Principal Component Analysis Based Method to Predict Temperatures

Kaitlyn Roberts, Western Connecticut State University
Advisor: Professor Xiaodi Wang, Western Connecticut State University

Abstract

Weather forecasting is an important tool in many industries around the world. Numerical models are useful in creating an accurate forecast, but the large amount of data involved in these calculations increases computational complexity dramatically. In this research, we apply principal component analysis (PCA) to reduce the dimensions of a data set used to train a neural network to predict temperatures. The data set includes temperature, dew point, air pressure, precipitation amounts, and wind speed for the past 72 hours in 3-hour increments. The goal is to predict temperatures. We then compare the performance of the model trained with reduced-dimension data versus that of the model trained with a full data set.

The Problem

Weather forecasting is a complicated, important process. Technology and machine learning have improved weather modelling and prediction accuracy, but require large amounts of both time and computing power.

Availability of Weather Models

Several large weather models, such as the North American Mesoscale Forecast System (NAM) and the Global Forecast System (GFS) are made available for free online, while others, like the European Centre For Medium-range Weather Forecasts (ECMWF) have limited data available for free, with more detailed information available via subscription.

Resources required to run in-depth prediction models are not readily available for most smaller operations.

For local forecasting, meteorologists must access a large model online and interpret that data as best as possible for their local region, adjusting the predicted values for many variables.

Research Question

Can dimensional deduction analysis, specifically PCA, produce a corresponding new dataset in a lower dimensional space that is small enough to be usable in modeling for everyday users, while retaining enough accuracy to provide usable predictions?

Experiment

- Collect historical weather data from National Centers for Environmental Information (NCEI) and format into matrix for importing into MATLAB
- Use MATLAB's built-in Classification Learner app to train and test two neural networks
  - Network 1: trained with complete dataset
  - Network 2: trained with PCA-based dataset

Dataset

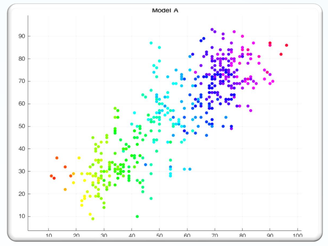
- Historical hourly weather observations from Danbury Municipal Airport (KDXR) from Jan. 1, 2021 to Nov. 9, 2021
- Matrix of previous 72 hours' observations in 3-hour intervals (120 observation columns total)
  - Temperature (˚ F)
  - Dew Point (˚ F)
  - Precipitation (in.)
  - Station Pressure (inHg)
  - Wind Speed (mph)

Experimental Procedure

- Original data processed in MATLAB
  - Standardized
  - Transformed via Transpose times Original
  - Eigenvectors and eigenvalues computed
- 18 principal components selected
  - Represented 81.17% of variance
- Original data projected onto eigenspace spanned by 18 principal components
- Tested neural networks using MATLAB Classification Learner app
- Trained with data from Jan. 4 to Sept. 30 (6479 timesteps)
- Tested with data from Oct. 1 to Nov. 9 (960 timesteps)

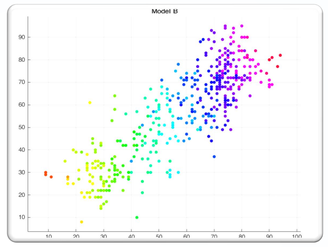
Neural Network Results

Network A



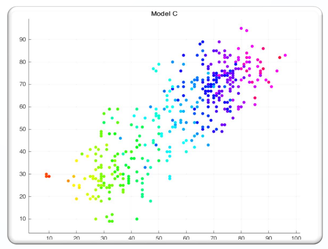
- Single-layer neural network
- 10 neurons in hidden layer
- Validation accuracy: 13.6%
- Testing accuracy: 10.0%

Network B



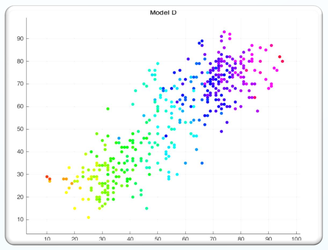
- Single-layer neural network
- 5 neurons in hidden layer
- Validation accuracy: 13.6%
- Testing accuracy: 11.0%

Network C



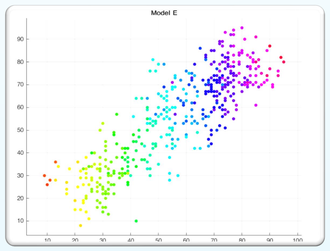
- Single-layer neural network
- 25 neurons in hidden layer
- Validation accuracy: 12.4%
- Testing accuracy: 7.4%

Network D



- Dual-layer neural network
- 10 neurons each in 2 hidden layers
- Validation accuracy: 13.0%
- Testing accuracy: 10.1%

Network E



- Dual-layer neural network
- 5 neurons each in 2 hidden layers
- Validation accuracy: 13.9%
- Testing accuracy: 7.0%

Chosen Network

Network B - PCA

- Single-layer neural network
- 5 neurons in hidden layer
- 18 principal components that represent 81.17% of data variance
- Validation accuracy: 8.8%
- Testing accuracy: 3.1%

Conclusion and Further Thoughts

The initial results of this experiment suggest a lack of accuracy. However, weather data is difficult to predict with complete accuracy, so these results are more promising than they may appear at first glance. Further investigation may yield a more accurate, fine-tuned result. Additional processing or dataset setup may alleviate the issues and result in better outcomes. For example, using 12-hour observations over multiple years could help alleviate issues caused by outlying datapoints. More computational power would also allow for a larger initial dataset, yielding more accurate results.

Conclusion and Further Thoughts

- Arnx, A. (2019, January 13). First Neural Network for beginners explained (with code). Medium.
- Jaruszewicz, M., & Mandziuk, J. (2002). Application of PCA method to weather prediction task. Proceedings of the 9th International Conference on Neural Information Processing, 2002. ICONIP '02.
- Raschka, S. (2015, January 27). Principal component analysis. Dr. Sebastian Raschka.