Precipitation Data Analysis

Yuanchen Lu ylu36
Fogo Tunde-Onadele oatundeo
Zhuo Li zli36
Chen Zhao czhao13

November, 2017

Outline

- Background
- Introduction
- Approach
- Data Processing
- Model Evaluation
- Sensor selection
- Conclusion

Weather prediction

- Humans and Industries rely on weather predictions for high performance

- In general, observations of the atmosphere initialize models that utilize fluid dynamics equations to predict future atmospheric state.

- Forecasting is complex
  - Many variables involved
  - Numerical equations lead to approximate results

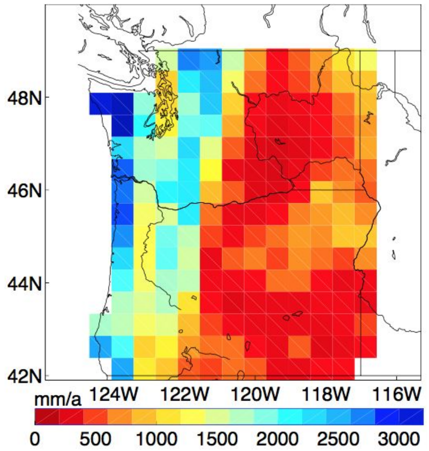
Project Introduction

Using methods learned from this course

- Goal 1: Predict precipitation levels of a Pacific NorthWest region of the US
- Goal 2: Determine best sensor locations for prediction

Source

- Researchers from the University of Washington
    - 46 years of daily precipitation data (1949-1994)
    - Map of observation stations



Widmann, M. and C. S. Bretherton, 2000: Validation of Mesoscale Precipitation in the NCEP Reanalysis Using a New Gridcell Dataset for the Northwestern United States. J. Clim., 13, 1936-1950
http://research.jisao.washington.edu/data_sets/widmann/

Approach

Understanding Data Format

Handling Missing Values

Supervised Learning for Weather prediction

- Linear Regression, Ridge Regression, Lasso Regression
- Artificial Neural Network

Unsupervised Learning for Sensor Selection

- K Means Clustering

Python Packages

Numpy -- array and matrices; handle input file

netCDF4 -- convert input file format to a more familiar format

Sklearn -- machine learning library

Data Processing

Original data format -> .nc

Step 1: extract data by attribute (nc.variables(X))

Step 2: append data to a .csv file

Time	Lat	Lon	Precipitation
1900	46.9039	-123.75	4.7
1900	46.9039	-123.125	1.9
1900	46.9039	-122.5	3.4
1900	46.9039	-121.875	6.3
1900	46.9039	-121.25	3.9

```
NetCDF Global Attributes:
NetCDF dimension information:
    Name: lat
        size: 17
        type: dtype('float32')
        title: 'Latitude'
        units: 'degrees_north'
        scale_factor: 1.0
        add_offset: 0.0
    Name: lon
        size: 16
        type: dtype('float32')
        title: 'Longitude'
        units: 'degrees_east'
        scale_factor: 1.0
        add_offset: 0.0
    Name: time
        size: 16801
        type: dtype('float64')
        title: 'Time'
        units: 'days   since 1949- 1- 1  0:
        scale_factor: 1.0
        add_offset: 0.0
NetCDF variable information:
    Name: data
        dimensions: ('time', 'lat', 'lon')
        size: 4569872
        type: dtype('int16')
        long_name: 'mm/day'
        add_offset: 0.0
        scale_factor: 0.1
        missing_value: 32767
        units: 'mm/day'
```

Data Analysis

Step 1: 5-fold training data and testing data

Step 2: fit training data to each model (LR, Lasso, Ridge, ANN)

Step 3: for each trained model, compute MSE with testing data

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2$$

Model Comparison

LR Error rates: 2.48396122971

ANN Error rates for layer 3: 2.48580597654

ANN Error rates for layer 2: 1.59623754444

ANN Error rates for layer 1: 2.48382232418

Ridge Error rates: 2.48396122971

Lasso Error rates: 2.48393840128

Sensor Selection

- Unsupervised learning

- K-means clustering

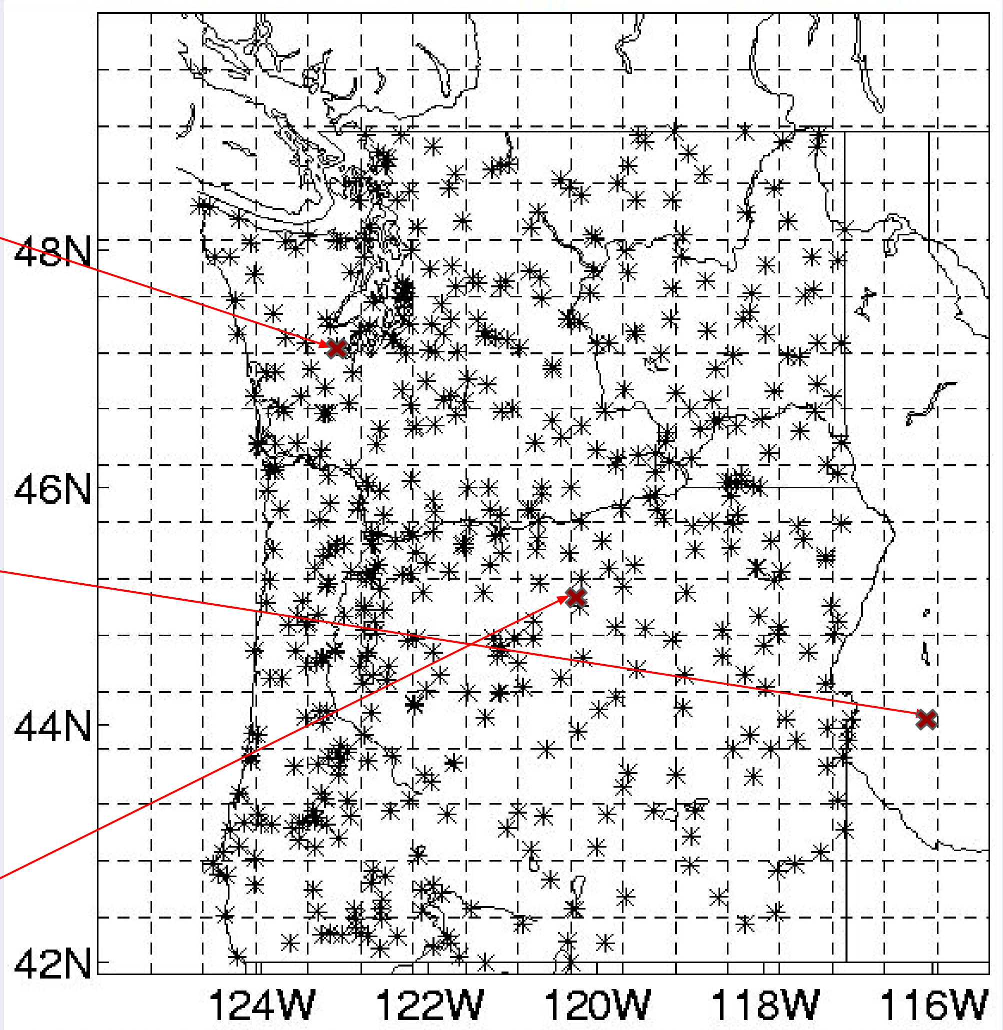
    Step 1: preprocess data

    Step 2: KMeans into 21 clusterings

Sensor Selection

```
[[ 47.67435709 -123.70737306]
 [ 44.93884639 -118.73325688]
 [ 43.37745139 -122.2127809 ]
 [ 48.01397463 -118.72384481]
 [ 46.05360835 -121.94698495]
 [ 43.70611638 -123.4399841 ]
 [ 48.5841914  -120.90481243]
 [ 44.88613152 -122.28392491]
 [ 47.30819377 -116.51189446]
 [ 44.38018404 -116.95361812]
 [ 42.63545137 -123.93049846]
 [ 46.39423132 -123.52350675]
 [ 48.59849092 -122.06961496]
 [ 42.54867135 -121.04210993]
 [ 47.44717742 -121.0198657 ]
 [ 46.01279307 -117.48340459]
 [ 48.54964365 -117.29367089]
 [ 42.73911987 -119.29924242]
 [ 45.01862586 -123.5549718 ]
 [ 45.12746901 -120.54056407]
 [ 47.32031932 -122.09779775]]
```

The location of stations contributing to the data set.



Conclusion

- Presented a 2-layer Artificial Neural Network to predict precipitation

- Proposed the best sites for placing sensors -- 21 calculated locations (lat, lon) of cluster centroids

- Future Development
  - Can explore other learning models
  - Better ways to store precipitation data

References

1. Widmann, M. Bretherton, C.S. (2000) Validation of Mesoscale Precipitation in the NCEP Reanalysis Using a New Gridcell Dataset for the Northwestern United States.
2. Daly, C., R. P. Neilson, and D. L. Phillips (1994) A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J. Appl. Meteor.,33, 140–158.
3. G. Taylor, and W. Gibson, (1997) The PRISM approach to mapping precipitation and temperature. Preprints, 10th Conf. on Applied Climatology, Reno, NV, Amer. Meteor. Soc., 10–12.