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.

Method #1: Nearby Lanes + LR

Nearby lanes usually provide useful information about flow measurements. For example, if flow of one lane is very high, then we expect the flow of another lane will also be high because vehicles tend to automatically balance the load of different lanes. We can use this relationship to predict the correct flow measurement, as given by a linear regression model:

$$\text{Predicted(flow)} = a*\text{Nearby\_Measured(flow)} + b$$

where (a,b) are LR model parameters. The confidence of this prediction can be estimated by the probability density of the nearby measurement, as:

$$\text{Confidence(flow)} = \text{Prob\_Desensity(Nearby(flow, speed, occupancy))}$$

Method #2: Nearby Timestamps + Weighted Sum

Measurements between consecutive time intervals are usually similar. Thus, we can predict the flow measurement by average of preceding and following time intervals, as given by:

$$\text{Predicted(flow)} = w1*\text{Preceding\_Measured(flow)} + w2*\text{Following\_Measured(flow)}$$

where (w1,w2) are weights between (0,1) such that w1+w2=1. We can calculate the (w1,w2) as:

$$w1 = c1/(c1+c2), \quad w2 = 1-w1$$

where c1 is probability density of preceding measurement (calculated in Lab 9), similar for c2:

$$c1 = \text{Prob\_Desensity(Preceding(flow, speed, occupancy))}$$

Finally, the confidence of this prediction can be estimated as: Confidence(flow) = min(c1,c2).

Method #3: Keep Measurement Unchanged

Most of the measurements are correct, so keeping all flow measurements unchanged might not lead to a result that is too bad. In this case, we simply predict a correct flow value with the measured flow value, as:

$$\text{Predicted(flow)} = \text{Measured(flow)}$$

The confidence of this prediction can be estimated by probability density of this point (calculated in Lab 9):

$$\text{Confidence(flow)} = \text{Prob\_Desensity((flow, speed, occupancy))}$$

Merging Multiple Predictions

We have described three methods to predict correct flow values. Each method outputs a predicted flow value, as well as confidence score of this prediction. We can thus merge multiple predictions by:

$$\text{Merged(flow)} = w1*\text{Predicted\_1(flow)} + \ldots + w3*\text{Predicted\_3(flow)}$$

where (w1,w2,w3) are weights between (0,1) such that w1+w2+w3=1. The w1 can be defined as (similar to w2 and w2):

$$w1 = c1/(c1+c2+c3)$$

where c1 is confidence of the flow prediction given by method 1.

Submission

For each of the zones (3445, 3532, 3451, 3232, 1160), generate one file named “zone_id.flow.txt”. It should contain predicted flow corresponding to given measured flow in flow.tsv.

- The zone_id.flow.txt file should have R rows, and C columns. When a measurement is missing, you should use empty string in the corresponding place.
- Submit a report (**pdf file**) named report.pdf describing all details.
- Submit source codes. It should also include a shell runner script named “clean_one_zone.sh”, which takes one argument (root directory of a zone) as input, and outputs zone_id.flow.txt. Example:
  $bash clean_one_zone.sh path/to/zone/3445
  This program should output 3445.flow.txt under current directory. You can assume folder 3445 has the same folder structure as given data (e.g. containing flow.tsv, timestamps.tsv, …, prob.tsv).
- Zip all the files, and submit one zipped file named “lab10.zip”. Your zipped file should have this folder structure:
  - Lab10
    - report.pdf
    - zone_id.flow.txt (5 files because we have 5 zones)
    - src (contain all source codes)
    - clean_one_zone.sh
- **We will execute your program under Lab10 folder, by: $bash clean_one_zone.sh /path/to/zone**
- You should optimize your codes, because we may also evaluate your work based on running time on our machines.