Capstone :Prediction of Hourly Rainfall

Author: Venkatesh Viswanathan

Date: 11/20/2015

Introduction

Rainfall is highly variable across space and time, making it notoriously tricky to measure. Rain gauges can be an effective measurement tool for a specific location, but it is impossible to have them everywhere. In order to have widespread coverage, data from weather radars is used to estimate rainfall nationwide. Unfortunately, these predictions never exactly match the measurements taken using rain gauges. In order to improve the efficiency of the rain fall measurements the weather department has installed dual polarimetry. Following report describes the problem stated in the project and proposes a solution. The data is used is from kaggle competition and relevant link is provided below.

https://www.kaggle.com/c/how-much-did-it-rain-ii

Problem

The key problem is to predict the amount of rainfall for remaining 10/11 days given the rainfall information for the 1st 20 days from April to August in the corn growing states. The existing data consists of Reflectivity, Reflectivity Composite, expected values and other data.

Reflectivity

Return echoes from targets ("reflectivity") are analyzed for their intensities to establish the precipitation rate in the scanned volume. The *Z*–*R* relationship developed by J. S. Marshall and W. M. Palmer (1948) consistent with an exponential distribution. The relationship is $Z = 200R^{1.6}$, where Z (mm^{6} m^{-3}) is the reflectivity factor and R (mm h^{-1}) is the rainfall rate. The relationship is sometimes generalized to the form $Z = a'\mathbf{R}^{b}$**, where *a* and *b* are adjustable parameters.** In general it represents the rainfall at the gauges and its related to amount of rainfall per hour

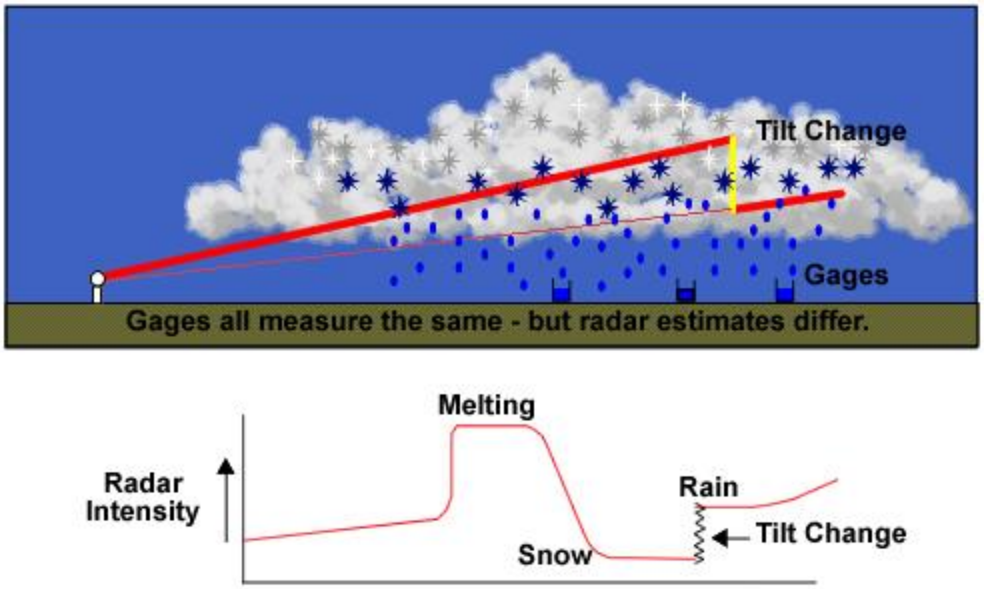

Fig 1. Illustration of the radar transmission and the reception by gauges

Solution

There are two type of data sets given.

1. Training data with Reflectivity information and Expected (Actual) rainfall information.
2. Test data has the Reflectivity information and the Expected values need to be found.

The approach consists of three parts

1. Use the Reflectivity and Expected information from the training data and come up with a Regression Model (ex: likely linear as log(Z)=log(a)+b*log(R))
2. The coefficients that is obtained in the previous step is leveraged to predict the Expected values in the test data.
3. Analyze the results using cross validation techniques. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *validation set* or *testing set*). To reduce variability, multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.