0.1 Get Raw Data

The first step was to get the raw weather data from the website : `http://www.ncdc.noaa.gov/orders/qclcd/`

This website is hosted by the National Center for Environmental Information which is part of the National Oceanic and Atmospheric Administration (NOAA). It provides hourly weather measures from weather station all over the US. The data are available from zipfiles updated daily.

Our goal was to get the weather data over the time period : 2008/12/21-2010/2/10. This data will be the inputs of our dataset (weather features) where the data from the Y2E2 building of Stanford will be the output (solar panel energy production).

To get an estimation of the weather conditions in Stanford (*zip5 = 94305*), we located the n = *3* closest weather station :

Key	Name	Distance (km)	City	State
23289	Airport of Santa Clara County	5.8	Palo Alto	CA
23244	Moffett Federal Field Airport	10.4	Mountain View	CA
93231	San Carlos Airport	12	San Carlos	CA

To get the weather data from these 3 weather station, we used a GitHub program available here : `https://github.com/sborgeson/local-weather`.

The data was captured using our Python program : *getWeatherDataStanford.py*. These program use 2 functions from the GitHub Python files :

- *stationList(zip5,2009,01,n,preferredDistKm)* : this function output the details about the 3 closest weather stations.
- *weatherMonths(zip5,start,end,hourly,subset,n,preferredDistKm)* : this function output a matrix of all the weather data examples for the 3 closest stations.

Our Python program output a csv file (*rawWeatherDataStanford.csv*) : the weather data examples are the rows of the csv file.

We pre-selected $n = 8$ weather features for the raw data :

Key	Weather feature	Unit	Remark
4	Sky condition	% range	"CLR", "FEW", "SCT", "BKN", "OVC"
6	Visibility	miles	
12	Temperature	C	"Dry Bulb" in the original files
20	Dew point	C	when vapor water starts to condensate
22	Relative humidity	%	ratio of the partial to the equilibrium pressure of vapor water
24	Wind speed	mph	
30	Station pressure	inchHg	inch of mercure
42	Altimeter	inchHg	altitude-pressure of the station

The key of a feature is its column index in the original NOAA files : the indexes of the preselected weather features where stored in the parameter *subset*.

0.2 Data processing

The raw data is not straight away usable.

- feature values are missing for some examples : we replace them by mean values. It just adds some noise to the data ;
- the value of the feature *Sky Condition* is a String : we need to convert it to a numeric value.

The corresponding processed feature is called *Cloud coverage* and is in %, of the sky which is covered by clouds.

A problem was that the resolution of measures were not coherent between the weather station : some of them had an hour resolution, some of them half an hour, and some of them didn't even have a regular resolution. Thus, the first step was to average the measures over an hour.

Then we need to spatially average every weather feature over the 3 closest weather station. We used the formula of the barycenter to take the distance of each station into account.

For a specific (at a given hour, day, month, year) feature x_j, the average value is :

$$x_j = \frac{d_A x_{j,A} + d_B x_{j,B} + d_C x_{j,C}}{d_A + d_B + d_C}$$

Where :

$$\begin{cases} d_A & : \quad \text{the distance from weather station A to Stanford} \\ x_{j,A} & : \quad \text{the value of feature } x_j \text{ measured by weather station A} \end{cases}$$

With this formula : the closest a weather station is, the more weight it has.

Cont.

Finally, after this data-cleaning process, we are left with 1 dataset example per hour per day for the 2008/12/21-2010/2/10 period.

$$m = (365 + 52) * 24$$

We have $m = \mathit{10\ 008}$ examples.

The End.