces of very large **dimensionality**.
- Computational efficiency and at the same time **robustness** in noise are key issues in designing algorithms.

What Machine Learning Is About?

- Learning through personal experience and knowledge, that propagates from generation to generation, is at the heart of human intelligence.
- Also, at the heart of any scientific field lies the development of models in order to explain the available experimental evidence. In other words, we always **learn from data**.
- Learning from data encompasses techniques that attempt to detect and unveil a possible hidden structure and regularity patterns associated with their generation mechanism. This information will in turn helps the analysis and our understanding the nature of the data, which in the sequel can be used in order to make predictions for the future.
- Besides modeling the underlying structure, a major direction is to develop **efficient** algorithms for designing the models and also for the analysis and prediction.
- Designing efficient algorithms is gaining in importance in the dawn of what we call big data era, where one has to deal with massive number of data, which may be represented in spaces of very large **dimensionality**.
- Computational efficiency and at the same time **robustness** in noise are key issues in designing algorithms.

What Machine Learning Is About?

- Learning through personal experience and knowledge, that propagates from generation to generation, is at the heart of human intelligence.
- Also, at the heart of any scientific field lies the development of models in order to explain the available experimental evidence. In other words, we always **learn from data**.
- Learning from data encompasses techniques that attempt to detect and unveil a possible hidden structure and regularity patterns associated with their generation mechanism. This information will in turn helps the analysis and our understanding the nature of the data, which in the sequel can be used in order to make predictions for the future.
- Besides modeling the underlying structure, a major direction is to develop **efficient** algorithms for designing the models and also for the analysis and prediction.
- Designing efficient algorithms is gaining in importance in the dawn of what we call big data era, where one has to deal with massive number of data, which may be represented in spaces of very large **dimensionality**.
- Computational efficiency and at the same time **robustness** in noise are key issues in designing algorithms.

What Machine Learning Is About?

- Learning through personal experience and knowledge, that propagates from generation to generation, is at the heart of human intelligence.
- Also, at the heart of any scientific field lies the development of models in order to explain the available experimental evidence. In other words, we always **learn from data**.
- Learning from data encompasses techniques that attempt to detect and unveil a possible hidden structure and regularity patterns associated with their generation mechanism. This information will in turn helps the analysis and our understanding the nature of the data, which in the sequel can be used in order to make predictions for the future.
- Besides modeling the underlying structure, a major direction is to develop **efficient** algorithms for designing the models and also for the analysis and prediction.
- Designing efficient algorithms is gaining in importance in the dawn of what we call big data era, where one has to deal with massive number of data, which may be represented in spaces of very large **dimensionality**.
- Computational efficiency and at the same time **robustness** in noise are key issues in designing algorithms.

What Machine Learning Is About?

- Learning through personal experience and knowledge, that propagates from generation to generation, is at the heart of human intelligence.
- Also, at the heart of any scientific field lies the development of models in order to explain the available experimental evidence. In other words, we always **learn from data**.
- Learning from data encompasses techniques that attempt to detect and unveil a possible hidden structure and regularity patterns associated with their generation mechanism. This information will in turn helps the analysis and our understanding the nature of the data, which in the sequel can be used in order to make predictions for the future.
- Besides modeling the underlying structure, a major direction is to develop **efficient** algorithms for designing the models and also for the analysis and prediction.
- Designing efficient algorithms is gaining in importance in the dawn of what we call big data era, where one has to deal with massive number of data, which may be represented in spaces of very large **dimensionality**.
- Computational efficiency and at the same time **robustness** in noise are key issues in designing algorithms.

Classification

- The goal in a classification task is to classify an unknown pattern to one out of a number of classes, which are considered to be known a-priori.
- For example, in X-ray mammography, we are given an image where a region indicates the existence of a tumor. The goal of a computer-aided diagnosis system is to predict whether this tumor corresponds to the benign or the malignant class.
- The first step in designing any Machine Learning task is to decide on how to **represent** each pattern in the computer. One has to encode related information that resides in the raw data in an efficient and information-rich way.
- Features and feature vectors: The data representation is, usually, done by transforming the raw data (measured by a sensing device) into a new space and each pattern is represented by a vector, $x \in \mathbb{R}^l$. This is known as the feature vector, and its l elements as the features. In this way, each patten becomes a single point in an l-dimensional space, known as the feature space or the input space.
- This preprocessing is known as **feature generation** stage. Usually, one starts with some large value, K, of features and selects the l most informative ones. The latter is known as the **feature selection** stage.

Classification

- The goal in a classification task is to classify an unknown pattern to one out of a number of classes, which are considered to be known a-priori.
- For example, in X-ray mammography, we are given an image where a region indicates the existence of a tumor. The goal of a computer-aided diagnosis system is to predict whether this tumor corresponds to the benign or the malignant class.
- The first step in designing any Machine Learning task is to decide on how to **represent** each pattern in the computer. One has to encode related information that resides in the raw data in an efficient and information-rich way.
- Features and feature vectors: The data representation is, usually, done by transforming the raw data (measured by a sensing device) into a new space and each pattern is represented by a vector, $x \in \mathbb{R}^l$. This is known as the feature vector, and its l elements as the features. In this way, each patten becomes a single point in an l-dimensional space, known as the feature space or the input space.
- This preprocessing is known as **feature generation** stage. Usually, one starts with some large value, K, of features and selects the l most informative ones. The latter is known as the **feature selection** stage.

Classification

- The goal in a classification task is to classify an unknown pattern to one out of a number of classes, which are considered to be known a-priori.
- For example, in X-ray mammography, we are given an image where a region indicates the existence of a tumor. The goal of a computer-aided diagnosis system is to predict whether this tumor corresponds to the benign or the malignant class.
- The first step in designing any Machine Learning task is to decide on how to **represent** each pattern in the computer. One has to encode related information that resides in the raw data in an efficient and information-rich way.
- Features and feature vectors: The data representation is, usually, done by transforming the raw data (measured by a sensing device) into a new space and each pattern is represented by a vector, $x \in \mathbb{R}^l$. This is known as the feature vector, and its l elements as the features. In this way, each patten becomes a single point in an l-dimensional space, known as the feature space or the input space.
- This preprocessing is known as **feature generation** stage. Usually, one starts with some large value, K, of features and selects the l most informative ones. The latter is known as the **feature selection** stage.

Classification

- The goal in a classification task is to classify an unknown pattern to one out of a number of classes, which are considered to be known a-priori.
- For example, in X-ray mammography, we are given an image where a region indicates the existence of a tumor. The goal of a computer-aided diagnosis system is to predict whether this tumor corresponds to the benign or the malignant class.
- The first step in designing any Machine Learning task is to decide on how to **represent** each pattern in the computer. One has to encode related information that resides in the raw data in an efficient and information-rich way.
- Features and feature vectors: The data representation is, usually, done by transforming the raw data (measured by a sensing device) into a new space and each pattern is represented by a vector, $\boldsymbol{x} \in \mathbb{R}^l$. This is known as the feature vector, and its l elements as the features. In this way, each patten becomes a single point in an l-dimensional space, known as the feature space or the input space.
- This preprocessing is known as **feature generation** stage. Usually, one starts with some large value, K, of features and selects the l most informative ones. The latter is known as the **feature selection** stage.

Classification

- The goal in a classification task is to classify an unknown pattern to one out of a number of classes, which are considered to be known a-priori.
- For example, in X-ray mammography, we are given an image where a region indicates the existence of a tumor. The goal of a computer-aided diagnosis system is to predict whether this tumor corresponds to the benign or the malignant class.
- The first step in designing any Machine Learning task is to decide on how to **represent** each pattern in the computer. One has to encode related information that resides in the raw data in an efficient and information-rich way.
- Features and feature vectors: The data representation is, usually, done by transforming the raw data (measured by a sensing device) into a new space and each pattern is represented by a vector, $\boldsymbol{x} \in \mathbb{R}^l$. This is known as the feature vector, and its l elements as the features. In this way, each patten becomes a single point in an l-dimensional space, known as the feature space or the input space.
- This preprocessing is known as **feature generation** stage. Usually, one starts with some large value, K, of features and selects the l most informative ones. The latter is known as the **feature selection** stage.

Classification

- Having decided upon the input space, one has to train a classifier. This is achieved by first selecting a set of data, whose class is known and comprise the **training set**. This is a set of pairs, $(y_n, \boldsymbol{x}_n),\ \ n = 1, \ldots, N$, where y_n is the (output) variable denoting the class in which $\boldsymbol{x}_n$ belongs, and it is known as the corresponding class label; the class labels, y, take values over a **discrete** set, e.g., $\{1, 2, \ldots, M\}$, for an M-class classification task.
- Based on the training data, one then designs a function, f, which **predicts** the output label, given the input. This function is known as the **classifier**. In general, we need to design a set of such functions, but for the time being let us keep our discussion simple.
- Once the classifier has been designed, the system is ready for predictions. Given an unknown pattern, we form the corresponding feature vector, x, from the raw data, and depending on the value of $\hat{y} = f(x)$, the pattern is classified into one of the classes.

Classification

- Having decided upon the input space, one has to train a classifier. This is achieved by first selecting a set of data, whose class is known and comprise the **training set**. This is a set of pairs, $(y_n, \boldsymbol{x}_n)$, $n = 1, \ldots, N$, where y_n is the (output) variable denoting the class in which $\boldsymbol{x}_n$ belongs, and it is known as the corresponding class label; the class labels, y, take values over a **discrete** set, e.g., $\{1, 2, \ldots, M\}$, for an M-class classification task.
- Based on the training data, one then designs a function, f, which **predicts** the output label, given the input. This function is known as the **classifier**. In general, we need to design a set of such functions, but for the time being let us keep our discussion simple.
- Once the classifier has been designed, the system is ready for predictions. Given an unknown pattern, we form the corresponding feature vector, $\boldsymbol{x}$, from the raw data, and depending on the value of $\hat{y} = f(\boldsymbol{x})$, the pattern is classified into one of the classes.

Classification

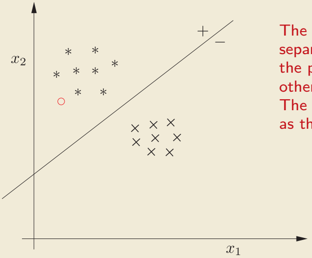
- Having decided upon the input space, one has to train a classifier. This is achieved by first selecting a set of data, whose class is known and comprise the **training set**. This is a set of pairs, $(y_n, \boldsymbol{x}_n),\ n = 1, \ldots, N$, where y_n is the (output) variable denoting the class in which $\boldsymbol{x}_n$ belongs, and it is known as the corresponding class label; the class labels, y, take values over a **discrete** set, e.g., $\{1, 2, \ldots, M\}$, for an M-class classification task.
- Based on the training data, one then designs a function, f, which **predicts** the output label, given the input. This function is known as the **classifier**. In general, we need to design a set of such functions, but for the time being let us keep our discussion simple.
- Once the classifier has been designed, the system is ready for predictions. Given an unknown pattern, we form the corresponding feature vector, $\boldsymbol{x}$, from the raw data, and depending on the value of $\hat{y} = f(\boldsymbol{x})$, the pattern is classified into one of the classes.

Classification

- The figure below illustrates the classification task. Initially, we are given the set of training points in the two-dimensional space (two features used, x_1, x_2). Stars belong to one class, say ω_1 and the crosses to the other, ω_2, in a two-class classification task. Based on these points, a classifier was learned; for our very simple case, this turned out to be a linear function, i.e.,

$$f(\boldsymbol{x}) = \theta_1 x_1 + \theta_2 x_2 + \theta_0,$$

whose graph for all the points such as: $f(\boldsymbol{x}) = 0$, is the straight line shown in the figure.



The classifier (linear in this simple case) has been designed so that to separate the training data in the two classes, having on its positive side the points coming from one class and on its negative side those of the other.
The "red" point, whose class is unknown, is classified to the same class as the "star" points, since it lies on the **positive side** of the classifier.

Supervised, Unsupervised and Semisupervised Learning

- The previously described type of learning based on the use of learning data is known as supervised learning. Note that the training data can be seen as the available previous experience, and based on this, one builds a model to make predictions for the future.
- In unsupervised/clustering no training data is available and the task is to recover the groups /clusters in which the available data are clustered together. Data in the same cluster are considered to be more similar than those belonging to different clusters.
- In semisupervised learning, there are some training data, but not enough to fully learn the model.

Supervised, Unsupervised and Semisupervised Learning

- The previously described type of learning based on the use of learning data is known as supervised learning. Note that the training data can be seen as the available previous experience, and based on this, one builds a model to make predictions for the future.
- In unsupervised/clustering no training data is available and the task is to recover the groups /clusters in which the available data are clustered together. Data in the same cluster are considered to be more similar than those belonging to different clusters.
- In semisupervised learning, there are some training data, but not enough to fully learn the model.

Supervised, Unsupervised and Semisupervised Learning

- The previously described type of learning based on the use of learning data is known as supervised learning. Note that the training data can be seen as the available previous experience, and based on this, one builds a model to make predictions for the future.
- In unsupervised/clustering no training data is available and the task is to recover the groups /clusters in which the available data are clustered together. Data in the same cluster are considered to be more similar than those belonging to different clusters.
- In semisupervised learning, there are some training data, but not enough to fully learn the model.

Regression

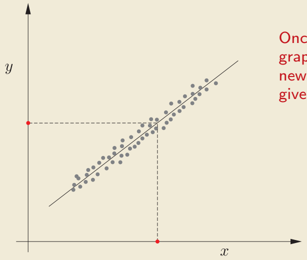
- The regression task shares, to a large extent, the feature generation/selection stage, as described before; however, now the output variable, y, is not discrete but it takes values, e.g., in an interval in the real axis or in a region in the complex numbers plane. The regression task is basically a curve fitting problem.
- We are given a set of training points, $(y_n, \boldsymbol{x}_n)$, $y_n \in \mathbb{R}$, $\boldsymbol{x}_n \in \mathbb{R}^l$, $n = 1, 2, \ldots, N$, and the task is to estimate a function, f, whose graph fits the data. Once we have found such a function, when an unknown point arrives we can predict its output value.

Regression

- The regression task shares, to a large extent, the feature generation/selection stage, as described before; however, now the output variable, y, is not discrete but it takes values, e.g., in an interval in the real axis or in a region in the complex numbers plane. The regression task is basically a curve fitting problem.
- We are given a set of training points, $(y_n, \boldsymbol{x}_n)$, $y_n \in \mathbb{R}$, $\boldsymbol{x}_n \in \mathbb{R}^l$, $n = 1, 2, \ldots, N$, and the task is to estimate a function, f, whose graph fits the data. Once we have found such a function, when an unknown point arrives we can predict its output value.

Regression

- The regression task shares, to a large extent, the feature generation/selection stage, as described before; however, now the output variable, y, is not discrete but it takes values, e.g., in an interval in the real axis or in a region in the complex numbers plane. The regression task is basically a curve fitting problem.
- We are given a set of training points, $(y_n, \boldsymbol{x}_n)$, $y_n \in \mathbb{R}$, $\boldsymbol{x}_n \in \mathbb{R}^l$, $n = 1, 2, \ldots, N$, and the task is to estimate a function, f, whose graph fits the data. Once we have found such a function, when an unknown point arrives we can predict its output value.
- This is shown in the following figure.



Once a function (linear in this case), f, has been designed, so as its graph to fit the available training data set in a regression task, given a new (red) point, x, the prediction of the associated output (red) value is given by $\hat{y} = f(x)$.

Regression

- The regression task is a generic task that embraces a number of problems. For example, in financial applications one can predict tomorrow's stock market price given current market conditions and all other related information. Each piece of information is a measured value of a corresponding feature.
- Signal and image restoration and denoising come under this common umbrella of regression tasks. Figure (a) shows a blurred image, taken by a moving camera, and Figure (b) the deblurred one. The de-blurred image is obtained as the output, by feeding the blurred one as input to an appropriately designed function.

Regression

- The regression task is a generic task that embraces a number of problems. For example, in financial applications one can predict tomorrow's stock market price given current market conditions and all other related information. Each piece of information is a measured value of a corresponding feature.
- Signal and image restoration and denoising come under this common umbrella of regression tasks. Figure (a) shows a blurred image, taken by a moving camera, and Figure (b) the deblurred one. The de-blurred image is obtained as the output, by feeding the blurred one as input to an appropriately designed function.

(a)

(b)