BWT- Data Science Task 11

Bisma Fajar

July 2024

Introduction to Machine Learning

1 What is ML?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.

2 Why do we need ML?

1. Problems for which existing solutions require a lot of hand-tuning or long lists of rules: one Machine Learning algorithm can often simplify code and perform bet- ter.
2. Complex problems for which there is no good solution at all using a traditional approach: the best Machine Learning techniques can find a solution.
3. Fluctuating environments: a Machine Learning system can adapt to new data.
4. Getting insights about complex problems and large amounts of data

3 Types of ML

Machine Learning systems can be classified according to the amount and type of supervision they get during training. There are four major categories: supervised learning, unsupervised learning, semisupervised learning, and Reinforcement Learn- ing.

3.1 Supervised Learning

In supervised learning, the training data you feed to the algorithm includes the desired solutions, called labels. A typical supervised learning task is classification. The spam filter is a good example of this: it is trained with many example emails along with their class (spam or ham), and it must learn how to classify

new emails.
Here are some of the most important supervised learning algorithms
• k-Nearest Neighbors
• Linear Regression
• Logistic Regression
• Support Vector Machines (SVMs)
• Decision Trees and Random Forests
• Neural networks

3.2 Unsupervised Learning

In unsupervised learning, as you might guess, the training data is unlabeled. The system tries to learn without a teacher.For example, say you have a lot of data about your blog's visitors. You may want to run a clustering algorithm to try to detect groups of similar visitors. At no point do you tell the algorithm which group a visitor belongs to: it finds those connections without your help. For example, it might notice that 40% of your visitors are males who love comic books and generally read your blog in the evening, while 20% are young sci-fi lovers who visit during the weekends, and so on. If you use a hierarchical clustering algorithm, it may also subdivide each group into smaller groups. This may help you target your posts for each group.
Here are some of the most important unsupervised learning algorithms:
• Clustering
— K-Means
— DBSCAN
— Hierarchical Cluster Analysis (HCA)
• Anomaly detection and novelty detection
— One-class SVM
— Isolation Forest
• Visualization and dimensionality reduction
— Principal Component Analysis (PCA)
— Kernel PCA
— Locally-Linear Embedding (LLE)
— t-distributed Stochastic Neighbor Embedding (t-SNE)
• Association rule learning
— Apriori
— Eclat

3.3 Semisupervised learning

Some algorithms can deal with partially labeled training data, usually a lot of unla- beled data and a little bit of labeled data. This is called semisupervised learning Most semisupervised learning algorithms are combinations of unsupervised and supervised algorithms. For example, deep belief networks (DBNs) are based on unsu- pervised components called restricted Boltzmann machines (RBMs) stacked on top of one another. RBMs are trained sequentially in an unsupervised manner, and then the whole system is fine-tuned using supervised learning techniques.

3.4 Reinforcement Learning

The learning system, called an agent in this context, can observe the environment, select and perform actions, and get rewards in return. It must then learn by itself what is the best strategy, called a policy, to get the most reward over time. A policy defines what action the agent should choose when it is in a given situation. For example, many robots implement Reinforcement Learning algorithms to learn how to walk. DeepMind's AlphaGo program is also a good example of Reinforcement Learning: it made the headlines in May 2017 when it beat the world champion Ke Jie at the game of Go. It learned its winning policy by analyzing millions of games, and then playing many games against itself. Note that learning was turned off during the games against the champion; AlphaGo was just applying the policy it had learned.

4 Batch and Online Learning

Another criterion used to classify Machine Learning systems is whether or not the system can learn incrementally from a stream of incoming data.

4.1 Batch learning

In batch learning, the system is incapable of learning incrementally: it must be trained using all the available data. This will generally take a lot of time and computing resources, so it is typically done offline. First the system is trained, and then it is launched into production and runs without learning anymore; it just applies what it has learned. This is called offline learning. If you want a batch learning system to know about new data (such as a new type of spam), you need to train a new version of the system from scratch on the full dataset (not just the new data, but also the old data), then stop the old system and replace it with the new one. Fortunately, the whole process of training, evaluating, and launching a Machine Learning system can be automated fairly easily, so even a batch learning system can adapt to change. Simply update the data and train a new version of the system from scratch as often as needed.

4.1.1 Disadvantages

1. Training on the full set of data requires a lot of computing resources (CPU, memory space, disk space, disk I/O, network I/O, etc.).

2. If you have a lot of data and you automate your system to train from scratch every day, it will end up costing you a lot of money.

3. If the amount of data is huge, it may even be impossible to use a batch learning algorithm

4.2 Online learning

In online learning, you train the system incrementally by feeding it data instances sequentially, either individually or by small groups called mini-batches. Each learning step is fast and cheap, so the system can learn about new data on the fly, as it arrives.
One important parameter of online learning systems is how fast they should adapt to changing data: this is called the **learning rate**. If you set a high learning rate, then your system will rapidly adapt to new data, but it will also tend to quickly forget the old data and vise versa.

4.2.1 Challenges

A big challenge with online learning is that if bad data is fed to the system, the sys- tem's performance will gradually decline. If we are talking about a live system, your clients will notice. For example, bad data could come from a malfunctioning sensor on a robot, or from someone spamming a search engine to try to rank high in search results.

4.2.2 Possible Solution

To reduce this risk, you need to monitor your system closely and promptly switch learning off (and possibly revert to a previously working state) if you detect a drop in performance. You may also want to monitor the input data and react to abnormal data (e.g., using an anomaly detection algorithm).

5 Instance Based vs Model Based Learning

One more way to categorize Machine Learning systems is by how they generalize. Most Machine Learning tasks are about making predictions. This means that given a number of training examples, the system needs to be able to generalize to examples it has never seen before. Having a good performance measure on the training data is good, but insufficient; the true goal is to perform well on new instances.

5.1 Instance-based learning

Possibly the most trivial form of learning is simply to learn by heart. If you were to create a spam filter this way, it would just flag all emails that are identical to emails that have already been flagged by users—not the worst solution, but certainly not the best. Instead of just flagging emails that are identical to known spam emails, your spam filter could be programmed to also flag emails that are very similar to known spam emails. This requires a measure of similarity between two emails. A (very basic) simi- larity measure between two emails could be to count the number of words they have in common. The system would flag an email as spam if it has many words in com- mon with a known spam email. This is called instance-based learning: the system learns the examples by heart, then generalizes to new cases by comparing them to the learned examples (or a subset of them), using a similarity measure. For example, the new instance would be classified as a triangle because the majority of the most similar instances belong to that class.

5.2 Model-based learning

Another way to generalize from a set of examples is to build a model of these exam- ples, then use that model to make predictions. This is called model-based learning. For example, suppose you want to know if money makes people happy, so you down- load the Better Life Index data from the OECD's website as well as stats about GDP per capita from the IMF's website. Then you join the tables and sort by GDP per cap- ita.

6 Main Challenges of Machine Learning

In short, since your main task is to select a learning algorithm and train it on some data, the two things that can go wrong are "bad algorithm" and "bad data." Let's start with examples of bad data.

1. Insufficient Quantity of Training Data
2. Nonrepresentative Training Data
3. Poor-Quality Data
4. Irrelevant Features
5. Overfitting the Training Data
6. Underfitting the Training Data