A Crash Course in Machine Learning Fundamentals

Valerie Ding
Stanford University
dingv@stanford.edu

December 22, 2017

Abstract

This is a high-level overview of fundamental terminology, concepts, and techniques in machine learning. I will start by describing each paradigm in layman's terms, then outline the mathematical intuition and practical applications.

Machine learning

Think back to when you were a child, first learning the alphabet. Perhaps you had stickers, magnets, or posters in your classroom. The letters might have been in different fonts, colors, or sizes, but over time, you were trained to recognize the similarities and classify the letters as A, B, C, etc. even if, for example, a black serif **A** looked wholly different from a red A.

Indeed, this was a seminal result in early machine learning: literally teaching a computer how to recognize characters in images (optical character recognition), emulating the process of teaching a child how to read. Now you can take a picture of a document with your phone and convert it to an electronic typed document, because we have figured out how to get machines to recognize characters from images.

Optical character recognition is just one, early example of machine learning. The takeaway is that we are developing learning models for machines by which the machines can use input information to learn about data, helping us make decisions, discoveries, and processes more efficient in a wide range of fields such as information security, medicine, and automated machinery.

Supervised learning

Supervised learning is one of the most common paradigms in machine learning. We give the computer a set of **training data**, which is a set of features and labels. For example, let's say we are trying to teach the computer to flag suspicious credit card activity to detect credit card fraud. Some features used in our learning model might be transaction amount, location, and vendor type. Then, each training example (e.g. 20 USD spent on groceries at 3PM at the Safeway in Menlo Park) is accompanied by a label (e.g. yes, this was an actual transaction by the credit card owner) in the training data. Now we can train our computer on these training examples: we can train our model to weight the features by importance by minimizing a loss function (also known as a cost or objective function, this means we are trying to optimally quantify a general model for our training data by bringing results as close as possible to the actual labels). Then, when we apply the model to **make predictions on unlabeled data**, we are putting our model out into the field and seeing how it performs on data where we don't know the answers (e.g. is this new transaction of 100 USD spent at 2AM on candles at a Walmart in Kentucky legitimate?) using the model that we tuned over the training set and validated with existing data. This is the foundation of supervised learning.

A central concept in supervised learning is whether the problem is one of **classification or regression**. Classification problems have discrete outputs (e.g. classifying characters in OCR, or boolean yes/no credit card authentication), while regression problems have continuous outputs (e.g. stock price predictions). Common fundamental paradigms in modeling regression problems are logistic regression and gradient descent; in classification, Naive Bayes (features assumed independent from each other) and Support Vector Machines (maximizes the margin attempting to separate classes); the important takeaway in understanding supervised learning approaches is that predictions are driven by maximizing the likelihood and minimizing the objective function, making predictions based on probabilistic models that are trained on existing data and known answers.

Unsupervised learning

In unsupervised learning, we are trying to find structure in data. Rather than making classification or regression predictions, we are trying to find patterns in the data itself. For example, the k-means **clustering** algorithm maintains a set of k cluster centroids that it iteratively updates until convergence, by assigning each data point to its closest centroid and then updating each centroid to be the mean of the data points assigned to it. Some applications of clustering are object recognition and in bioinformatics such as cancer detection.

Deep learning

Deep learning techniques are foundationally neural networks with multiple hidden **layers**. This means that features are processed and optimized over multiple internal layers, allowing us to parse data with greater complexity and find structure at different levels of abstraction. This is especially helpful in dealing with large-scale unlabeled datasets. A fundamental paradigm in deep learning is backpropagation. To understand this conceptually, first understand that forward propagation is the process of transforming input to final output; the input is passed through each layer of the deep neural network to land on an output; then, to minimize the error, we use backpropagation to determine each layer's contribution to the overall error. Now we have landed on an optimization problem to optimize weights to minimize loss.