Using a Deep Autoencoder Model to Characterize Air Pollution

Nicolas Shannon, Bishnu Timalsena, Raul Chavez, Vandana Nunna Lakshmi

July 13, 2021

Abstract

Pollution Autoencoders uses a deep autoencoder neural network to characterize cities in the United States based on time-series readings of various polluting gases and particulates. We compared the Autoencoders method with Principle Component Analysis (PCA), an older technique for reducing the dimensionality of a data set. The main advantage of Autoencoders have over PCA is in their capability of discovering non-linear relationships between polluting gases. Autoencoders are able to capture more of the explained variance because of this increased flexibility. We measured the explained variance across 190 dimensions for each of the 8 polluting gases by performing a simple linear regression on the compressed data generated by both the PCA and Autoencoder methods. As a control, we also performed linear regression on the uncompressed data.

1 Introduction

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

2 Methodology

The main use of an undercomplete autoencoder is not in its ability to perfectly reconstruct the input, but in the useful features it is capable of distilling.

In our case, we wanted to see what properties the autoencoder model could extract when given time series air pollution data for various cities throughout the United States. Using data from OpenWeather's Air Pollution API, we constructed data frames for eight polluting gases and particulates. We obtained the daily averages of each pollutant six and a half months for just over eighteen-thousand cities in the U.S. and other U.S. territories.

After obtaining the requisite data, we cleaned and preprocessed the data by removing cities that had the same name and any entries with missing values. While this made the data quickly useable, a more ideal solution would be to add a state/country code to distinguish between cities and to impute any missing values. For more on this see the future improvements section.

We compared the autoencoder model to Principle Component Analysis, a much older technique used for dimensionality reduction that has been a part of statistical literature since the early twentieth century.[2] The main distinction between autoencoders and PCA is that autoencoders can span a nonlinear subspace, whereas PCA learns the principle subspace.

Autoencoders that employ nonlinar encoder and decoder functions can create a more powerful generalization than PCA. However, an autoencoder model that is allowed too much capacity is at risk of reconstructing the input "too perfectly" without extracting any useful information.[3]

As Goodfellow, Bengio, and Courville (2016) say, there's no reason... Jolliffe (2002) Agency (2009)

References

Agency, U.S. Environmental Protection (2009). "Carbon Monoxide NAAQS: Scope and Methods Plan for Health Risk and Exposure Assessment". In: `https://www.epa.gov/sites/production/files/2020-07/documents/2009_04_coscopeandmethodsplan.pdf`.

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

Jolliffe, I.T. (2002). *Principle Component Analysis – 2nd Ed.* Library of Congress.