CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

Part I

Density-Ratio Approach to Machine Learning

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

1

Introduction

The goal of *machine learning* is to extract useful information hidden in data (Hastie et al., 2001; Schölkopf and Smola, 2002; Bishop, 2006). This chapter is devoted to describing a brief overview of the machine learning field and showing our focus in this book – *density-ratio methods*. In Section 1.1, fundamental machine learning frameworks of *supervised learning*, *unsupervised learning*, and *reinforcement learning* are briefly reviewed. Then we show examples of machine learning problems to which the density-ratio methods can be applied in Section 1.2 and briefly review methods of density-ratio estimation in Section 1.3. A brief overview of theoretical aspects of density-ratio estimation is given in Section 1.4. Finally, the organization of this book is described in Section 1.5.

1.1 Machine Learning

Depending on the type of data and the purpose of the analysis, machine learning tasks can be classified into three categories:

Supervised learning: An input–output relation is learned from input–output samples.

Unsupervised learning: Some interesting "structure" is found from input-only samples.

Reinforcement learning: A decision-making policy is learned from reward samples.

In this section we briefly review each of these tasks.

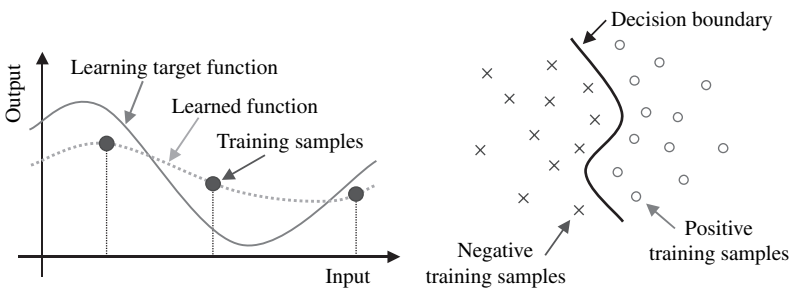
1.1.1 Supervised Learning

In the *supervised learning* scenario, data samples take the form of input–output pairs and the goal is to infer the input–output relation behind the data. Typical examples of supervised learning problems are *regression* and *classification* (Figure 1.1):

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information



(a) Regression (b) Classification

Figure 1.1. Regression and classification tasks in supervised learning. The goal of regression is to learn the target function from training samples, while the goal of classification is to learn the decision boundary from training samples.

Regression: Real-valued output values are predicted. A distance structure exists in the output space, thus making it important that the prediction be "close" to the true output.

Classification: Categorical output values are predicted. No distance structure exists in the output space, and thus the only thing that matters is whether the prediction is correct.

Designing learning algorithms for making good predictions is the central research topic in supervised learning. Beyond that, there are various challenging research issues such as *model selection*, *active learning*, and *dimensionality reduction* (Figure 1.2):

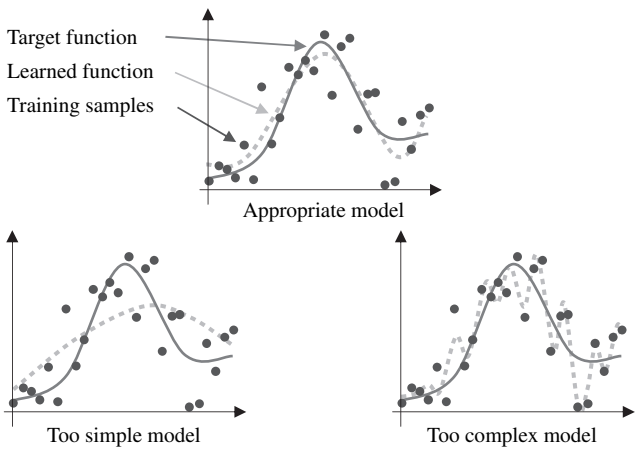
Model selection: To obtain good prediction performance in supervised learning, it is critical to control the *complexity* of models appropriately. Here a *model* refers to a set of functions from which a learned function is searched (Akaike, 1970, 1974, 1980; Mallows, 1973; Takeuchi, 1976; Schwarz, 1978; Rissanen, 1978; Craven and Wahba, 1979; Rissanen, 1987; Shibata, 1989; Wahba, 1990; Efron and Tibshirani, 1993; Murata et al., 1994; Konishi and Kitagawa, 1996; Ishiguro et al., 1997; Vapnik, 1998; Sugiyama and Ogawa, 2001b; Sugiyama and Müller, 2002; Sugiyama et al., 2004).

Active learning: When users are allowed to design the location of training input points, it is desirable to design the location so that the prediction performance is maximized (Fedorov, 1972; Pukelsheim, 1993; Cohn et al., 1996; Fukumizu, 2000; Wiens, 2000; Sugiyama and Ogawa, 2000, 2001a; Kanamori and Shimodaira, 2003; Sugiyama, 2006; Kanamori, 2007; Sugiyama and Nakajima, 2009). Active learning is also called the *experiment design* in statistics. The challenging problem of solving active learning and model selection simultaneously has also been explored (Sugiyama and Ogawa, 2003; Sugiyama and Rubens, 2008).

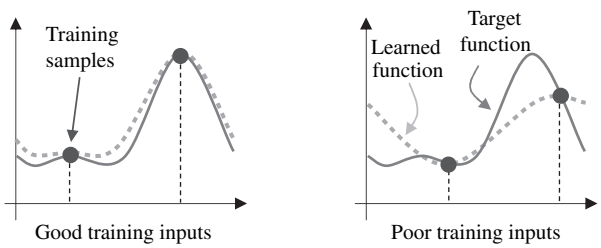
© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

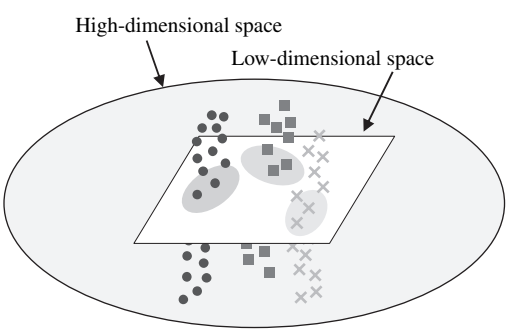
Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information



(a) Model selection



(b) Active learning



(c) Dimensionality reduction

Figure 1.2. Research topics in supervised learning. The goal of model selection is to appropriately control the complexity of a function class from which a learned function is searched. The goal of active learning is to find good location of training input points. The goal of dimensionality reduction is to find a suitable low-dimensional expression of training samples for predicting output values.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

Dimensionality reduction: As the dimensionality of input variables gets higher, the supervised learning problem becomes more and more difficult. Because of its hardness, this phenomenon is often referred to as the *curse of dimensionality* (Bellman, 1961; Vapnik, 1998). One way to mitigate the curse of dimensionality is to assume that the data at hand are redundant in some sense and try to remove such redundant components. Various dimensionality reduction techniques have been developed (Fukunaga, 1990; Li, 1991, 1992; Fukumizu et al., 2004; Weinberger et al., 2006; Globerson and Roweis, 2006; Sugiyama, 2007, 2009; Davis et al., 2007; Song et al., 2007b; Suzuki and Sugiyama, 2010; Sugiyama et al., 2010c). Dimensionality reduction is also referred to as *feature extraction*. If the reduction is confined to choosing a subset of attributes of the original high-dimensional input vector, it is called *feature selection* or *variable selection*. Unsupervised dimensionality reduction methods (i.e., output information is not utilized) are also often employed, even in supervised learning scenarios (see Section 1.1.2).

1.1.2 Unsupervised Learning

In contrast to supervised learning, where input–output samples are provided, data samples containing only input information are available in an unsupervised learning scenario. The goal of *unsupervised learning* is to discover some "interesting" structure hidden in the input-only data.

Typical examples of unsupervised learning problems are (Figure 1.3)

Visualization: High-dimensional data samples are projected onto a space with dimension no more than 3. Essentially, data visualization is almost the same as dimensionality reduction (Friedman and Tukey, 1974; Huber, 1985; Friedman, 1987; Oja, 1982, 1989; Jolliffe, 1986; Schölkopf et al., 1998; Roweis and Saul, 2000; Tenenbaum et al., 2000; Saul and Roweis, 2003; Belkin and Niyogi, 2003; Donoho and Grimes, 2003; He and Niyogi, 2004; Hinton and Salakhutdinov, 2006; Blanchard et al., 2006; Kawanabe et al., 2007).

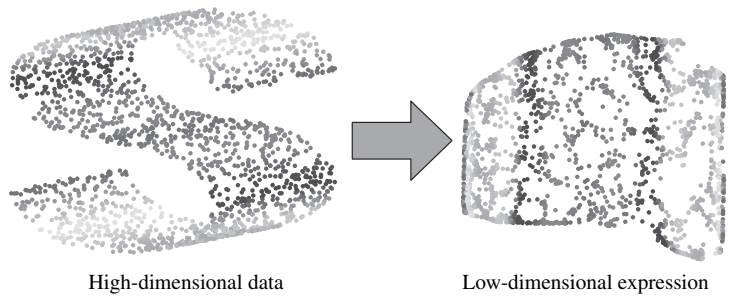
Clustering: Data samples are grouped into several clusters based on their similarity/distance (MacQueen 1967; Antoniak 1974; Hartigan 1975; Kohonen 1988; Jain and Dubes 1988; Buhmann 1995; Kohonen 1995; Shi and Malik 2000; Meila and Heckerman 2001; Ng et al. 2002; Bach and Jordan 2006; Dhillon et al. 2004; Xu et al. 2005; Bach and Harchaoui 2008; Zelnik-Manor and Perona 2005; Blei and Jordan 2006; Song et al. 2007a; Faivishevsky and Goldberger 2010; Kimura and Sugiyama 2011; Agakov andBarber 2006; Gomes et al. 2010; Sugiyama et al. 2011d).

Outlier detection: In real-world problems, data samples often contain *outliers* (i.e., "irregular" samples) because of measurement error or human mislabeling. The goal of outlier detection is to identify outliers in a dataset (Breunig et al., 2000; Schölkopf et al., 2001; Tax and Duin, 2004; Hodge and Austin, 2004; Hido et al., 2011). The problem of outlier detection is also referred to as *anomaly detection*, *novelty detection*, and *single-class classification*.

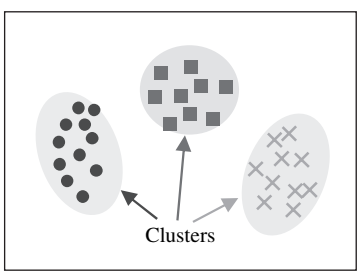
© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

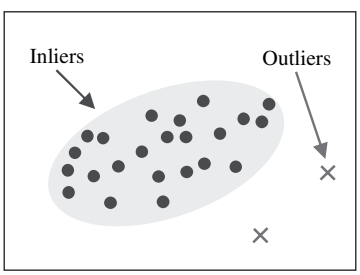
Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information



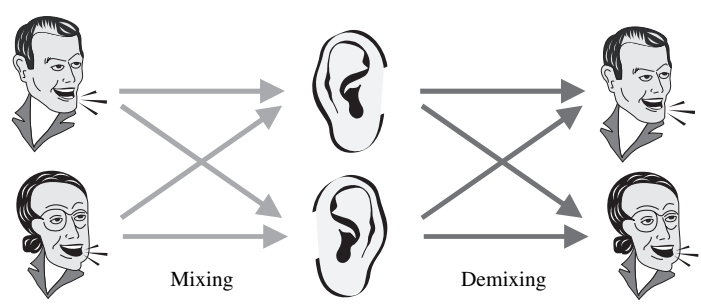
(a) Visualization



(b) Clustering



(c) Outlier detection



(d) Independent component analysis

Figure 1.3. Research topics in unsupervised learning. The goal of visualization is to find a low-dimensional expression of samples that provides us some intuition behind the data. The goal of clustering is to group samples into several clusters. The goal of outlier detection is to identify "irregular" samples. The goal of independent component analysis is to extract original source signals from their mixed signals.

Independent component analysis: Independent component analysis is a machine learning approach to the problem of *blind source separation*, which is also known as the *cocktail party problem*. The goal of blind source separation is to extract the original source signals from their mixed signals. Independent component analysis methods tackle this problem based on the statistical independence among the source signals (Jutten and Herault, 1991; Cardoso and Souloumiac,

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

1993; Comon, 1994; Amari et al., 1996; Amari, 1998, 2000; Hyvaerinen, 1999; Cardoso, 1999; Lee et al., 1999; Hyvärinen et al., 2001; Bach and Jordan, 2002; Hulle, 2008; Suzuki and Sugiyama, 2011).

An intermediate situation between supervised and unsupervised learning called *semi-supervised learning* has also become a major topic of interest recently, where input–output samples and input-only samples are given (Chapelle et al., 2006). The goal of semi-supervised learning is still the same as supervised learning. Thus, the only difference between semi-supervised learning and supervised learning is that, in semi-supervised learning, additional input-only samples are provided, based on which we expect that prediction performance can be improved.

1.1.3 Reinforcement Learning

Reinforcement learning (Sutton and Barto, 1998) is a framework of learning a decision-making policy for computer agents through interaction with the surrounding environment. Since the goal is to learn a policy function, it is similar to supervised learning. However, no explicit supervision is available, that is, an action to take is not explicitly provided. Thus, it is also similar to unsupervised learning.

Without any supervision, it is not possible to learn a policy function in a meaningful way. A key assumption in reinforcement learning is that implicit supervision called *rewards* is provided. The reward information reflects the appropriateness of the action the agent takes at that time. Thus, intuitively, the action with the maximum reward is regarded as the best choice. However, a greedy strategy of maximizing the immediate reward does not necessarily lead to the maximization of the *long-term cumulative rewards*. For example, a prior investment can produce more profits in the long run. The goal of reinforcement learning is to let the agent learn the control policy that maximizes the long-term cumulative rewards (Figure 1.4).

Regression methods are often utilized to solve efficiently the reinforcement learning problem (Lagoudakis and Parr, 2003). Furthermore, various machine learning techniques such as clustering, active learning, and dimensionality

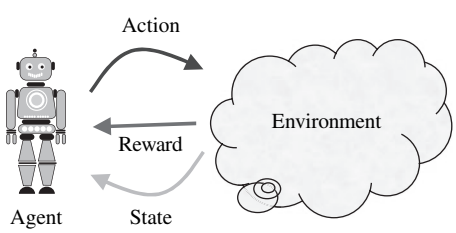


Figure 1.4. Reinforcement learning. If the agent takes some action, then the next state and the reward are given from the environment. The goal of reinforcement learning is to let the agent learn the control policy that maximizes the long-term cumulative rewards.

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

reduction are highly useful for solving realistic reinforcement learning problems. Thus, reinforcement learning can be regarded as a challenging application of machine learning methods. Following the rapid advancement of machine learning techniques and computer environments in the last decade, reinforcement learning algorithms have become capable of handling large-scale, complex, real-world problems. For this reason, reinforcement learning has gathered considerable attention recently in the machine learning community.

1.2 Density-Ratio Approach to Machine Learning

As described in the previous section, machine learning includes various kinds of important and practical data-processing tasks. For this reason, machine learning become one of the most challenging and growing research fields in the areas of computer science and related data processing fields. Among the various approaches, statistical methods have particularly achieved great success in the last decade, when the size, dimension, and complexity of the data have grown explosively.

In a statistical approach, the data samples are assumed to be drawn from an underlying probability distribution. Thus, the most fundamental statistical approach tries to estimate the underlying probability distribution from samples. Indeed, virtually all machine learning problems can be solved via probability distribution estimation. However, since probability distribution estimation is known to be one of the most difficult problems (Vapnik, 1998), solving a target machine learning task via probability distribution estimation can be highly erroneous.

Recently, an alternative framework of machine learning based on the *ratio* of probability densities has been proposed and has gathered a great deal of attention (Sugiyama et al., 2009). The purpose of this book is to give an overview of the density-ratio framework including algorithms, applications, and theories.

The density-ratio framework of machine learning includes various statistical data-processing tasks, which are extensively covered in Part III of this book.

Non-stationarity adaptation (Section 9.1): Ordinary supervised learning methods have been developed under the assumption that samples used for training a regressor/classifier follow the same probability distribution as test samples whose outputs are predicted. Under the common distribution assumption, ordinary learning algorithms are designed to have good theoretical properties such as *consistency* and *asymptotic unbiasedness*.

However, this fundamental assumption is often violated in real-world problems such as robot control, speech recognition, image recognition, brain-signal analysis, natural language processing, bioinformatics, and computational chemistry. The goal of non-stationarity adaptation research is to develop learning algorithms that perform well even when the training and test samples follow different probability distributions. If the problem domains of the training and test data are different, the adaptation problem is called *domain adaptation* or

© in this web service Cambridge University Press www.cambridge.org

CAMBRIDGE

Cambridge University Press
978-0-521-19017-6 - Density Ratio Estimation in Machine Learning
Masashi Sugiyama, Taiji Suzuki and Takafumi Kanamori
Excerpt
More information

transfer learning. In econometrics, this problem has been studied under the name of *sample selection bias* (Heckman, 1979).

When the training and test distributions have nothing in common, it is not possible to learn anything about the test distribution from the training samples. Thus, some similarity between training and test distributions needs to be assumed to make the discussion meaningful. Here we focus on the situation called a *covariate shift*, where the training and test *input* distributions are different but the conditional distribution of outputs given inputs is common to the training and test samples (Shimodaira, 2000). Note that "covariate" refers to an input variable in statistics.

Under the covariate shift setup, the density ratio between the training and test densities can be used as the measure of the "importance" of each training sample in the test domain. This approach can be regarded as an application of the *importance sampling* technique in statistics (Fishman, 1996). By weighting the training loss function according to the importance value, ordinary supervised learning techniques can be systematically modified so that suitable theoretical properties such as consistency and asymptotic unbiasedness can be properly achieved even under covariate shift (Shimodaira, 2000; Zadrozny, 2004; Sugiyama and Müller, 2005; Sugiyama et al., 2007, 2008; Storkey and Sugiyama, 2007; Huang et al., 2007; Yamazaki et al., 2007; Bickel et al., 2007; Kanamori et al., 2009; Quiñonero-Candela et al., 2009; Sugiyama and Kawanabe, 2011).

Multi-task learning (Section 9.2): When one wants to solve many supervised learning problems, each of which contains a small number of training samples, solving them simultaneously could be more promising than solving them independently if the learning problems possess some similarity. The goal of multi-task learning is to solve multiple related learning problems accurately based on task similarity (Caruana et al., 1997; Baxter, 1997, 2000; Ben-David et al., 2002; Bakker and Heskes, 2003; Ben-David and Schuller, 2003; Evgeniou and Pontil, 2004; Micchelli and Pontil, 2005; Yu et al., 2005; Ando and Zhang, 2005; Xue et al., 2007; Kato et al., 2010).

The essence of multi-task learning is data sharing among different tasks, which can also be carried out systematically by using the *importance sampling* technique (Bickel et al., 2008). Thus, a multi-task learning problem can be handled in the same way as non-stationarity adaptation in the density ratio framework.

Outlier detection (Section 10.1): The goal of outlier detection is to identify outliers in a dataset. In principle, outlier-detection problems can be solved in a supervised way by learning a classification rule between the outliers and inliers based on outlier and inlier samples. However, because outlier patterns are often so diverse and their tendency may change over time in practice, such a supervised learning approach is not necessarily appropriate. *Semi-supervised* approaches are also being explored (Gao et al., 2006a, 2006b) but they still suffer from the same limitation. For this reason, it is common to tackle the

© in this web service Cambridge University Press www.cambridge.org