IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 1, January 2015.
www.ijiset.com
ISSN 2348 – 7968

A Brief STUDIES on Machine Learning Techniques

Deepika Trivedi[1], Seema Bissa Harsh[2] and Rajesh Kumar[3]

[1] Department of Computer Science / Arya Institute of Engineering & Technology, Jaipur, Rajasthan 302012/ India

[2] Department of Computer Science, ,B.J.S. Rampuria jain college Bikaner, Rajasthan / India

[3] Department of Computer Science / Arya Institute of Engineering & Technology, Jaipur, Rajasthan 302012/ India

Abstract

Machine-learning Techniques [MLT] has been making great progress in many directions. This article summarizes these learning Techniques. In the field of *Machine Learning* Techniques one considers the important cross-examine of how to make machines able to "cognize". This article provides a concise and personal view of the discipline that has emerged as Machine Learning and the three machine learning methods we employed (Naive Bayes, maximum entropy Classification and support vector machines). The machine learning fields and mathematical programming are increasingly interweave.

Keywords: Machine Learning Techniques (MLT), Naive Bayes, maximum entropy classification, support vector machines.

1. Introduction

Machine Learning is the study of computational methods for improving performance by mechanizing the acquisition of knowledge from experience. Machine learning methods have been used to study the genotype-phenotype relationship in genetic epidemiology. Machine learning systems automatically learn programs from data. Machine learning is a scientific discipline that deals with the construction and study of algorithms that can learn from data. Machine learning can be considered a subfield statistics data analysis.

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed. Machine learning focuses on the development of computer programs that can teach themselves to grow and change when exposed to new data.

The process of machine learning is similar to that of data mining. Both systems search through data to look for patterns. However, instead of extracting data for human comprehension - as is the case in data mining applications - machine learning uses that data to improve the program's own understanding. Machine learning programs detect patterns in data and adjust program actions accordingly.

For example, Facebook's News Feed changes according to the user's personal interactions with other users. If a user frequently tags a friend in photos, writes on his wall or "likes" his links, the News Feed will show more of that friend's activity in the user's News Feed due to presumed closeness.

2. Machine Learning Methods

Two of the most widely adopted machine learning methods are **supervised** and **unsupervised**. Most machine learning – about 70 percent – is supervised learning. Unsupervised learning accounts for 10 to 20 percent.

2.1 Supervised learning algorithms are trained using labelled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labelled either "F" (failed) or "R" (runs). The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with the correct outputs to find errors. It then modifies the model accordingly. Through methods like classification, regression,

ISSN 2348 – 7968

prediction and gradient boosting, supervised learning uses patterns to predict the values of the label on additional unlabeled data. Supervised learning is commonly used in applications where historical data predicts likely future events. For example, it can anticipate when credit card transactions are likely to be fraudulent or which insurance customer is likely to file a claim.

2.2 Unsupervised learning is used against data that has no historical labels. The system is not told the "right answer." The algorithm must figure out for itself what's being shown. The goal is to explore the data and find some structure within. Unsupervised learning works well on transactional data. For example, it can identify segments of customers with similar attributes who can then be treated similarly in marketing campaigns. Or it can find the main attributes that separate customer segments from each other. Popular techniques include self-organizing maps, nearest-neighbor mapping, k-means clustering and singular value decomposition. These algorithms are also used to segment text topics, recommend items and identify data outliers.

Semi-supervised learning and **reinforcement learning** are two other techniques that are sometimes used.

- **Semi-supervised learning** is used for the same applications as supervised learning. But it uses both labelled and unlabeled data for training – typically a small amount of labelled data with a large amount of unlabeled data (because unlabeled data takes less effort and is less expensive to acquire). This type of learning can be used with methods such as classification, regression and prediction. Semi-supervised learning is useful when the cost associated with labelling is too high to allow for a fully labelled training process. Early examples of this include identifying a person's face on a web cam.

- **Reinforcement learning** is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers for itself through trial and error which actions yield the greatest rewards. This type of learning has three primary components: the agent (the learner or decision maker), the environment (everything the agent interacts with), and actions (what the agent can do). The objective is for the agent to choose actions that maximize the expected reward over a given amount of time. The agent will reach the goal much faster by following a good policy. So the goal in reinforcement learning is to learn the best policy.

3. Naive Bayes

The Naive Bayes classifier is a simple probabilistic classifier which is based on Bayes theorem with strong and naïve independence assumptions. It is one of the most basic text classification techniques with various applications in email spam detection, personal email sorting, document categorization, sexually explicit content detection, language detection and sentiment detection. Despite the naïve design and oversimplified assumptions that this technique uses, Naive Bayes performs well in many complex real-world problems.

One approach to text classification is to assign to a given document d the class $c^*=\arg\max_c P(c|d)$. We drive the naive Bayes(NB) classifier by first observing that by Baye's rule,

$$P(c|d)=P(c)P(c|d)/P(d),$$

Where P(d) plays no role in selecting c*. To estimate the term P(c|d), Naive Baye's decomposes it by assuming the f_i's are conditionally independent given d's class:

ISSN 2348 – 7968

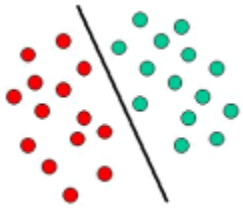
$$P_{NB}(c|d):=P(c)(II_{i=1}{}^{m}\, P(f_i|c)^{ni()d})/P(d)$$

Our training method consists of relative-frequency estimation of P(c) and $P(f_{i|}\ c)$, using add-one smoothing.
Despite its simplicity and the fact that its conditional independence assumption clearly does not hold in real-world situations, Naive Bayes-based text categorization still tends to perform surprisingly well(Lewis,1998);indeed, domingos and Pazzani(1997) show that Naive Bayes is optimal for certain problem classes with highly dependent features. On the other hand, more sophisticated algorithm might(and often do) yield better results; we examine two such algorithms next.

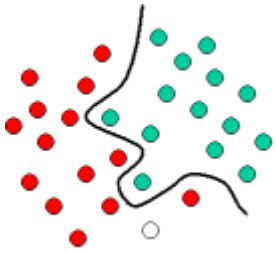
4. Support Vector Machines (SVM)

This method performs regression and classification tasks by constructing nonlinear decision boundaries. Because of the nature of the feature space in which these boundaries are found, Support Vector Machines can exhibit a large degree of flexibility in handling classification and regression tasks of varied complexities. There are several types of Support Vector models including linear, polynomial, RBF, and sigmoid.

Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane is one that separates between a set of objects having different class memberships. A schematic example is shown in the illustration below. In this example, the objects belong either to class GREEN or RED. The separating line defines a boundary on the right side of which all objects are GREEN and to the left of which all objects are RED. Any new object (white circle) falling to the right is labeled, i.e., classified, as GREEN (or classified as RED should it fall to the left of the separating line).



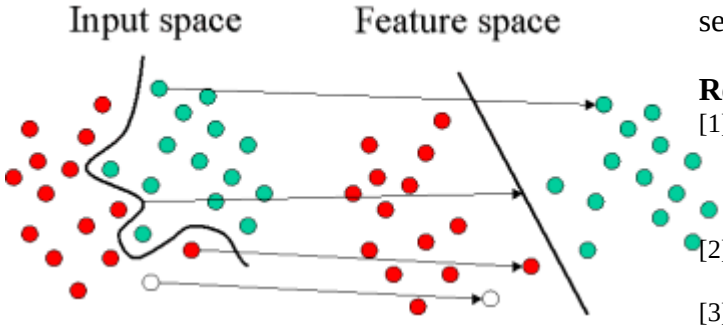
The above is a classic example of a linear classifier, i.e., a classifier that separates a set of objects into their respective groups (GREEN and RED in this case) with a line. Most classification tasks, however, are not that simple, and often more complex structures are needed in order to make an optimal separation, i.e., correctly classify new objects (test cases) on the basis of the examples that are available (train cases). This situation is depicted in the illustration below. Compared to the previous schematic, it is clear that a full separation of the GREEN and RED objects would require a curve (which is more complex than a line). Classification tasks based on drawing separating lines to distinguish between objects of different class memberships are known as hyperplane classifiers. Support Vector Machines are particularly suited to handle such tasks.



The illustration below shows the basic idea behind Support Vector Machines. Here we see the original objects (left side of the schematic) mapped, i.e., rearranged, using a set of

ISSN 2348 – 7968

mathematical functions, known as kernels. The process of rearranging the objects is known as mapping (transformation). Note that in this new setting, the mapped objects (right side of the schematic) is linearly separable and, thus, instead of constructing the complex curve (left schematic), all we have to do is to find an optimal line that can separate the GREEN and the RED objects.



5. *k*-Nearest Neighbors

k-Nearest Neighbors is a memory-based method that, in contrast to other statistical methods, requires no training (i.e., no model to fit). It falls into the category of Prototype Methods. It functions on the intuitive idea that close objects are more likely to be in the same category. Thus, in KNN, predictions are based on a set of prototype examples that are used to predict new (i.e., unseen) data based on the majority vote (for classification tasks) and averaging (for regression) over a set of *k*-nearest prototypes (hence the name *k*-nearest neighbors).

6. Conclusions

This paper provides a review of machine learning approaches and documents representation techniques. This paper describes the best-known supervised techniques in relative detail. Machine learning offers a cornucopia of useful ways to approach problems that otherwise defy manual solution. This article summarized some of the most salient items.

Our study also encourages that no one technique can be classified as being the perfect machine learning technique. In this paper, a study oriented work is defined regarding the sentiment analysis. This analysis process is defined in terms of different approaches for sentiment analysis. The work also includes the exploration of the sentiment analysis.

References

[1] “S. B. Kotsiantis Department of Computer Science and Technology University of Peloponnese” Supervised Machine Learning: A Review of Classification Techniques 2007.

[2] Kiri L. Wagstaff, “Machine Learning that Matters,2012.

[3] Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan Department of Computer and Information Science, Universiti Teknologi PETRONAS, Tronoh, Malaysia. “A Review of Machine Learning Algorithms for Text-Documents Classification”. 2010.

[4] Pedro Domingos Department of Computer Science and Engineering University of Washington Seattle, “A Few Useful Things to Know about Machine Learning”. 2012.

[5] Yogesh Singh, Pradeep Kumar Bhatia & Omprakash Sangwan “A REVIEW OF STUDIES ON MACHINE LEARNING TECHNIQUES”.

[6] Tom M. Mitchell “The Discipline of Machine Learning”. 2006.

[7] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan “Thumbs up? Sentiment Classification using Machine Learning Techniques”. 2012.