Title: Manifold hypothesis

URL: https://en.wikipedia.org/wiki/Manifold_hypothesis

PageID: 68581881

Categories: Category:Machine learning, Category:Theoretical computer science

-----

The manifold hypothesis posits that many high-dimensional data sets that occur in the real world actually lie along low-dimensional latent manifolds inside that high-dimensional space. [1] [2] [3] [4] As a consequence of the manifold hypothesis, many data sets that appear to initially require many variables to describe, can actually be described by a comparatively small number of variables, linked to the local coordinate system of the underlying manifold. It is suggested that this principle underpins the effectiveness of machine learning algorithms in describing high-dimensional data sets by considering a few common features.

The manifold hypothesis is related to the effectiveness of nonlinear dimensionality reduction techniques in machine learning. Many techniques of dimensional reduction make the assumption that data lies along a low-dimensional submanifold, such as manifold sculpting , manifold alignment , and manifold regularization .

The major implications of this hypothesis is that

Machine learning models only have to fit relatively simple, low-dimensional, highly structured subspaces within their potential input space (latent manifolds).

Within one of these manifolds, it’s always possible to interpolate between two inputs, that is to say, morph one into another via a continuous path along which all points fall on the manifold.

The ability to interpolate between samples is the key to generalization in deep learning . [5]

The information geometry of statistical manifolds

An empirically-motivated approach to the manifold hypothesis focuses on its correspondence with an effective theory for manifold learning under the assumption that robust machine learning requires encoding the dataset of interest using methods for data compression. This perspective gradually emerged using the tools of information geometry thanks to the coordinated effort of scientists working on the efficient coding hypothesis , predictive coding and variational Bayesian methods .

The argument for reasoning about the information geometry on the latent space of distributions rests upon the existence and uniqueness of the Fisher information metric . [6] In this general setting, we are trying to find a stochastic embedding of a statistical manifold. From the perspective of dynamical systems, in the big data regime this manifold generally exhibits certain properties such as homeostasis:

We can sample large amounts of data from the underlying generative process.

Machine Learning experiments are reproducible, so the statistics of the generating process exhibit stationarity.

In a sense made precise by theoretical neuroscientists working on the free energy principle , the statistical manifold in question possesses a Markov blanket . [7]

See also

Kolmogorov complexity

Minimum description length

Solomonoff's theory of inductive inference

References

Further reading

Brown, Bradley C. A.; Caterini, Anthony L.; Ross, Brendan Leigh; Cresswell, Jesse C.; Loaiza-Ganem, Gabriel (2023). The Union of Manifolds Hypothesis and its Implications for Deep Generative Modelling . The Eleventh International Conference on Learning Representations. arXiv : 2207.02862 .

Lee, Yonghyeon (2023). A Geometric Perspective on Autoencoders . arXiv : 2309.08247 .