Deep Learning

AlejandroMllo

This document serves as a very brief summary of the topics covered in each chapter of the book Deep Learning [1].

Disclaimer: This document is completely extracted from [1], the author does not attribute any ownership over the material.

4 Numerical Computation

- Algorithms that solve mathematical problems by methods that update estimates of the solution via an iterative process.
- Underflow: numbers near zero that are rounded to zero.
- Overflow: numbers with large magnitude are approximated as ∞ or -∞.
- Conditioning: how rapidly a function changes w.r.t. small changes in its inputs.
- Optimization: minimize or maximize an objective function $f(\boldsymbol{x})$ by altering $\boldsymbol{x}$. Sometimes, the value that minimizes or maximizes a function, is denoted with a superscript $*$: $\boldsymbol{x}^* = arg\,min\ f(\boldsymbol{x})$.
- Gradient Descent: reduce $f(x)$ by moving x in small steps with the opposite sign of the derivative. $f(x - \epsilon \texttt{sign}(f'(x))$.
- Critical point: point with zero slope.
- In the context of Deep Learning it is tried to find a value of f that is very low but not necessarily minimal in any formal sense.
- The partial derivative $\dfrac{\partial}{\partial x_i} f(\boldsymbol{x})$ measures how f changes as only the variable x_i increases at point $\boldsymbol{x}$. The gradient generalizes the notion of derivative to the case where the derivative is w.r.t. a vector: the gradient of f is the vector containing all the partial derivatives, denoted $\nabla_x f(\boldsymbol{x})$.
- The directional derivative in direction$\boldsymbol{u}$ (a unit vector) is the slope of the function f in direction u. The minimization occurs when the gradient points directly uphill, and the negative gradient points directly downhill.
- It is possible to decrease f by moving in the direction of the negative gradient. This is the method of steepest descent, or gradient descent: $\boldsymbol{x}' = \boldsymbol{x} - \epsilon \nabla_x f(\boldsymbol{x})$, where ϵ is the learning rate, a positive scalar determining the size of the step. This method converges when every element of the gradient is zero (or very close to zero). To jump directly to the critical point, solve: $\nabla_x f(\boldsymbol{x}) = 0$.
  Hill climbing is the generalization of this method for discrete spaces.
- If there is the function $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^n$, then the Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{n \times m}$ of $\boldsymbol{f}$ is defined such that $J_{i,j} = \frac{\partial}{\partial x_j} f(\boldsymbol{x})_i$.
- The Hessian matrix $\boldsymbol{H}(f)(\boldsymbol{x})$ is defined such that:

$$\boldsymbol{H}(f)(\boldsymbol{x})_{i,j} = \frac{\partial^2}{\partial x_i x_j} f(\boldsymbol{x}).$$

  Equivalently, the Hessian is the Jacobian of the gradient. Also, it is symmetric (if the function is continuous at such points). The second derivative in a specific direction represented by a unit vector $\boldsymbol{d}$ is given by $\boldsymbol{d}^\top \boldsymbol{H} \boldsymbol{d}$; when $\boldsymbol{d}$ is an eigenvector of $\boldsymbol{H}$, the second derivative in that direction is the corresponding eigenvalue.
  To the extent that the function to be minimized can be approximated well by a quadratic function, the eigenvalues of the Hessian thus determine the scale of the learning rate.
  Using the eigendecomposition of the Hessian matrix, it is possible to generalize the second derivative test to multiple dimensions.
- Newton's Method: uses a second-order Taylor series expansion to approximate $f(\boldsymbol{x})$ near some point $\boldsymbol{x}^{(0)}$.

- In Deep Learning, sometimes the functions used are restricted to those that are either Lipschitz continuous or have Lipschitz continuous derivatives. A Lipschitz continuous function is a function f whose rate of change is bounded by a Lipschitz constant $\mathcal{L} : \forall \boldsymbol{x}, \forall \boldsymbol{y}, |f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \mathcal{L}||\boldsymbol{x} - \boldsymbol{y}||_2$.
- Constrained Optimization: used to find the maximal or minimal value of $f(\boldsymbol{x})$ for values of $\boldsymbol{x}$ in some set $\mathbb{S}$.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.