Predicting Traffic Measurement

Cleaning II

Task Overview

- Description

  Given a collection of erroneous measurement data (e.g. flow, speed, occupancy), where most of the measurement is correct. You are asked to predict the correct **flow** measurement.

- Example Erroneous Measurement

	A	B	C	D	E	F	G
1	trial_id	lane_id	measurement_start	speed	flow	occupancy	quality
2	c_06_09_000000000	12	2006-09-01T00:00:07-04:00	65	0	0	0
3	c_06_09_000000001	13	2006-09-01T00:00:07-04:00	63	3	2	0
4	c_06_09_000000002	14	2006-09-01T00:00:07-04:00	64	-2	1	0
5	c_06_09_000000003	15	2006-09-01T00:00:07-04:00	59	4	3	0
6	c_06_09_000000004	16	2006-09-01T00:00:07-04:00	66	5	1	0
7	c_06_09_000000005	17	2006-09-01T00:00:07-04:00	0	255	4	0
8	c_06_09_000000006	18	2006-09-01T00:00:07-04:00	67	13	7	0
9	c_06_09_000000007	19	2006-09-01T00:00:07-04:00	61	4	1	0
10	c_06_09_000000008	20	2006-09-01T00:00:07-04:00	65	0	0	0

Data (Same as Cleaning I)

Measurements are divided by zones, where each zone can have one or more detectors. Detectors in the same zone are geographically next to each other. For each zone, you are given the following data:

1	77	132
2	84	144
3	78	115
4	91	141
5	96	149

flow.tsv

1	5	9
2	4	10
3	5	9
4	4	8
5	6	12

occupancy.tsv

1	68.9000015259	59.0
2	66.4000015259	55.2999992371
3	68.9000015259	52.0999984741
4	72.0	62.7000007629
5	68.3000030518	50.2999992371

speed.tsv

1	2013-06-18T13:41:07
2	2013-06-18T13:47:26
3	2013-06-18T13:53:01
4	2013-06-18T13:59:28
5	2013-06-18T14:04:04

timestamp.tsv

- R = #columns = #lanes: Each column is corresponding to one lane (e.g. data by the same detector).
- C = #rows = #timestamp: Each row represents measurement at specific time given by timestamp.tsv
- Missing data: flow, occupancy and speed can have missing data. If a measurement of specific lane at specific timestamp is missing, then that corresponding field is empty.
- Discontinuous timestamps: most of the time, the timestamp increases with fixed interval. But, this is not guaranteed. You should NOT assume nearby rows are measured in nearby time intervals. Always check the timestamp to see if they are continuous or not.

Data (Output of Cleaning I)

In Cleaning I (Lab 9), you are asked to predict the probability density for a specific measurement being correct. So you can output like:

1	0.0001	0.003
2	0.0003	0.004
3	0.0004	0.0002
4	0.0005	0.0003
5	0.0007	0.0003

prob.tsv

Where each value in the cell is the probability (density) that the corresponding measurement is being correct. So higher value means the corresponding measurement is more reliable. In this lab, we will make use of the **reliable** data to predict correct flow values.