CS670/470 Team Project Phase 1: Data Pre-processing

UMass-Boston 11/06/2017

1 Educational Goal

Practice how to construct data samples in proper format for machine learning algorithms from raw data.

2 Details

Project goal: Construct data samples in proper format that can be used as input data for various machine learning algorithms. The construction procedure includes feature construction and class label assignment.
Due Date: 4:00 pm, November 16, 2017
Programming language: Python 2.7.

3 Problem Formulation

Let S be a set of locations, m be a vector of meteorological variables, M be a set of m, and $\mathcal{C}$ be the precipitation risk level. For each location s, its precipitation risk level at a specific time t is denoted as $\mathcal{C}_t^s$, and its historical climate information in a fixed time-period q is given as $M^s_{(t-q+1)\sim t} = \{m^s_{t-q+1}, m^s_{t-q+2}, ..., m^s_t\}$, where $m^s_{t_i}$ presents the vector of variables collected in the location s at time t_i. And the historical climate information of all locations in the fixed time-period q is denoted as $M^S_{(t-q+1)\sim t} = \{M^{s_1}_{(t-q+1)\sim t}, M^{s_2}_{(t-q+2)\sim t}, ...\}$. Given the lead time as p, then the long-lead extreme precipitation forecasting can be formulated as to predict $\mathcal{C}^s_{t+p}$, based on the history $M^S_{(t-q+1)\sim t}$

4 Raw Data

The data we will use for machine-learning experiments is the historical meteorological data (including 9 variables) of the whole Northern Hemisphere (5,328 locations) over 30 years (1980-2010) and the historical spatial average precipitation data of the state Iowa from the same time period.

4.1 10 variables

Nine meteorological variables + the historical spatial average precipitation data of the state Iowa.

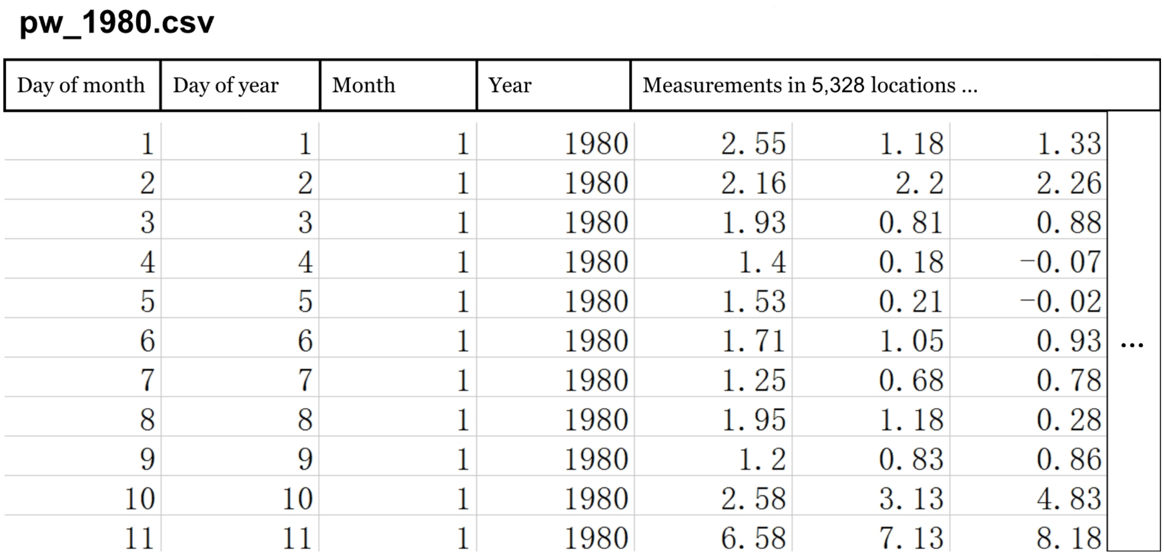
4.2 Data files

Each variable has 31 CSV files (one file per year).

4.3 Nine meteorological variables

Meteorological Variables	
Z300	300hPa Geopotential Height
Z500	500hPa Geopotential Height
Z1000	1000hPa Geopotential Height
U300	300hPa Zonal Wind
V300	300hPa Meridional Wind
U850	850hPa Zonal Wind
V850	850hPa Meridional Wind
T850	850hPa Temperature
PW	Precipitable water

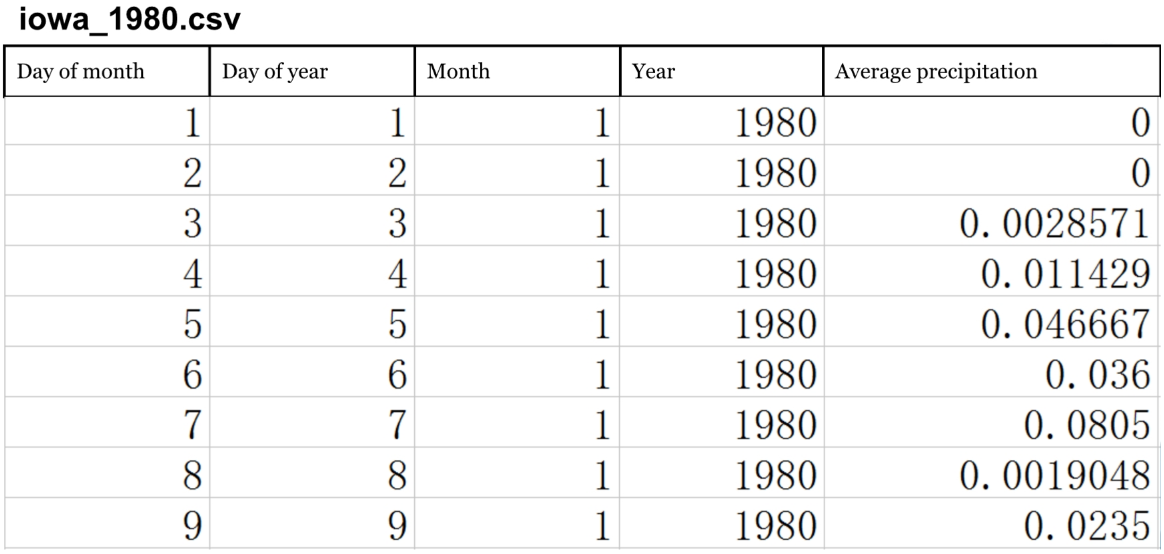
4.4 A CSV file example of the nine meteorological variables



pw_1980.csv

Day of month	Day of year	Month	Year	Measurements in 5,328 locations ...			
1	1	1	1980	2.55	1.18	1.33	
2	2	1	1980	2.16	2.2	2.26	
3	3	1	1980	1.93	0.81	0.88	
4	4	1	1980	1.4	0.18	−0.07	
5	5	1	1980	1.53	0.21	−0.02	
6	6	1	1980	1.71	1.05	0.93	...
7	7	1	1980	1.25	0.68	0.78	
8	8	1	1980	1.95	1.18	0.28	
9	9	1	1980	1.2	0.83	0.86	
10	10	1	1980	2.58	3.13	4.83	
11	11	1	1980	6.58	7.13	8.18	

4.5 A CSV file example of the historical spatial average precipitation data of the state Iowa



iowa_1980.csv

Day of month	Day of year	Month	Year	Average precipitation
1	1	1	1980	0
2	2	1	1980	0
3	3	1	1980	0.0028571
4	4	1	1980	0.011429
5	5	1	1980	0.046667
6	6	1	1980	0.036
7	7	1	1980	0.0805
8	8	1	1980	0.0019048
9	9	1	1980	0.0235

5 Data Pre-Processing

5.1 Features

Suppose the set of locations $S = 5328$, the vector of meteorological variables $m = 9$, then the historical climate information in a fixed time period $q = 10$ days is:

$$M^S_{(t-9)\sim t} = \underbrace{\begin{array}{|c|c|c|c|c|c|c|c|c|c|}\hline m^S_{t-9} & m^S_{t-8} & m^S_{t-7} & m^S_{t-6} & m^S_{t-5} & m^S_{t-4} & m^S_{t-3} & m^S_{t-2} & m^S_{t-1} & m^S_{t} \\ \hline\end{array}}_{\textbf{5328* 9 * 10 features}}$$

5.2 Labels

We label a time window as being a period of extreme precipitation if the total rainfall in that period reaches a historically high level (above the 95^{th} percentile of the historical record). For this project, we'll be using a window size of 15 days. Since each window has a unique start date, we can assign the label of the window starting on that date to the date itself. So if the period from July 1st to July 15^{th} is above the 95^{th} percentile for rainfall, we mark July 1 as a positive example, as the beginning of an extreme precipitation period.

Suppose the lead time $p = 5$ days, the target area s is Iowa, the data we used to create labels is the historical spatial average precipitation data of the state Iowa, then the problem formulation can be presented as follows:

$$M^S_{(t-9)\sim t} = \begin{array}{c}\textbf{5328 * 9 * 10 features} \\ \begin{array}{|c|c|c|c|c|c|c|c|c|c|}\hline m^S_{t-9} & m^S_{t-8} & m^S_{t-7} & m^S_{t-6} & m^S_{t-5} & m^S_{t-4} & m^S_{t-3} & m^S_{t-2} & m^S_{t-1} & m^S_{t} \\ \hline\end{array} \\ \Downarrow \\ \mathcal{C}^s_{t+5}\end{array}$$

where $\mathcal{C}^s_{t+5} = 1$, if $\sum_{t+5}^{t+19} PW_{Iowa}$ is above 95 percentile of any sum of 15 days' precipitations in the historical records. Otherwise, $\mathcal{C}^s_{t+5} = 0$.

6 Tasks

Finish the data pre-processing to create the experimental samples (features and labels) to be used for next term project phase 2.

6.1 Examples

$Sample_1$: $M^S_{(1-1-1980)\sim(1-10-1980)} \Rightarrow \mathcal{C}^s_{(1-15-1980)}$
$Sample_2$: $M^S_{(1-2-1980)\sim(1-11-1980)} \Rightarrow \mathcal{C}^s_{(1-16-1980)}$
$Sample_3$: $M^S_{(1-3-1980)\sim(1-12-1980)} \Rightarrow \mathcal{C}^s_{(1-17-1980)}$
$\vdots$

7 Submission Requirements

Submit the code and a brief explanation of how you generate the samples through your UMassOnline account. One team can submit one solution through the team lead's UMassOnline account, but the submission should include the team members' name list.