What are the limitations of derivative-based models for optimization in machine learning?

Faris Chaudhry

August 2, 2021

Abstract

Most machine learning problems can be transposed into optimization problems with the goal being finding the global minima or maxima to minimize loss or maximize potential. Each of the main learning methodologies (namely supervised, unsupervised and reinforcement) along with the models to represent and solve optimization problems have limitations - computationally and theoretically - that have to be identified and mitigated against to create an effective model. The focus here are objective functions that are continuously differentiable and thus derivative-based solutions are used.

Contents

Chapter 1

Introduction to Machine Learning and Optimization

1.1 What is Machine Learning?

Machine learning (ML) is a subfield of artificial intelligence (AI) which, broadly speaking, is the use of computational methods and models to improve performance and predictions through experience [2, p. 1]. Unlike humans, this learning is based entirely on data and statistics and experience is gained through interaction with a training set of data or an environment of some kind.

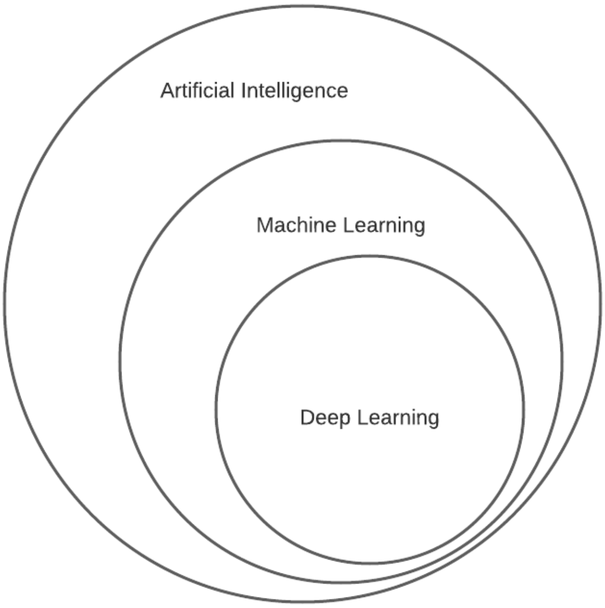

Figure 1.1: subfields of AI

There are 3 primary categories (supervised, unsupervised and reinforcement) of learning philosophies for ML models, with other hybrid models being combinations of these. Each type of learning lends to itself to certain types of problems due to the limitations that each one has. Supervised learning is used for classifying images and extrapolating data. Unsupervised learning takes raw data and finds patterns such as the overall distribution or groups with similar attributes. Reinforcement is used in complex systems which many changing variables that would be computationally difficult to solve otherwise, like chess.

1.2 Prerequisite Conditions for Derivative-Based Optimization

Optimization revolves around minimizing the loss or maximizing the value of a function. In the context of ML, the process of optimizing is vital to ensure modelling produces the greatest accuracy. The goal is to optimize an objective function, which is the representation of the variables being simulated. The solution to the objective function will be either a minimum (minima) or maximum (maxima) point (collectively called the set of extrema) as this is when the value of a function is highest or lowest.

The derivative is a linear approximation (tangent) to a function at a point. Suppose there was a function $f(x)$ then, intuitively, the derivative with respect to x would be the how much the value of $f(x)$ changed with a small nudge in the x direction. It is important to note that, by Fermat's theorem on stationary points [7], all critical points (extrema and saddle points) have first derivative equal to 0. Visually, this is because the tangent to any turning point will have a gradient of 0. See the function $y = x^2$ (fig. 1.2) which has a minima at $(0, 0)$.

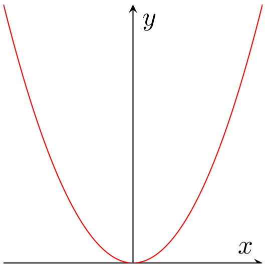

Figure 1.2: graph of x^2

So the objective becomes to find all critical point locations and the nature of these points - what kind of critical point it is.

1.2.1 Continuity and Differentiability

The most essential requirement to using derivative-based methods will be that the objective function must be continuous and twice-differentiable (the derivative of the function must also be differentiable) over the interval that contains the solution.

This is because to find the location and nature of critical points, the first and second derivative of a function are required [6].

For a function $f(x)$ to be continuous over the interval $I = [a, b]$

$$\forall k \in I, \lim_{x \to k} f(x) = f(k) \tag{1}$$

This means that, given any number in the interval, as x approaches that number it would be equal to putting the number into the function. This prevents any discontinuity since the limit wouldn't exist at discontinuous points. In fig. 1.3, $\lim_{x \to 0^+} = +\infty$ and $\lim_{x \to 0^-} = -\infty$. These values contradict meaning the limit isn't defined.

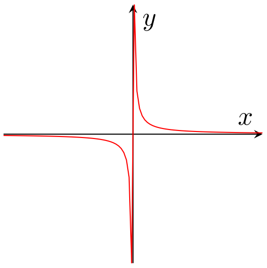

Figure 1.3: graph of $\frac{1}{x}$

For a single-valued function, $f(x)$, the derivative, $f'(x)$ exists iff the following limits exists.

$$\frac{df}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{2}$$

However, most objective functions will be multi-valued to account for all the variables so this definition must be extended. This is the same principle but a nudge in a specific direction. Suppose there is a function $f(x_1, \cdots, x_i)$ then the derivative with respect to a certain variable, x_n, will be.

$$\frac{\partial f}{\partial x} = \lim_{x \to 0} \frac{f(x_1, \cdots, x_n + \Delta x, \cdots, x_i) - f(x_1, \cdots, x_n, \cdots, x_i)}{\Delta x} \tag{3}$$

In practice these rigorous definitions are not used but the concept of continuous differentiability is important.

- The first and second derivative must exist for an objective function to be solvable in this method, which is the major limiting factor. Although derivative-free methods do exist, they tend to be approximations of the exact values and heuristic in theory.
- Although many functions discontinuities, like asymptotes or singularities, many times these are removable either by defining an interval without them or assigning an arbitrary value at a point for continuity.

1.2.2 Concavity and Convexity

When a function has only 1 minima or maxima over an interval it becomes much easier to find the global minimum or maximum due the lack of a need to check which point is a local extremum and which is the global extremum. Functions like these are called convex and concave where convex functions have a minimum point and concave functions have a maximum point. A convex function [5] can visually be described as having all its points below a line segment drawn between (fig. 1.4) any 2 points while a concave function has all points above.

It is important to note that concavity and convexity are not opposites. A function can be concave, non-concave, convex or non-convex. In addition, reflecting a function in the x-axis will reverse its concavity or convexity. Suppose $f(x)$ is convex then $-f(x)$ is concave.

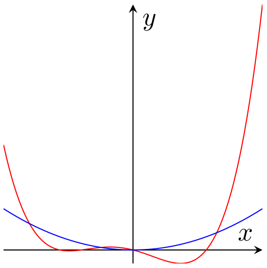

Figure 1.4: convex (blue) and non-convex (red)

If the objective function is a non-convex or non-concave function this doesn't prevent the use of derivative-based optimization. However it does restrict the range of methods that can be used to find the global solution. For example, iterative methods to find extrema might not always work since they could get closer to the local extrema while neglecting other possible values. Moreover, it will increase the complexity of the problem computationally since there will be range of possible global extrema that have to be checked - which can be particularly difficult when certain derivative tests are inconclusive.

Chapter 2

Supervised Learning

2.1 Application of Supervised Learning

The philosophy of supervised learning is to use labelled training data to map between an input vector and a target vector. In this case, the model is given data with input variables and the correct associated target values corresponding with them [1, p. 105]. Effectively, the model is creating a pattern out of which inputs cause certain outputs so that, given new inputs, the correct outputs can be predicted.

Supervised learning problems are split into 2 distinct categories: classification problems and regression problems.

2.1.1 Classification

Classification problems are about predicting the class labels of an object. A common example of classification is assigning an digit label to a handwritten digit. However, these objects could be anything that can be labelled such as sentences or sounds.

Let x_n be a feature/parameter of the object and l_n be a label where $n \in \mathbb{Z}$
Then a general classification function can be described as the mapping:

$$[x_1, x_2, \cdots] \mapsto [l_1, l_2, \cdots]$$

Given a particular feature vector (a vector of the parameters of the classification function) the goal is to assign a set of class labels.

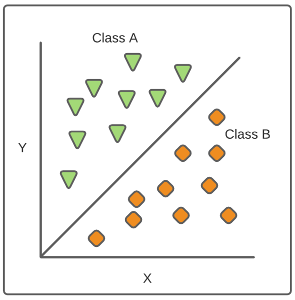

Figure 2.1: example of classification

2.1.2 Regression

Regression problems involve predicting a numerical value from the feature vector of an object. For example, given many variables about a stock (past history), predict the future value of the stock.

Let x_n be a feature/parameter of the object and k be the numerical value associated with it where $n \in \mathbb{Z}$ Then a general regression function can be described as the mapping:

$$[x_1, x_2, \cdots] \mapsto k$$

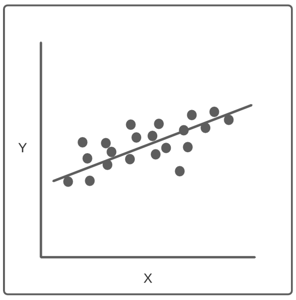

Figure 2.2: example of regression

2.2 General Optimization of Supervised Learning Problems

The optimization of a supervised learning problem is to minimize the average of the loss function using the training samples. This produces the most accurate approximation to the underlying function to extrapolate values.

The general equation [4, p. 3] for this can be written as:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i, \theta)) \tag{4}$$

where N is the number of training samples, θ is the parameter of the mapping function, x^i is a feature vector and y^i is the array of labels associated with that feature vector.

The problem with using training samples is that the resulting function might be over fitted to the given data. This would mean that, although the model is accurate for the training data it has been given, accuracy is reduced on new objects. The method to deal with this is through a regularization item, λ:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} L(y^i, f(x^i, \theta)) + \lambda ||\theta||_2^2 \tag{5}$$

Regularization fundamentally discourages learning complex models and will be covered more in depth when talking about limitations.

2.3 Limitations of Supervised Learning

2.3.1 Data Preprocessing and Quality

2.3.2 Over Fitting and Regularization

Models don't generalize well from observed, training data to unseen data [8]

2.3.3 Computational Resources

Chapter 3

Unsupervised Learning

3.1 Application of Unsupervised Learning

Unsupervised learning is different to supervised learning in the way that it uses unlabelled data [1, p. 105]; instead of learning from a mapping of inputs to a know output, the model is given only the inputs to learn from and has to make sense of the data without guidance. As a result of this, unsupervised learning revolves around extracting relationships from the data without the inherent human biases caused by choosing the correct output beforehand.

Unsupervised learning problems strive to solve 1 of 2 problems: finding clusters of similar data and summarizing the distribution of the data (density estimation).

3.1.1 Clustering

Unlike classification, where the classes are predefined, clustering requires the model to define its own cluster of data based on the similarites of the features.

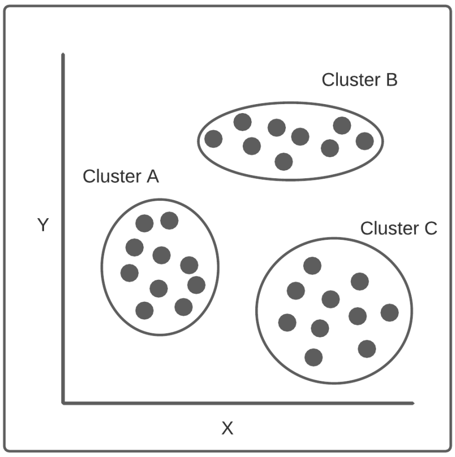

Figure 3.1: example of clustering

In general, optimization will involve making the variance of each cluster as small as possible, which will be equivilant to the distnace from the center of the cluster.

$$\min_{s} \sum_{k=1}^{K} \sum_{x \in S_k} \|x - \mu_k\|_2^2 \tag{6}$$

3.1.2 Density Estimation

The assumption is that there exists some probability istribution to describe the relationship between the variables [3]. Density estimation is a useful asset in modelling to estimate the properties of a given data set (variance, skewness, type of distribution).

Suppose there exists a set of continous random variables, $(x_1, \cdots, x_n)$, then there is a probability distribution that the set models, $P(x_1, \cdots, x_n)$. The goal is to find a continous probability density function (PDF) that can describe the mapping: $\{x_1, \cdots, x_n\} \rightarrow P(x_1, \cdots, x_n)$. The assumption that this will PDF will be continous is valid] since our objective function has the precondition of being continous.

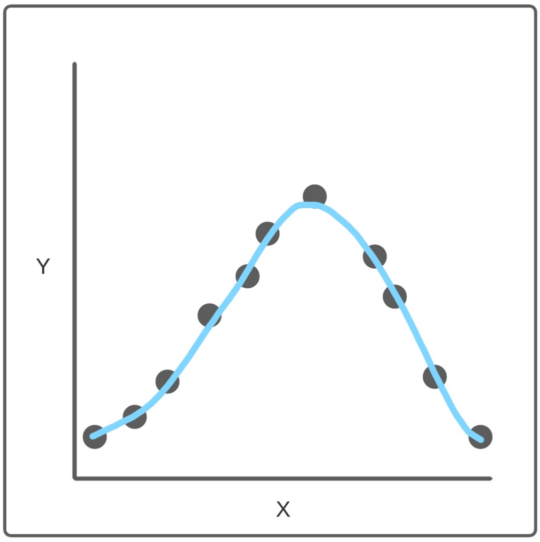

Figure 3.2: example of density estimation

Chapter 4

Reinforcement Learning

Chapter 5

Mathematical Models for Optimization

5.1 Creating an Objective Function

5.1.1 Constraints and the Lagrangian Multiplier

Chapter 6

Conclusion

References

[1] Yoshua Bengio, Ian Goodfellow, and Aaron Courville. *Deep learning.* Vol. 1. MIT press Massachusetts, USA: 2017.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning.* MIT press, 2018.

[3] Simon J Sheather. "Density estimation". In: *Statistical science* (2004), pp. 588–597.

[4] Shiliang Sun et al. "A survey of optimization methods from a machine learning perspective". In: *IEEE transactions on cybernetics* 50.8 (2019), pp. 3668–3681.

[5] Lieven Vandenberghe and Stephen Boyd. *Convex optimization.* Vol. 1. Cambridge University Press Cambridge, 2004.

[6] Eric W. Weisstein. *Second Derivative Test. From MathWorld–A Wolfram Web Resource.* URL: `https://mathworld.wolfram.com/SecondDerivativeTest.html`.

[7] Eric W. Weisstein. *Stationary Point. From MathWorld–A Wolfram Web Resource.* URL: `https://mathworld.wolfram.com/StationaryPoint.html`.

[8] Xue Ying. "An overview of overfitting and its solutions". In: *Journal of Physics: Conference Series.* Vol. 1168. 2. IOP Publishing. 2019, p. 022022.